Bayesian inference and forecasting methods in public transit systems

Xiaoxu Chen

Department of Civil Engineering
McGill University
Montreal, Quebec, Canada

July 18, 2024

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Doctor of Philosophy

Abstract

Public transit is crucial for reducing congestion and emissions, improving access to essential services, and promoting sustainable urban growth. To enhance transit services and attract ridership, transit agencies must address critical research problems within public transit systems, such as forecasting bus travel time/passenger occupancy and estimating origin-destination (OD) matrices. With the advent of automatic data collection in transit systems, e.g., automated fare collection (AFC) systems, automated vehicle location (AVL) systems, and automated passenger counts (APC) systems, both practitioners and scholars are increasingly focusing on data-driven approaches to address the above problems and enhance transit services. Most previous data-driven approaches for transit studies have predominantly relied on deterministic models, which only provide point estimation and fail to offer uncertainty quantification. To fill this gap, this thesis aims to develop probabilistic models based on Bayesian statistics for four important research problems in public transit systems: (1) link travel time correlation inference, (2) bus travel time forecasting, (3) bus passenger occupancy forecasting, and (4) transit OD matrices inference.

The proposed Bayesian inference and forecasting models in this thesis correspond to four published or under-review manuscripts. All the models developed in the thesis are tested by real-world transit data. Firstly, this thesis develops a Bayesian Gaussian model to estimate the link travel time correlation matrix of a bus route using smart-card-like data. This method overcomes the small-sample-size problem in correlation matrix estimation by borrowing/integrating those incomplete observations (i.e., with missing/ragged values and overlapped link segments) from other bus routes. Numerical experiments are conducted to evaluate model performance and results show that the proposed method can make an accurate estimation for travel time correlations with credible intervals. Secondly, this thesis proposes a Bayesian Gaussian mixture model for probabilistic forecasting of bus travel time. This approach can naturally capture the interactions between adjacent buses (e.g., correlated speed and smooth variation of headway), handle missing values in data, and depict the multimodality in bus travel time distributions. An efficient algorithm is proposed to obtain the posterior distributions of model parameters and make probabilistic forecasting. Results show that our approach significantly outperforms baseline models that overlook bus-to-bus interactions, in terms of both predictive means and distributions. Thirdly, this thesis develops a Bayesian Markov regime-switching vector autoregressive model to jointly forecast both bus travel time and passenger occupancy with uncertainty. The proposed approach naturally captures the intricate interactions among adjacent buses and adapts to the multimodality and skewness of real-world bus travel time and passenger occupancy observations. With this framework, the estimation of downstream bus travel time and passenger occupancy is transformed into a multivariate time series forecasting problem conditional on partially observed outcomes. Experimental validation using real-world data demonstrates the superiority of our proposed model in terms of both predictive means and uncertainty quantification compared to the baseline models. Finally, this thesis proposes a temporal Bayesian model to estimate transit OD matrices at the individual bus level using counts of boarding and alighting passengers at each stop. Specifically, the number of alighting passengers at subsequent bus stops, given a boarding stop, is modeled by a multinomial distribution. This method uses a latent variable matrix to parameterize the time-varying multinomial distributions through the softmax transformation and employs matrix factorization to parameterize it into a mapping factor matrix and a temporal factor matrix. Gaussian process priors are imposed on the columns of the temporal factor matrix. The model is validated using real-world data of three bus routes (short, medium, long) and results demonstrate that the proposed model can achieve accurate estimation and outperforms the iterative proportional fitting method. Moreover, this model can provide uncertainty quantification associated with estimation and parameter interpretation.

In summary, this thesis uses Bayesian statistics to develop probabilistic inference and forecasting models for the above problems in public transit systems. The proposed models can improve the accuracy of inference/forecasting and provide uncertainty quantification, which is crucial for transit agencies to optimize the management and operation of transit systems.

Résumé

Les transports en commun sont essentiels pour réduire la congestion et les émissions, améliorer l'accès aux services essentiels et promouvoir une croissance urbaine durable. Pour améliorer les services de transport et attirer les usagers, les agences de transit doivent aborder des problèmes de recherche critiques au sein des systèmes de transport public, tels que la prévision du temps de trajet / de l'occupation des passagers des bus et l'estimation des matrices origine-destination (OD). Avec l'avènement de la collecte automatique des données dans les systèmes de transport, par exemple, les systèmes de collecte automatique des tarifs (AFC), les systèmes de localisation automatique des véhicules (AVL) et les systèmes de comptage automatique des passagers (APC), les praticiens et les universitaires se concentrent de plus en plus sur des approches basées sur les données pour aborder les problèmes ci-dessus et améliorer les services de transport. La plupart des approches précédentes basées sur les données pour les études de transport ont principalement reposé sur des modèles déterministes, qui ne fournissent qu'une estimation ponctuelle et échouent à offrir une quantification de l'incertitude. Pour combler cette lacune, cette thèse vise à développer des modèles probabilistes basés sur la statistique bayésienne pour quatre problèmes de recherche importants dans les systèmes de transport public : (1) l'inférence de corrélation du temps de trajet entre les liens, (2) la prévision du temps de trajet des bus, (3) la prévision de l'occupation des passagers des bus, et (4) l'inférence des matrices de transit OD.

Les modèles bayésiens d'inférence et de prévision proposés dans cette thèse correspondent à quatre manuscrits publiés ou en cours de révision. Tous les modèles développés dans la thèse sont testés par des données de transport réelles. Premièrement, cette thèse développe un modèle gaussien bayésien pour estimer la matrice de corrélation du temps de trajet entre les liens d'une route de bus en utilisant des données semblables à des cartes intelligentes. Cette méthode surmonte le problème de la petite taille d'échantillon dans l'estimation de la matrice de corrélation en empruntant/intégrant ces observations incomplètes (c'est-à-dire avec des valeurs manquantes/déchirées et des segments de lien chevauchants) d'autres routes de bus. Des expériences numériques sont menées pour évaluer la performance du modèle et les résultats montrent que la méthode proposée peut faire une estimation précise des corrélations de temps de trajet avec des intervalles crédibles. Deuxièmement, cette thèse propose un modèle de mélange gaussien bayésien pour la prévision probabiliste du temps de trajet des bus. Cette approche peut naturellement capturer les interactions entre les bus adjacents (par exemple, vitesse corrélée et variation douce de l'intervalle), gérer les valeurs manquantes dans les données et dépeindre la multimodalité dans les distributions de temps de trajet des bus. Un algorithme efficace est proposé pour obtenir les distributions postérieures des paramètres du modèle et faire des prévisions probabilistes. Les résultats montrent que notre approche surpasse significativement les modèles de base qui négligent les interactions bus à bus, tant en termes de moyennes prédictives que de distributions. Troisièmement, cette thèse développe un modèle autorégressif vectoriel à changement de régime markovien bayésien pour prévoir conjointement le temps de trajet des bus et l'occupation des passagers avec incertitude. L'approche proposée capture naturellement les interactions complexes entre les bus adjacents et s'adapte à la multimodalité et à l'asymétrie des observations réelles de temps de trajet des bus et d'occupation des passagers. Avec ce cadre, l'estimation du temps de trajet des bus en aval et de l'occupation des passagers est transformée en un problème de prévision de séries temporelles multivariées conditionné à des résultats partiellement observés. La validation expérimentale utilisant des données réelles démontre la supériorité de notre modèle proposé en termes de moyennes prédictives et de quantification de l'incertitude par rapport aux modèles de base. Enfin, cette thèse propose un modèle bayésien temporel pour estimer les matrices OD de transit au niveau des bus individuels en utilisant les comptages des passagers montant et descendant à chaque arrêt. Spécifiquement, le nombre de passagers descendant aux arrêts de bus subséquents, étant donné un arrêt de montée, est modélisé par une distribution multinomiale. Cette méthode utilise une matrice de variables latentes pour paramétrer les distributions multinomiales variables dans le temps à travers la transformation softmax et emploie la factorisation matricielle pour la paramétrer en une matrice de facteur de cartographie et une matrice de facteur temporel. Des a priori de processus gaussien sont imposés sur les colonnes de la matrice de facteur temporel. Le modèle est validé en utilisant des données réelles de trois itinéraires de bus (court, moyen, long) et les résultats démontrent que le modèle proposé peut réaliser une estimation précise et surpasse la méthode de réglage proportionnel itératif. De plus, ce modèle peut fournir une quantification de l'incertitude associée à l'estimation et à l'interprétation des paramètres.

En résumé, cette thèse utilise la statistique bayésienne pour développer des modèles d'inférence et de prévision probabilistes pour les problèmes ci-dessus dans les systèmes de transport public. Les modèles proposés peuvent améliorer la précision de l'inférence/la prévision et fournir une quantification de l'incertitude, ce qui est crucial pour les agences de transit afin d'optimiser la gestion et l'opération des systèmes de transport.

Acknowledgements

As I stand on the precipice of graduation, the days that once seemed endless have passed all too quickly. Reflecting on this journey, I am filled with nostalgia for the vibrant and enriching experiences at McGill University, and I am deeply grateful for the myriad of individuals who have been part of this significant chapter of my life.

I extend my deepest appreciation to my supervisor, Prof. Lijun Sun. His mentorship plays a critically important role in shaping my academic journey—from ideating research concepts to the meticulous preparation of my papers. Prof. Sun's invaluable advice and rigorous academic standards not only enhance the quality of my dissertation but also profoundly develop my scholarly skills. His support and encouragement are crucial to my accomplishments, and his influence extends beyond academia into my personal and professional growth. Similarly, I am thankful to Prof. Saeid Saidi, who mentored me during my visiting time at the University of Calgary, enriching both my studies and my life with his guidance.

Next, I would like to express my gratitude to my thesis committee members, Prof. Luc Chouinard, Prof. Adrian Liu, Prof. Jiangbo Yu, Prof. Yazhou (Tim) Xie, Prof. Zachary Patterson, and Prof. Zhenliang Ma, for their invaluable assistance and insightful feedback. I am especially thankful to my thesis examiners, Prof. Jiangbo Yu and Prof. Zhenliang Ma, whose detailed comments significantly enhanced the quality of this thesis.

I am also grateful to McGill University and the China Scholarship Council for their financial support, which was essential for the completion of my Ph.D. project.

Being a part of my research group was a privilege. The group's camaraderie, humor, and intellect made my time at McGill truly rewarding. I am fortunate to have worked alongside such exceptional colleagues and friends, including Yuankai Wu, Zhanhong Cheng, Xinyu Chen, Dingyi Zhuang, Wenshuo Wang, Jiawei Wang, Jingbo Tian, Fuqiang Liu, Zhihao Zheng, Xudong Wang, Sicong Jiang, Qiujia Liu, Xinghang Zhu, Fan Wu, Xiting Zhang, Jian Yuan, Xian Chen, Mengyi Sha, Mengying Lei, Chengyuan Zhang, Tianyu Shi, Seongjin Choi, Sohyeong Kim, Alicia Qiao, Kehua Chen, Kevin Hou, Mengying Zhu, Dan

Qiang, Jiawen Liu, Ziyu Chen, Tianyang Luo, and many others whose fellowship enriched my daily life.

Heartfelt thanks to my parents, whose boundless love and sacrifices made all this possible. I also owe a debt of gratitude to my brother, Xiaotong Chen, whose support has been a constant in my life. Lastly, I am profoundly thankful to my girlfriend, Shan Li, whose love and support have been my sanctuary; I am devoted to making our shared future just as meaningful.

To all who have accompanied me on this journey, your roles in my academic, personal, and professional spheres have been indispensable. Thank you for being part of my story.

Contribution of Authors

This is an article-based thesis and the main contents from Chapter 3 to Chapter 6 are four journal articles. Details of the four articles are listed below.

- **Chen, X.**, Cheng, Z., Sun, L. 2022. Bayesian inference for link travel time correlation of a bus route. arXiv preprint arXiv:2202.09485. (Under review in Transportmetrica B: Transport Dynamics)

- **Chen, X.**, Cheng, Z., Jin, J. G., Trépanier, M., Sun, L. 2023. Probabilistic forecasting of bus travel time with a Bayesian Gaussian mixture model. *Transportation Science*, 57(6), 1516-1535.

- **Chen, X.**, Cheng, Z., Schmidt, A. M., Sun, L. 2024. Conditional forecasting of bus travel time and passenger occupancy with Bayesian Markov regime-switching vector autoregression. arXiv preprint arXiv:2401.17387. (Under review in Transportation Research Part B: Methodological)

- **Chen, X.**, Cheng, Z., Sun, L. 2024. Bayesian inference of time-varying origin-destination matrices for transit services. arXiv preprint arXiv:2403.04742. (Under review in Transportation Science)

I declare that I am the first author of the four articles. My contributions to the four articles include designing models and experiments, implementing and validating models, and writing manuscripts. The other authors are Prof. Lijun Sun (my supervisor), Prof. Martin Trépanier, Prof. Alexandra M. Schmidt, Prof. Jiangang Jin, and Dr. Zhanhong Cheng. They offer guidance, insightful comments, and editorial revisions for the articles. Besides the above four articles, the rest part of the thesis is completed by me. Thanks to my supervisor for helping me proofread the thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

Public transit plays a crucial role in modern urban environments by providing an efficient, cost-effective, and environmentally friendly alternative to personal vehicles. Its significance extends beyond mere transportation; it facilitates greater accessibility to employment, education, and healthcare for a diverse demographic, including those without access to a car. By reducing the number of vehicles on the road, public transit helps decrease traffic congestion, lower greenhouse gas emissions, and reduce the urban carbon footprint. In the 2030 Agenda for Sustainable Development, the United Nations has emphasized the critical role of public transportation in shaping a sustainable society (United Nations, 2015). However, despite the growing investment in infrastructure, North American cities have not seen rapid growth and even observed a decline in ridership in recent years, even before the COVID-19 pandemic (Erhardt et al., 2022). Therefore, transit agencies are continually striving to enhance their services to attract more passengers.

To provide better transit services, transit agencies concentrate on addressing two primary types of problems: (1) Demand/supply inference and forecasting. It includes inferring and forecasting passenger flow (demand) and travel time (supply). Accurate forecasting of passenger flow and travel time can help travelers make informed travel plans in terms of mode choice, route choice, and even vehicle choice (e.g., waiting for a less crowded bus or boarding a full vehicle). Furthermore, the insights gained from inferring and forecasting passenger flow and travel times serve as foundational data for addressing the subsequent type of problem, demonstrating the integral role in the overall optimization of transit services. (2) Transit design and optimization. It involves route/network design, scheduling (frequency or timetable), and resource allocation. Through employing advanced optimization algorithms and simulation techniques, agencies can expand service coverage, minimize waiting times for passengers, and enhance the reliability of the entire transit network.

With the advent of automatic data collection in transit systems, both practitioners and scholars are increasingly focusing on data-driven approaches to address the above problems and enhance transit services. The data collection systems include automated fare collection (AFC) systems, automated vehicle location (AVL) systems, and automated passenger counts (APC) systems. AFC data typically includes passenger boarding times and locations, which provide valuable insights into passenger travel patterns and help transit agencies optimize service operations. AVL data records positions and speeds for all transit vehicles, which is crucial for transit agencies to understand the travel time variability and provide accurate estimated time of arrivals. APC data includes passenger boarding/alighting counts at each stop, which can provide crowdedness information for passengers and help agencies control crowd levels and optimize scheduling.

Most previous data-driven approaches for transit services have predominantly relied on deterministic models. For example, there are many deterministic forecasting models for bus travel time, such as Artificial Neural Network (Gurmu and Fan, 2014), Support Vector Machine (Yu et al., 2011; Kumar et al., 2013), K-nearest neighbors model (Kumar et al., 2019), Long-Short-Term Memory neural network (He et al., 2018), and various hybrid models (Yu et al., 2018). However, such deterministic forecasting models only can provide point estimation and lack the capacity to quantify uncertainty. Transit systems have complicated operations and suffer serious uncertainty caused by many factors such as driving behaviors, traffic congestion, accidents, road conditions, weather, special events, etc. Therefore, it is necessary to develop probabilistic forecasting models for bus travel time, which provide not only a point estimate but also a forecasted probability distribution. For other problems, most solutions are also based on deterministic methods and we should consider revisiting them from a probabilistic perspective. Bayesian statistics have a long history of development and application across a broad range of domains including time series forecasting, pattern recognition, causal inference, and even optimization, etc. By developing Bayesian models for transit problems, researchers and practitioners can incorporate uncertainty quantification to yield more robust and reliable transit service optimization.

1.2 Bayesian Inference and Forecasting

We here introduce two important concepts in Bayesian statistics: Bayesian inference and Bayesian forecasting.

1.2.1 Bayesian Inference

Bayesian inference is a method of statistical inference that updates the probability for parameters as more evidence or information becomes available (Gelman et al., 2013). It is based on Bayes' theorem, which provides a way to update prior beliefs with available data to form posterior beliefs. This approach allows for the incorporation of prior knowledge into the analysis, making it particularly useful in situations where data may be limited. More importantly, Bayesian inference also provides a powerful framework for quantifying uncertainty, making it an essential tool in various scientific and practical applications where understanding uncertainty is critical.

In Bayesian inference, we start with a prior distribution that represents our initial beliefs about the parameters of interest. Next, we calculate the likelihood of the observed data given the parameters. By applying Bayes' theorem, we combine the prior distribution and the likelihood to obtain the posterior distribution, which represents our updated beliefs about the parameters after considering the evidence from observations. In general, it is difficult to derive analytical solutions for the posterior distributions of parameters with complex priors and likelihood structures. A common solution is Markov chain Monte Carlo (MCMC) sampling, which can approximate the posterior distribution when it is computationally infeasible to calculate it directly.

1.2.2 Bayesian Forecasting

Bayesian forecasting is an extension of Bayesian inference that focuses on predicting future observations or outcomes based on current data and updated beliefs. It involves using the posterior distribution obtained from Bayesian inference to make probabilistic predictions about future events. This approach allows for the incorporation of uncertainty in the forecasts, providing a more comprehensive understanding of potential future scenarios.

In Bayesian forecasting, the posterior predictive distribution is derived by integrating over the parameters using the posterior distribution. This distribution captures the uncertainty in the parameter estimates and propagates it into the predictions, resulting in a range of possible future outcomes with associated probabilities. Bayesian forecasting is particularly valuable in time series analysis, where it can be used to model and predict future values of a time series based on historical data.

In this thesis, we aim to develop Bayesian statistical methods for four important research problems in public transit systems: (1) Bayesian inference for link travel time correlation of a bus route, (2) Bayesian forecasting for bus travel time, (3) Bayesian forecasting for passenger occupancy, and (4) Bayesian inference of time-varying origin-destination matrices from boarding/alighting counts.

1.3 Research Objectives

This thesis aims to develop Bayesian inference and forecasting methods for the research problems in public transit systems. The overview of research objectives is shown in Figure 1.1.

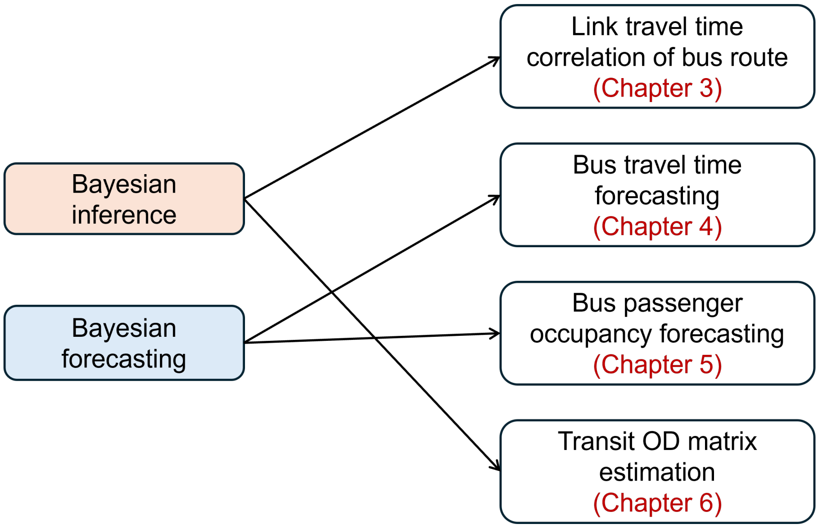

Bayesian inference
Bayesian forecasting
Link travel time correlation of bus route (Chapter 3)
Bus travel time forecasting (Chapter 4)
Bus passenger occupancy forecasting (Chapter 5)
Transit OD matrix estimation (Chapter 6)

Figure 1.1: Overview of research objectives

The details of the objectives are summarized as follows:

- **Objective 1: Develop a Bayesian model for inferring link travel time correlation of a bus route**. Link travel time correlation of a bus route is important for bus operation applications, such as scheduling and travel time forecasting. Most previous studies rely on either independent assumptions or simplified local spatial correlation structures. In the real world, however, link travel time on a bus route could exhibit complex correlation structures, such as long-range correlations (e.g., a delayed bus is more likely to be further delayed due to bus bunching), negative correlations (e.g., a bus that goes ahead of schedule may intentionally slow down to follow a pre-defined timetable), and time-varying correlations (e.g., different correlation patterns for peak and off-peak hours). Therefore, Chapter 3 aims to develop a Bayesian model to infer the link travel time correlation of a bus route, which could help to understand the characteristics of link travel time and further provide the foundation for bus travel time forecasting.

- **Objective 2: Develop a Bayesian model for bus travel time forecasting**. Accurate forecasting of bus travel time and its uncertainty is critical to service quality and operation of transit systems: it can help passengers make informed decisions on departure time, route choice, and even transport mode choice, and also support transit operators on tasks such as crew/vehicle scheduling and timetabling. However, most previous studies on bus travel time forecasting mainly center on making point estimation (i.e., deterministic forecasting) but ignore the importance of travel time uncertainty. Chapter 4 thus attempts to develop a Bayesian probabilistic model for bus travel time forecasting, which can provide predicted distributions of bus travel time.

- **Objective 3: Develop a Bayesian model for bus passenger occupancy forecasting**. Accurate occupancy forecasting along with uncertainty is important to travelers to make informed travel planning in terms of mode choice, route choice, and even vehicle choice (e.g., waiting for a less crowded bus or boarding a full vehicle). For transit agencies/operators, probabilistic forecasting could benefit the design of robust bus management strategies, such as bus route design, bus crowding control, and timetable design. Previous studies on forecasting bus passenger occupancy have predominantly employed deterministic approaches and they overlooked the strong correlations between passenger occupancy and travel time. In response to these challenges, Chapter 5 focuses on developing a joint Bayesian forecasting model for bus travel time and passenger occupancy.

- **Objective 4: Develop a Bayesian model for inferring time-varying origin-destination (OD) matrices from boarding/alighting counts**. OD demand matrices are crucial for transit agencies to optimize the management and operation of transit systems. Estimating OD matrices for transit systems from boarding/alighting counts data has been a long-standing research question for both practitioners and researchers.

This problem is quite challenging due to its underdetermined nature. Therefore, Chapter 6 develops a temporal Bayesian model for inferring transit OD matrices, which could make full use of prior information in observations.

1.4 Thesis Contributions

The detailed contributions of each model/application are provided individually in each chapter. The following is the high-level summary of the contributions of this thesis:

- **Contribution 1**: Most previous studies on transit problems have predominantly relied on deterministic models, which overlooked the uncertainty of the complex transit systems caused by stochastic factors such as traffic conditions and passenger behaviors. This thesis highlights that Bayesian statistics could be applied to model the uncertainties in transit systems. In other words, the Bayesian models could offer a more realistic representation of transit systems, enabling a better understanding of underlying patterns and relationships. This thesis develops Bayesian inference and forecasting methods for four important problems in transit systems, demonstrating the strong ability of Bayesian statistics to make probabilistic inference/forecasting and provide uncertainty quantification.

- **Contribution 2**: This thesis proposes several Bayesian inference and forecasting models for transit systems with improved accuracy. Specifically, the proposed time-dependent Bayesian Gaussian mixture model in Chapter 4 improves the forecasting accuracy of link and trip travel time. The developed Bayesian Markov Regime-switching vector autoregression model in Chapter 5 enhances the forecasting accuracy of bus travel time and passenger occupancy. The Bayesian inference model proposed in Chapter 6 achieves an improved performance for OD matrices estimation and provides a good uncertainty quantification. With these better models, this thesis helps to improve the forecasting and estimation to enhance transit services.

- **Contribution 3**: The proposed Bayesian inference and forecasting methods in this thesis are validated using real-world data. Besides the improved performance and uncertainty quantification, results demonstrate that estimated model parameters have good interpretations, which can help to better understand the latent patterns of passenger behaviors and transit operations. Specifically, the estimated correlation matrix of link travel time in Chapter 3 presents the complex characteristics such as long-range correlations and negative correlations. The patterns of travel time and passenger occupancy in Chapter 4 and 5 show that the interaction between adjacent buses like bus bunching has a strong influence on the travel time and passenger occupancy. The estimated patterns of passenger behaviors in OD matrix estimation in Chapter 6 present smoothly time-varying characteristics.

1.5 Thesis Organization

This is a manuscript-based thesis with seven chapters, where Chapter 3 to Chapter 6 are based on articles that were either submitted or published by peer-reviewed journals. The organization of the thesis is as follows:

- **Chapter 1** introduces the background, objectives, and contributions of this thesis.
- **Chapter 2** summarizes the Bayesian inference and forecasting methods, and their applications in public transit systems.
- **Chapter 3** presents a Bayesian Gaussian model to estimate the link travel time correlation matrix of a bus route using smart-card-like data. This method can overcome the small-sample-size problem in correlation matrix estimation by borrowing those incomplete observations from other bus routes. This chapter shows that link travel times of a bus route have both local and long-range correlations.
- **Chapter 4** proposes a Bayesian Gaussian mixture model for probabilistic forecasting of bus travel time and estimated time of arrival. This chapter shows that modeling the interaction between adjacent buses can significantly improve forecasting performance.
- **Chapter 5** presents a Bayesian Markov regime-switching vector autoregressive model to jointly forecast both bus travel time and passenger occupancy with uncertainty. This approach can capture the intricate interactions among adjacent buses and adapts to the multimodality and skewness of real-world bus travel time and passenger occupancy observations. This chapter shows that the joint forecasting of passenger occupancy and bus travel time can achieve better performance.
- **Chapter 6** proposes a temporal Bayesian model to estimate transit OD matrices at the individual bus level using counts of boarding and alighting passengers at each stop. This chapter shows that the proposed time-varying model outperforms the

static estimation model and can provide uncertainty quantification associated with estimation.

- **Chapter** 7 summarizes the thesis with final conclusions and future directions.

Chapter 2

Literature Review and Preliminary

This thesis aims to develop Bayesian inference and forecasting methods for the research problems outlined in Section 1.3. Detailed reviews and discussions related to these research problems will be provided in subsequent chapters. The purpose of this chapter is to summarize the methodological developments in Bayesian inference and forecasting methods, and their applications in public transit systems. In this chapter, we will review uncertainty quantification, Bayesian inference methods, and Bayesian forecasting methods.

2.1 Uncertainty Quantification

Uncertainty quantification (UQ) is an interdisciplinary field focused on the systematic assessment and management of uncertainty in computational models (Sullivan, 2015). UQ aims to identify the various sources of uncertainty, characterize them mathematically, and evaluate their impact on model predictions. It plays an important role in ensuring that models and predictions are reliable and robust, especially in complex systems where variability and incomplete knowledge are inherent (Abdar et al., 2021). UQ helps researchers and practitioners understand the limits of their models, identify potential uncertainty, and make better-informed decisions.

2.1.1 Aleatory and Epistemic Uncertainties

Two primary categories of uncertainty in research problems are *aleatory* and *epistemic* uncertainties, each with distinct characteristics and implications for modeling and decision-making (Soize, 2017).

Aleatory Uncertainty

Aleatory uncertainty, also known as stochastic or inherent uncertainty, arises from the inherent variability and randomness in natural systems and processes. This type of uncertainty is irreducible, meaning that no matter how much additional data we gather or how sophisticated our models become, the underlying randomness remains. In transportation systems, aleatory uncertainty arises from the natural fluctuations and unpredictable events that affect transportation operations and outcomes (Li et al., 2020). Here are some examples of aleatory uncertainty in public transit systems:

- Adverse weather, such as rain and snow, can cause sudden changes in road conditions, leading to slower traffic speeds and higher travel times (Lam et al., 2008). The occurrence and intensity of rain or snow are inherently random and can vary widely.
- Traffic accidents are random events that can significantly disrupt traffic flow (Bao et al., 2020). The timing, location, and severity of accidents are unpredictable, leading to aleatory uncertainty in transit management and operation.

While aleatory uncertainty cannot be eliminated, understanding and managing this type of uncertainty is crucial for maintaining the efficiency and reliability of public transit operations. By employing robust design, real-time monitoring, and adaptive management strategies, public transportation planners and operators can better deal with the unpredictable nature of aleatory uncertainties and enhance the overall resilience of transit systems (Ibarra-Rojas et al., 2015).

Epistemic Uncertainty

Epistemic uncertainty, also known as systematic or reducible uncertainty, stems from a lack of knowledge or information about the system or process being modeled. This type of uncertainty can, in principle, be reduced or eliminated through additional research, better data collection, improved measurement techniques, or enhanced modeling methods. Epistemic uncertainty in public transit systems arises from a lack of knowledge or incomplete information about the system (Li et al., 2020). Here are some examples of epistemic uncertainty in public transit systems:

- Surveys conducted to understand travel behavior might suffer from low response rates or biased samples, leading to incomplete data. This uncertainty affects the accuracy of models that predict route choices and mode preferences.

- Models predicting future transit demand might not fully capture the impact of emerging trends such as telecommuting and ride-sharing services. This limitation can lead to uncertainties in long-term transit demand forecasting.

Epistemic uncertainty is often addressed through Bayesian methods, which allow for the updating of probability distributions as new data/knowledge becomes available. Sensitivity analysis is also used to identify which parameters or assumptions contribute most to the uncertainty, guiding efforts to reduce these uncertainties (Sullivan, 2015).

2.1.2 Components of Uncertainty Quantification

For a transit system, we often pay more attention to understanding epistemic uncertainty and aim to provide robust forecasting and operation decisions. In general, the components of uncertainty quantification are categorized into three types: parameter uncertainty, model uncertainty, and data uncertainty (Soize, 2017).

Parameter Uncertainty

This type of uncertainty arises from a lack of precise knowledge about the values of the model parameters. For instance, in a bus demand prediction model, parameters such as the arrival rate of passengers might be known only approximately. To manage parameter uncertainty, UQ techniques often involve treating these parameters as random variables characterized by probability distributions. Bayesian inference methods are particularly useful here, allowing for the updating of these distributions as new data becomes available, thus refining the model's predictions (Kennedy and O'Hagan, 2001).

Model Uncertainty

Different models or modeling approaches could yield varying results when applied to the same problem (Soize, 2017). This difference is due to simplifications, assumptions, and inherent limitations within the models. UQ addresses model uncertainty by comparing multiple models and often using Bayesian methods to integrate the predictions from different models. This approach provides a more comprehensive prediction by considering the strengths and weaknesses of each model, thereby offering a more robust understanding of the uncertainty involved.

Data Uncertainty

Observational data, which is used to calibrate and validate models, is often noisy, incomplete, or subject to measurement errors. This data uncertainty can significantly affect model estimation and predictions (Chatfield, 1995). UQ incorporates this type of uncertainty by using likelihood functions that describe the probability of observing the data given the model parameters. This probabilistic approach ensures that the uncertainty inherent in the data is reflected in the model outputs, leading to more credible and reliable predictions.

2.1.3 Probabilistic Uncertainty Quantification Scores

Several metrics are commonly used to assess the quality of the probabilistic estimations/predictions, including the Continuous Ranked Probability Score (CRPS) (Gneiting and Raftery, 2007), the logarithmic score (LogS) (Jordan et al., 2017), the interval score (INT) (Gneiting and Raftery, 2007), and coverage (CVG) (Heaton et al., 2019). Here, we introduce CRPS and LogS, which are commonly used in the evaluation of probabilistic models.

Continuous Ranked Probability Score (CRPS)

CRPS is a measure used to evaluate the accuracy of probabilistic forecasts/estimates. It compares the entire predictive distribution to the actual outcome, providing a single score that reflects both the sharpness and the reliability of the forecast (Gneiting and Raftery, 2007). CRPS is defined for a single observation as the integral of the squared difference between the cumulative distribution function (CDF) of the forecast and the CDF of the observed value. Let X be a random variable, F be the CDF of X (i.e., $F(t) = p(X \leqslant t)$), and x be the observation, the CRPS between x and F is given by:

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \left[F(t) - \mathbb{1}(t \geqslant x)\right]^2 dy, \tag{2.1}$$

where $\mathbb{1}(t \geqslant x)$ is the indicator function that equals 1 if $t \geqslant x$ and 0 otherwise. A lower CRPS value indicates a better probabilistic forecast or estimate. CRPS takes into account both the distance between the predicted and observed values and the spread of the predictive distribution. It rewards predictions/estimates that are both sharp (narrow predictive intervals) and well-calibrated (accurate coverage of the observed values).

Logarithmic Score (LogS)

The LogS, also known as the log score or negative log-likelihood, assesses the quality of probabilistic predictions by evaluating how likely the observed outcomes are under the predicted probability distribution (Jordan et al., 2017). LogS is defined as the logarithm of the probability assigned to the observed value by the predictive distribution. It penalizes forecasts/estimates that assign low probabilities to the observed values. For a predictive probability density function (PDF) f and an observed value x, the LogS is given by:

$$\mathrm{LogS}\left(f, x\right) = -\log f\left(x\right). \tag{2.2}$$

A lower LogS value indicates a better probabilistic forecast/estimate. LogS directly evaluates the likelihood assigned to the observed value by the model, with lower scores corresponding to higher assigned probabilities.

2.2 Bayesian Inference Methods

Bayesian inference is a powerful statistical approach that provides a framework for inferring the probability distributions of parameters using available data and prior knowledge (Gelman et al., 2013). It combines prior beliefs with available data to form a posterior distribution, offering a coherent method for estimation and uncertainty quantification.

2.2.1 Bayes' theorem

The core of Bayesian inference lies in Bayes' theorem, which is expressed as:

$$p\left(\theta \mid \mathcal{D}\right) = \frac{p\left(\mathcal{D} \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{D}\right)}, \tag{2.3}$$

where $p\left(\theta \mid \mathcal{D}\right)$ is the posterior distribution of the parameters θ given the data $\mathcal{D}$; $p\left(\mathcal{D} \mid \theta\right)$ is the likelihood of the data given the parameters; $p\left(\theta\right)$ is the prior distribution of the parameters; $p\left(\mathcal{D}\right)$ is the marginal likelihood or evidence, which ensures the posterior distribution sums to one.

Prior Distribution

The prior distribution $p\left(\theta\right)$ represents the initial beliefs about the parameters before observing the data. Priors can be informative, incorporating expert knowledge, or noninformative, reflecting a lack of prior knowledge.

Likelihood

The likelihood $p\left(\mathcal{D} \mid \theta\right)$ represents the probability of observing the data given the parameters. It encapsulates the relationship between parameters and the observed data.

Posterior Distribution

The posterior distribution $p\left(\theta \mid \mathcal{D}\right)$ combines the prior distribution and the likelihood, updating our beliefs about the parameters in light of the data. This distribution is the cornerstone of Bayesian inference, providing a complete probabilistic description of the parameters after considering the evidence.

2.2.2 Markov Chain Monte Carlo (MCMC)

Bayesian inference often involves complex posterior distributions that are not analytically tractable (Gelman et al., 2013). Therefore, computational methods are essential for approximating these distributions. Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings algorithm and Gibbs sampling, are widely used to approximate the posterior distribution by generating samples from it. MCMC methods originated in the field of physics, specifically in statistical mechanics (Metropolis and Ulam, 1949; Metropolis et al., 1953), and it was only towards the end of the 1980s that they started to have a significant impact in the field of statistics (Bishop, 2006).

MCMC methods are utilized to approximate complex posterior distributions, such as $p\left(\theta \mid \mathcal{D}\right)$, when deriving these explicitly is challenging or infeasible. These methods involve sequentially drawing samples from a series of related distributions, with each new sample adjusted based on the one preceding it. This sequential approach ensures that the constructed Markov chain's stationary distribution closely aligns with the target distribution (Gelman et al., 2013). The term "Markov" in MCMC highlights that each iteration's parameter value, $\theta^{(k)}$, depends solely on the value from the previous iteration, $\theta^{(k-1)}$. The sampling process at each iteration involves a specific transition distribution to move towards the target distribution. Here we introduce several important MCMC methods including the Metropolis-Hastings algorithm, the Gibbs sampling, slice sampling, and elliptical slice sampling.

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm (Hastings, 1970) is a generalization of the Metropolis algorithm (Metropolis et al., 1953). It constructs a Markov chain by proposing new states based on a proposal distribution and accepting or rejecting these states based on an acceptance probability. In the k-th sampling iteration, the Markov states of θ is updated through:

- Propose a new state θ' from a proposal distribution $q\left(\theta' \mid \theta^{(k-1)}\right)$.
- Calculate the acceptance probability α as:

$$\alpha = \min\left(1, \frac{p\left(\theta' \mid \mathcal{D}\right) q\left(\theta^{(k-1)} \mid \theta'\right)}{p\left(\theta^{(k-1)} \mid \mathcal{D}\right) q\left(\theta' \mid \theta^{(k-1)}\right)}\right). \tag{2.4}$$

- Accept the new state with probability α. If accepted, set $\theta^{(k)} = \theta'$; if rejected, set $\theta^{(k)} = \theta^{(k-1)}$.

A notable challenge with the Metropolis-Hastings method is managing the acceptance rate, which can be notably low for multidimensional data or parameters. This issue is often exacerbated by difficulties related to the choice of step size in the proposal distribution. In the Metropolis-Hastings algorithm, each proposed move must be accepted or rejected based on the acceptance probability. When dealing with high-dimensional parameters, the algorithm may struggle to efficiently explore the parameter space. To address the challenges, several variants of Metropolis-Hasting have been proposed, including the random walk Metropolis (Gelman et al., 1997), reversible-jump algorithms (Green, 1995; Richardson and Green, 1997), and delayed-rejection (Green and Mira, 2001). On the other hand, the step size determines how far a proposed move is likely to be from the current position, and choosing an appropriate step size can be tricky: If the step size is too small, the algorithm will explore the parameter space very slowly, making many incremental moves that are likely to be accepted but do not cover much ground. This leads to slow convergence and can require many iterations to adequately explore the distribution. Conversely, if the step size is too large, the algorithm may frequently propose moves to low-probability areas, resulting in a high rejection rate. This also impedes efficient exploration of the parameter space, as the chain can become stuck or only make occasional large jumps (Bishop, 2006). To deal with this challenge, the technique of slice sampling (Neal, 2003) provides an adaptive step size that is automatically adjusted to match the distribution, which will be introduced later in this section.

Gibbs Sampling

Gibbs sampling (Geman and Geman, 1984; Tanner and Wong, 1987; Gelfand and Smith, 1990) is a special case of the Metropolis-Hastings algorithm where the proposal distribution is derived from the full conditional distributions of each parameter. This method simplifies the sampling process by exploiting the structure of the conditional distributions. Gibbs sampling iteratively updates each variable in turn, conditioning on the current values of all other variables. Suppose we have the parameters $(\theta_1, \theta_2, \ldots, \theta_n)$. For each iteration, update each parameter θ_i in sequence, sampling from the conditional distribution of θ_i given all other parameters:

$$\theta_i^{(k)} \sim p\left(\theta_i \mid \theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{i-1}^{(k)}, \theta_{i+1}^{(k-1)}, \ldots, \theta_n^{(k-1)}\right). \tag{2.5}$$

This step is repeated for each parameter in the model, cycling through all the parameters systematically. After many iterations, the distribution of samples converges to the joint distribution of the parameters. The initial set of samples (known as the "burn-in period") is usually discarded to ensure that the samples come from the targeted distribution. Gibbs sampling is often considered the simplest MCMC method and is typically recommended as the first option for models that are conditionally conjugate (Bishop, 2006). In conditionally conjugate models, the parameters can be directly sampled from their conditional posterior distributions, which simplifies the computation significantly.

Slice Sampling

Slice sampling (Neal, 2003) can address the challenge that the Metropolis-Hastings algorithm is sensitive to step size. Slice sampling operates by defining a region or "slice" where the probability density of the target distribution is above a certain threshold. It then samples uniformly from this region. This method can effectively explore the target distribution without the need for a finely tuned proposal distribution. For the parameter θ, the basic steps of slice sampling are as follows:

- The slice is defined for a current parameter value θ by $u < p(\theta)$, where u is is a vertical level uniformly chosen below the curve $p(\theta)$ (the target density function). This defines a horizontal 'slice': $S = \{\theta : u < p(\theta)\}$.

- Find an interval around θ that contains a significant portion of the slice and sample θ' uniformly from this interval. If the new point θ' falls within the slice $(u < p(\theta'))$, it is accepted as the new sample. If not, the interval is shrunk, and the process is repeated until a valid sample is drawn.

Elliptical Slice Sampling

Elliptical Slice Sampling (Murray et al., 2010) is a variant of the slice sampling method designed specifically to sample from distributions that can be expressed as a Gaussian "prior" multiplied by a likelihood that is expensive or difficult to compute. This method is particularly useful for Bayesian inference problems in which the likelihood can be computed for a given parameter vector, but where the parameter vector itself is correlated and is modeled using a Gaussian process (Williams and Rasmussen, 2006). Elliptical Slice Sampling combines the benefits of slice sampling with the properties of elliptical distributions (i.e., multivariate Gaussian distributions) to efficiently explore parameter spaces that have strong correlations. For the parameter vector $\boldsymbol{\theta}$, the process of elliptical slice sampling is as follows:

- Draw an auxiliary variable $\boldsymbol{\beta}$ from the Gaussian prior.
- Choose a slice level u similar to slice sampling for the current parameter $\boldsymbol{\theta}$, $u \sim \mathcal{U}\left(0, p(\boldsymbol{\theta})\right)$, where $p(\boldsymbol{\theta})$ is the likelihood function times the prior evaluated at $\boldsymbol{\theta}$.
- Construct an ellipse using $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ such that any point on the ellipse can be written as:
$$\boldsymbol{\theta}\left(\phi\right) = \boldsymbol{\theta}\cos\left(\phi\right) + \boldsymbol{\beta}\sin\left(\phi\right), \tag{2.6}$$
where ϕ is the angel parameter. Randomly select an angle ϕ and define a bracket of angles around this angle. This bracket will shrink in subsequent steps.
- If the new point $\boldsymbol{\theta}(\phi)$ does not satisfy $p\left(\boldsymbol{\theta}\left(\phi\right)\right) > u$, adjust the bracket to exclude ϕ and sample a new ϕ from the remaining bracket. This step is repeated until a satisfactory $\boldsymbol{\theta}\left(\phi\right)$ is found.

This sampling method can significantly simplify the computation and improve the efficiency of the sampling process in complex Bayesian inference tasks involving correlated parameters and computationally intensive likelihood evaluations.

In real-world applications, it is often the case that models incorporate a mix of parameter types–some with conditional posterior distributions that are easily sampled directly and others that are not. To address these varied sampling requirements efficiently, combining Gibbs sampling with other methods like Metropolis-Hastings, slice sampling, and elliptical slice sampling provides a powerful framework for handling complex Bayesian models, enhancing both the efficiency and effectiveness of the sampling process (Bishop, 2006).

2.2.3 Applications in Public Transit

Bayesian inference has been applied to various research problems in public transit systems and this section will review the related previous literature. There are some existing works focusing on using Bayesian inference to estimate transit OD matrix based on counts of the passengers boarding and alighting at each stop. Li (2009) applied the Markov chain model to capture the relationships between the entries of the transit route OD matrix, and to reduce the total number of unknown parameters. Bayesian inference was performed to estimate the unknown parameters of the Markov model and this method derived a closed-form solution, which is computationally efficient. Hazelton (2010) introduced a novel Bayesian model for OD matrix estimation and developed a two-stage sampling algorithm for Bayesian inference using the MCMC method. The first stage samples latent OD matrices using the Markov model by Li (2009) as the proposal distribution. The second stage samples model parameters conditional on the OD matrices in the first stage. Blume et al. (2022) developed a Bayesian inference approach to estimate the static OD matrix in large-scale networked transit systems but considering elements as continuous random variables. This problem is approached as an inverse linear regression, and the posterior distributions of OD matrix entries are estimated using Hamiltonian Monte Carlo. Another important topic is using Bayesian inference for transit passenger assignment problems. Sun et al. (2015) proposed an integrated Bayesian statistical inference framework to characterize the transit assignment model. To estimate the high-dimensional parameters, they developed the variable-at-a-time Metropolis sampling algorithm to make Bayesian inference. Rahbar et al. (2018) proposed a Bayesian hierarchical model to estimate travel time components and to calibrate a transit assignment model. Route choices are represented by a multinomial logit model, and the parameters are estimated through the MCMC method. Besides the above research problems, Zhu et al. (2018) proposed a model for inferring the probability distribution of the number of times a passenger is left behind at stations in congested metro systems. They applied Bayesian inference methods to estimate the left-behind probability for a given station and period. Sun et al. (2021) proposed a Bayesian framework to infer the passenger demand profile conditional on the observed bus dwell times. They applied Hamiltonian Monte Carlo sampling to approximate the posterior distribution for the model parameters. Cheng et al. (2021) proposed a Bayesian topic model to infer trip destination in smart card data with only tap-in records. They applied Gibbs sampling to approximate the posterior distributions of model parameters.

2.3 Bayesian Forecasting Methods

2.3.1 Bayesian Forecasting

Bayesian forecasting involves using Bayesian methods to predict future data points based on existing data (Gelman et al., 2013). It incorporates prior distributions, likelihoods, and posterior distributions to make predictions. In forecasting problems, train data and test data play crucial roles: the train data is used to fit the Bayesian model and learn the parameters, while the test data is used to evaluate the model's predictive performance.

Bayesian Forecasting Process

- Model specification: Develop a probabilistic model that describes the relationship between the train data $\mathcal{D}$ and parameters θ. Specify the likelihood function $p(\mathcal{D} \mid \theta)$, which describes the probability of the observed data given the model parameters. Common forecasting models include linear regression, time series models, and state-space models.

- Prior specification: Specify prior distribution $p(\theta)$ for the model parameters. The prior represents prior beliefs about the parameters before observing the data.

- Bayesian inference: Use Bayes' theorem and Bayesian inference methods to update the prior distribution with the train data to obtain the posterior distribution $p(\theta \mid \mathcal{D})$ of the model parameters using Eq. (2.3).

- Bayesian forecasting: Integrate over the posterior distribution of the model parameters to obtain the predictive distribution of future observations. The posterior predictive distribution of a future data point y^* can be expressed as:

$$p\left(y^* \mid \mathcal{D}\right) = \int p\left(y^* \mid \theta\right) p\left(\theta \mid \mathcal{D}\right) d\theta. \tag{2.7}$$

Common Bayesian Forecasting Models

- Bayesian linear regression: Bayesian linear regression (Mitchell and Beauchamp, 1988) is a statistical method that combines linear regression with Bayesian inference, which can be applied to make probabilistic forecasting. It provides a probabilistic approach to linear regression, incorporating prior knowledge and updating this knowledge with observed data.

- Bayesian vector autoregression (VAR): Bayesian VAR (Bańbura et al., 2010) is a powerful extension of the traditional VAR model for probabilistic forecasting. It incorporates Bayesian inference to improve the estimation of the model parameters, especially when dealing with small sample sizes or highly parameterized models.
- Bayesian state-space model: Bayesian state-space models (Triantafyllopoulos et al., 2021), such as hidden Markov model (Scott, 2002) and linear dynamic systems (Linderman et al., 2017), are powerful tools for modeling time series data, especially when dealing with hidden or latent variables that evolve over time. These models combine state-space representation with Bayesian inference, allowing for robust estimation and prediction while accounting for uncertainty in model parameters and state variables.
- Bayesian neural network (BNN): BNN (Kononenko, 1989) incorporates Bayesian inference to estimate the distribution of the network's parameters. Unlike traditional neural networks, which provide point estimates of the parameters, BNNs provide a distribution over the parameters, allowing for uncertainty quantification in predictions.

Bayesian framework is highly flexible, allowing for the integration of Bayesian inference with various forecasting models to produce probabilistic forecasts. By treating model parameters as random variables with associated probability distributions, the Bayesian approach provides a comprehensive measure of uncertainty in forecasts. This probabilistic perspective not only enhances the interpretability of the predictions but also improves decision-making processes by quantifying the confidence in different outcomes.

2.3.2 Applications in Public Transit

Bayesian forecasting methods have been effectively applied in public transit systems to enhance various aspects of their operations. They can provide probabilistic predictions that account for uncertainty and improve decision-making. Transit demand prediction is a very important research problem, which is fundamental and critical to public transit planning and management. Li et al. (2020) proposed a Bayesian graph convolution model to provide probabilistic forecasting of transit OD demand. Roos et al. (2017) proposed a dynamic Bayesian network approach for short-term passenger flow forecasting. Zhao et al. (2018) developed a method for predicting daily individual mobility represented as a chain of trips and they proposed a Bayesian n-gram model to predict trip attributes. There are some existing works using Bayesian forecasting methods for transit travel time prediction. Ma et al. (2017) proposed a generalized Markov chain approach for estimating the probability distribution of bus trip travel times. Huang et al. (2021) proposed a Bayesian Support Vector Regression to forecast the distribution of bus travel time. Büchel and Corman (2022b) proposed a hidden Markov chain framework to forecast bus travel time distribution. More reviews and discussions on Bayesian forecasting of transit travel time are detailed in Chapter 4.3. Overall, probabilistic/Bayesian forecasting for transit systems has received significant attention in recent years due to its potential to enhance the efficiency and reliability of public transportation.

Chapter 3

Bayesian Inference for Link Travel Time Correlation

This chapter is a research article submitted to *Transportmetrica B: Transport Dynamics*:

- **Chen, X.**, Cheng, Z., Sun, L., 2022. Bayesian inference for link travel time correlation of a bus route. arXiv preprint arXiv:2202.09485.

This chapter corresponds to the Bayesian inference method for the link travel time correlation of a bus route. The understanding of the link travel time correlation is important for bus travel time forecasting in the subsequent chapters.

3.1 Abstract

Estimation of link travel time correlation of a bus route is essential to many bus operation applications, such as timetable scheduling, travel time forecasting and transit service assessment/improvement. Most previous studies rely on either independent assumptions or simplified local spatial correlation structures. In the real world, however, link travel time on a bus route could exhibit complex correlation structures, such as long-range correlations (e.g., a delayed bus is more likely to be further delayed due to bus bunching), negative correlations (e.g., a bus that goes ahead of schedule may intentionally slow down to follow a pre-defined timetable), and time-varying correlations (e.g., different correlation patterns for peak and off-peak hours). Therefore, before introducing strong assumptions, it is essential to empirically quantify and examine the correlation structure of link travel time from real-world bus operation data. To this end, this paper develops a Bayesian Gaussian model to estimate the link travel time correlation matrix of a bus route using smart-card-like data. Our method overcomes the small-sample-size problem in correlation matrix estimation by borrowing/integrating those incomplete observations (i.e., with missing/ragged values and overlapped link segments) from other bus routes. Next, we propose an efficient Gibbs sampling framework to marginalize over the missing and ragged values and obtain the posterior distribution of the correlation matrix. Three numerical experiments are conducted to evaluate model performance. We first conduct a synthetic experiment and our results show that the proposed method produces an accurate estimation for travel time correlations with credible intervals. Next, we perform experiments on a real-world bus route with in-out-stop record data; our results show that both local and long-range correlations exist on this bus route. Finally, we demonstrate an application of using the estimated covariance matrix to make probabilistic forecasting of link and trip travel time.

3.2 Introduction

Understanding travel time characteristics of buses is not only vital in providing better services for passengers (e.g., better travel time estimation), but also essential for transit agencies to design efficient and economical operation strategies (e.g., better route and timetable optimization) (Liao et al., 2020). A bus route can be viewed as a directed chain network, where each node represents a bus stop and each link represents the road section between two adjacent bus stops. Link travel time correlation of a bus route is essential to understanding the characteristics of the bus route and improving bus travel time estimation (Dai et al., 2019).

However, most existing studies on estimating link travel time and the corresponding correlation structure mainly centers on passenger car traffic, while such analysis is often inappropriate for bus systems due to the unique operational properties of bus services. A major limitation that prevents us from using existing link travel time analysis models is that an accurate estimation of link travel time correlation matrices requires a large number of complete observations, while buses are essentially sparse in general road traffic and incomplete trip observations are pretty normal from bus operations, particularly for a long route with many links. As a result, the scale of many studies is confined to only a few links (e.g., Gajewski and Rilett, 2005). To better utilize the limited data, another approach is to use simplified/parsimonious correlation structure to model link travel time. For instance, many statistical models assume only the travel times of adjacent/near links are correlated (e.g., Chen et al., 2012; Jenelius and Koutsopoulos, 2013; Srinivasan et al., 2014). Although this is an intuitively reasonable assumption, it is inappropriate for bus travel time since link travel time correlations of a bus route have much more complex spatiotemporal characteristics. First, the link travel time of a bus route may have long-range correlations due to factors such as bus bunching (e.g., a delayed bus tends to be further delayed). Second, the correlation might be negative; for example, a bus that goes ahead of schedule may intentionally slow down to follow a pre-defined timetable. Moreover, link travel time correlations vary in different periods; the correlations of peak hours and the off-peak period could be completely different due to the time-varying service frequency and road traffic. This further limits the available sample size to conduct link travel time analysis for a bus route over a pre-defined time window. In summary, it is difficult to adapt existing link travel time analysis for car traffic to bus operation due to (1) small sample size: the limited number of complete observations are usually insufficient to estimate the link travel time correlation accurately; (2) oversimplified assumptions on the correlation structure: assuming only local spatial correlation is insufficient to capture the complex characteristics of actual link travel time for bus operation.

If having access to a large amount of bus operation data (e.g., automatic vehicle location data, smart card data), we can infer the arrival time of a bus at a bus stop. Then, we can build a vector of link travel time using such in-out-stop records for each service run (from the first stop to the last stop), and then estimate the mean and covariance from samples for multiple service runs in a similar way as for car traffic. However, such data in practice is often not readily available due to the following issues. First, the arrival time at a stop becomes inaccessible when there is no boarding/alighting passenger or when the stop is skipped by a bus, bringing many unknown values in arrival time (and thus link travel time). A unique property in bus operation is that, for the stop-skipping case we still can obtain the sum of travel time of several adjacent links from the arrival time of upstream and downstream stops; we refer to this special type of missing values as *ragged values*. Ragged values are quite common in bus systems and contain valuable information for enhancing link travel time correlation estimation. Secondly, buses are essentially sparse in traffic, and the number of operational buses per bus route per day is usually very small. For instance, a high-frequency bus route with a 10-min headway will only generate a sample size of 6/hour (if fully observed), which is much smaller than that of general car traffic. The small sample size and the ragged pattern prevent us from having a robust estimation of link travel time correlation, especially when quantifying link travel time correlations of a specific period (such as morning peak).

To address the above issues, in this paper we develop a Bayesian probabilistic model to estimate the link travel time correlations in a bus route. In particular, we aim to address the missing/ragged value problem and limited sample size problem for a target bus route by incorporating data from other bus routes that have overlapped links/stops. We assume the travel time of links in a bus route follows a multivariate Gaussian distribution. The task is to estimate the covariance matrix and the Bayesian credible interval of each entry in the matrix (correlation matrix can be obtained from the covariance matrix). In particular, our method makes use of incomplete observations with missing, ragged values and route segments from multiple bus routes. We point out that the conditional distribution of missing and ragged values can be viewed as a multivariate Gaussian distribution truncated on the intersection with a hyperplane. Next, we develop an efficient Markov chain Monte Carlo (MCMC) sampling algorithm to marginalize over the missing and ragged values and obtain the posterior distribution of the covariance matrix. In a test with synthetic data, we found our method produces accurate estimation for link travel time covariance. The MCMC scheme also allows us to exploit the posterior distribution of each entry in the covariance/correlation matrix. In addition, the incorporation of incomplete data substantially improves the estimation. Moreover, we use our model to empirically quantify the link travel time correlations of a twenty-link bus route in Guangzhou, China; results reveal strong local and long-range correlation patterns in link travel time of the bus route. Finally, we demonstrate an example of probabilistic forecasting of link/trip travel time in a bus route using the estimated covariance matrix; our forecasting method is more accurate than the historical average.

The contribution of this paper is twofold. First, we propose a Bayesian model that can use incomplete/corrupted vectors of link travel time from multiple bus routes to estimate the link travel time correlations of a target bus route. This model overcomes the small sample size of a single bus route by integrating incomplete data that are unusable in other models. A Gibbs sampling algorithm is developed to obtain the posterior distribution of the covariance/correlation matrix. Second, we verify the robustness and applicability of the proposed model by a synthetic example and a real-world case study. Results show our model can accurately estimate the link travel time correlations with incomplete observations, and the model applies to problems at a practical scale. The estimated correlations are beneficial to system understanding/evaluation and bus travel time estimation/forecasting.

The remainder of this paper is organized as follows. In Section 3.3, we review previous studies on link travel time correlation. In Section 3.4, we describe the problem of the link travel time correlation estimation in a bus route and introduce notations. In Section 3.5, we present the Bayesian probabilistic model and the inference method based on MCMC. Next, in Section 3.6, we demonstrate the capability of our model through three experiments. Finally, we conclude our study, summarize our main findings, and discuss future research directions in Section 3.7.

3.3 Related Work

Most previous studies have concentrated on link travel time correlation or covariance matrix estimation for car trips. During the 1990s, Advanced Traveler Information System (ATIS) was deployed rapidly, which aims to provide information to assist surface transportation travelers in moving from a starting location (origin) to their desired destination (Schofer et al., 1993). This system collects data from probe automobiles, prompting the emerging research in travel time estimation and forecasting. Sen et al. (1999) pointed out the covariance of link travel times which are close together, may not be zero, and they proposed estimating the correlation matrix of link travel time as an open problem for future research. A straightforward solution is to infer the correlation matrix using asymptotic theory (i.e., correlation formula), which is the traditional estimation of correlation. Bernard et al. (2006) used the straight method to estimate link travel speed correlations, which are similar to link travel time correlations. Nevertheless, Gajewski and Rilett (2005) figured out that the classical estimation method lacks interpretability and is complicated due to the nonparametric nature of the estimator and the covariance between links. Then they adopted a Bayesian approach to estimate link travel time correlation, which had benefits in terms of interpretation and ease of use. The authors only experimented on three links because they could collect many full observations of the three links. However, for one road network with many links, the number of full observations dramatically decreases, and this simple Bayesian model can not estimate the correlation matrix accurately.

Link travel time correlation is essential to the stochastic routing problem, as it helps to consider the reliability of travel time. Many early studies (Cheung, 1998; Miller-Hooks, 2001; Seshadri and Srinivasan, 2010) ignored link travel time correlation because of the low computational efficiency for large networks. Some studies use origin-destination trip data to estimate link travel times and they usually make the independent link travel time assumptions (e.g., Hunter et al., 2009; Zhan et al., 2013, 2016; Sun et al., 2015). Many studies focus on estimating the distribution of route travel time, but their methods do not model link travel time correlations (e.g., Rakha et al., 2006; Jenelius and Koutsopoulos, 2017; Woodard et al., 2017; Huang et al., 2021). Next, correlations between link travel times are explored in Waller and Ziliaskopoulos (2002) and Fan and Nie (2006). Both of them consider local spatial correlations between adjacent links. Many studies followed this assumption that only the adjacent link travel times are correlated. Although the local spatial correlation assumption is strong, it seems reasonable because we would expect the impact of a link on another decreases with the increase of distance, and it becomes a popular choice in the literature (Chen et al., 2012; Srinivasan et al., 2014) partially due to the model simplicity and the lack of empirical evidence. Rachtan et al. (2013) adopted three regression models to estimate the correlation by considering various combinations of variables, including spatial distance, temporal distance, traffic state, and the number of lanes, and they found that the primary factor in correlation is spatial distance. Zeng et al. (2015) also incorporated the spatial correlation of link travel time in finding the reliable path of stochastic networks. Geroliminis and Skabardonis (2006) estimated the variance of route travel time; they used full observations of six links to directly compute the covariance and correlation. Ramezani and Geroliminis (2012) applied Markov chains to estimate the route travel time distribution considering the correlation between successive links. They first established a two-dimensional (2D) diagram with data points representing travel times of two consecutive links; then used a heuristic grid clustering method to cluster the 2D diagram to different spaces (states). With a Markov chain procedure, they can integrate the correlation between states of 2D diagrams for successive links. Jenelius and Koutsopoulos (2013) incorporated the spatial link travel time correlation into travel time estimation for urban road networks; they used a spatial moving average (SMA) structure to model link correlation by assuming that the stochastic component of each link is expressed as an independent term with zero mean and variance, plus a linear combination of the independent terms of the other links. Westgate et al. (2016) proposed the method for estimating large-network travel time distribution; they model travel time at the trip level instead of the link level, but they consider the dependency between links by incorporating explanatory factors like the road class, speed limit, one-way road. Rodriguez et al. (2017) used the multi-output Gaussian Processes to estimate network-wide travel time distribution. They considered the squared exponential (SE) kernel to capture correlations between any pair of time points and they applied Graph/Laplacian kernel to model correlations between two link travel times. Copula functions can describe the dependence between random variables. Chen et al. (2017) and Chen et al. (2019) developed a copula-based approach to model the link travel time correlation. The approach applied a two-dimensional Gaussian copula function to fit the link travel time distribution of two adjacent links. Qin et al. (2020) proposed a pair-copula mixture model for estimating urban arterial travel time distribution, and it can reduce the computational complexity. The copula-based models are limited to expensive computations, especially for many links; thus, they are applied on a few links and only consider the correlation between adjacent links.

In summary, most studies about link travel time correlation are for car travel time estimation. There are only a few studies for bus travel time estimation with link travel time correlation being considered. Uno et al. (2009) estimated the variance of individual path travel time by aggregating the covariance between link travel time; thus, the travel time distribution of one path can be estimated by summing up directly observed multiple links mean travel time with their covariance based on bus probe data. Dai et al. (2019) attempted to estimate the bus path travel time distribution using GPS probe and smart card data. They considered that path travel time distribution could be estimated by statistically summarizing link travel time distributions and dwell-time distributions at bus stops. Therefore, both studies need to obtain the correlation or covariance matrices, but they only compute the correlation matrix directly using many full observations without considering the temporal difference. In fact, the correlation matrix should be time-varying due to the temporal variations in bus operation and road traffic. To the best of our knowledge, little attention was paid to quantifying time-varying link travel time correlation in the literature.

3.4 Problem Description

For a bus route with n links (i.e., $n + 1$ bus stops), we define a bus link as the directional road segment between two adjacent bus stops. For conciseness, we often omit the word "bus" and simply use "route" and "link" in the following. Denote a random variable x_i to be the travel time of the i-th link in the route, whose value is observed by the time difference between the arrival of a bus at the two adjacent bus stops. Next, the travel time of all the links of the route can be represented by multivariate random variable $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^\top$.

This study uses the data from in-out-stop record systems. When a bus arrives at or leaves a bus stop, the system registers vehicle ID, route ID, action type (arrival/departure), together with a timestamp. We can thus calculate the link travel time from the bus in-out-stop data. Each bus has a vector for its link travel times on the route. Besides bus in-out-stop record data, smart card data—a more common type of data—can be equivalently used to obtain the link travel time. However, in practice, the link travel time vector $\boldsymbol{x}$ from a bus run on the target route or on a related route is often incomplete due to several reasons. Figure 3.1 uses a simple bus network to illustrate the issue of missing and ragged values in the data. Route 1 is a bus route of interest with seven links. We assume all buses go through the same bus link (during the same period of the day) have the same link travel time distribution, regardless of which route they belong; this allows us to use the data from Route 2 and Route 3 to improve the estimation of link travel time correlations in Route 1. An ideal observation for Route 1 is Figure 3.1 (a), where the travel times of all links are obtained. However, we cannot obtain the travel time of both link #3 and link #4 if a bus of Route 1 skipped stop D; we can instead observe the total travel time (x_3+x_4) from stop C to stop E, and this is the case of ragged value shown in Figure 3.1 (b). Moreover, Route 2 goes through six out of seven links, resulting in observations with missing values as shown in Figure 3.1 (c). The last example is Route 3, which goes through six links (B to H) and has no stop at F, bringing incomplete observations with both missing and ragged values as shown in Figure 3.1 (d). Essentially, data from other relevant bus routes can be considered as a general type of incomplete observations with both missing or ragged values.

We denote $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,n}]^\top$, which is a sample of the random variable $\boldsymbol{x}$, to be the link travel time of the i-th bus during the study period. Since not all entries of $\boldsymbol{x}_i$ are always available from data, we denote $\boldsymbol{r}_i = \mathbf{G}_i \boldsymbol{x}_i \in \mathbb{R}^{n_i}$ to be a vector of observed information attached with $\boldsymbol{x}_i$, where $\mathbf{G}_i \in \{0, 1\}^{n_i \times n}$ is a binary matrix encodes the missing and ragged positions of the i-th bus. We call a $\boldsymbol{r}_i$ a *recording vector* and $\mathbf{G}_i$ an *alignment matrix*. For example, we have $\boldsymbol{x}_i = \boldsymbol{r}_i$ and $\mathbf{G}_i$ being an identity matrix for a bus with a

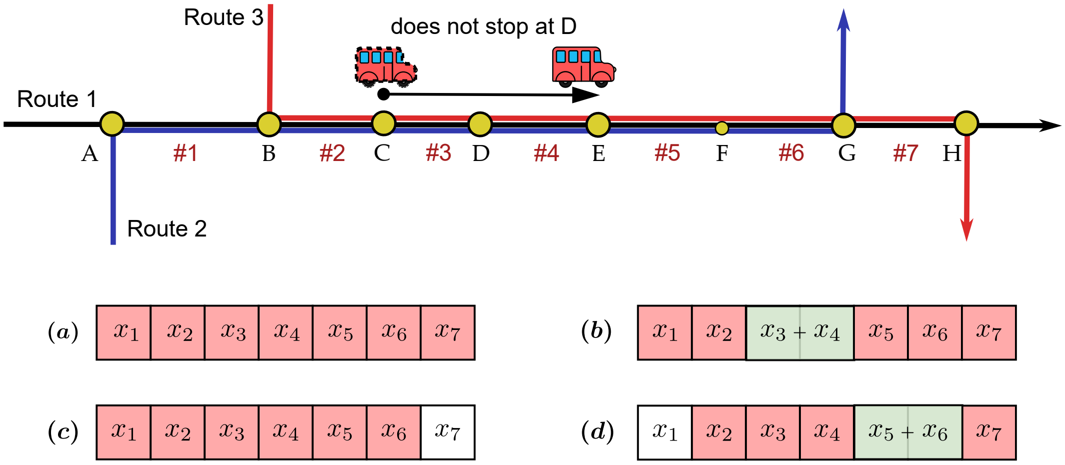

Route 3
does not stop at D
Route 1
A
#1
B
#2
C
#3
D
#4
E
#5
F
#6
G
#7
H
Route 2
(a)
x_1 x_2 x_3 x_4 x_5 x_6 x_7
(b)
x_1 x_2 $x_3 + x_4$ x_5 x_6 x_7
(c)
x_1 x_2 x_3 x_4 x_5 x_6 x_7
(d)
x_1 x_2 x_3 x_4 $x_5 + x_6$ x_7

Figure 3.1: Graphical illustration of full observations and incomplete observations. (a) a complete observation of Route 1; (b) an observation of Route 1 with ragged values; (c) an observation of Route 2, where x_7 is inaccessible (missing); (d) an observation of Route 3, where x_1 is inaccessible (missing), x_5 and x_6 are ragged.

complete observation; for a bus i in the case of Figure 3.1 (d), the link travel time and its missing and ragged values can be represented as

$$\underset{\mathbf{G}_i}{\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}} \underset{\boldsymbol{x}_i}{\begin{bmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \\ x_{i,4} \\ x_{i,5} \\ x_{i,6} \\ x_{i,7} \end{bmatrix}} = \begin{bmatrix} x_{i,2} \\ x_{i,3} \\ x_{i,4} \\ x_{i,5} + x_{i,6} \\ x_{i,7} \end{bmatrix} = \underset{\boldsymbol{r}_i}{\begin{bmatrix} r_{i,1} \\ r_{i,2} \\ r_{i,3} \\ r_{i,4} \\ r_{i,5} \end{bmatrix}}. \tag{3.1}$$

In the real world, the number of incomplete samples can be even greater than that of complete samples. However, such "incomplete" observations should not be discarded as they also encode valuable information in estimating link travel time correlations. Assuming there are m buses that went through a target bus route during a study period, the goal of this research is to incorporate all recording vectors $\{\boldsymbol{r}_i\}_{i=1}^{m}$ and alignment matrices $\{\mathbf{G}_i\}_{i=1}^{m}$ to quantify the link travel time correlation matrix $\mathbf{C} = \operatorname{Corr}[\boldsymbol{x}]$. Note that the correlation matrix could vary for different periods. Therefore, we divide a day into several periods and estimate a correlation matrix for each period separately.

3.5 Methodology

3.5.1 Multivariate Gaussian Model

Gaussian distribution offers numerous analytical and computational advantages and has been extensively used in modeling travel time distribution (e.g., Smeed and Jeffcoate, 1971; Li, 2004; Seshadri and Srinivasan, 2010). In this paper, we assume the joint probability of link travel times in a bus route follows a multivariate Gaussian distribution with the probability density function

$$p\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\left(2\pi\right)^{n/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\top} \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right], \tag{3.2}$$

where $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{x}\right] \in \mathbb{R}^n$ is a mean vector, and $\boldsymbol{\Sigma} = \operatorname{Cov}\left[\boldsymbol{x}\right]$ is an $n \times n$ covariance matrix, and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{\Sigma}$ and its relationship with the correlation matrix $\mathbf{C}$ are shown in Eq. (3.3) and Eq. (3.4), respectively.

$$\boldsymbol{\Sigma} = \mathbb{E}\Big[\left(\boldsymbol{x}-\mathbb{E}\left[\boldsymbol{x}\right]\right)\left(\boldsymbol{x}-\mathbb{E}\left[\boldsymbol{x}\right]\right)^{\top}\Big] = \begin{bmatrix} \mathbb{V}\left[x_1\right] & \operatorname{Cov}\left[x_1, x_2\right] & \cdots & \operatorname{Cov}\left[x_1, x_n\right] \\ \operatorname{Cov}\left[x_2, x_1\right] & \mathbb{V}\left[x_2\right] & \cdots & \operatorname{Cov}\left[x_2, x_n\right] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}\left[x_n, x_1\right] & \operatorname{Cov}\left[x_n, x_2\right] & \cdots & \mathbb{V}\left[x_n\right] \end{bmatrix}, \tag{3.3}$$

$$\mathbf{C} = \left(\operatorname{diag}\left(\boldsymbol{\Sigma}\right)\right)^{-\frac{1}{2}} \boldsymbol{\Sigma} \left(\operatorname{diag}\left(\boldsymbol{\Sigma}\right)\right)^{-\frac{1}{2}} = \begin{bmatrix} 1 & \frac{\operatorname{Cov}[x_1,x_2]}{\sqrt{\mathbb{V}[x_1]\mathbb{V}[x_2]}} & \cdots & \frac{\operatorname{Cov}[x_1,x_n]}{\sqrt{\mathbb{V}[x_1]\mathbb{V}[x_n]}} \\ \frac{\operatorname{Cov}[x_2,x_1]}{\sqrt{\mathbb{V}[x_2]\mathbb{V}[x_1]}} & 1 & \cdots & \frac{\operatorname{Cov}[x_2,x_n]}{\sqrt{\mathbb{V}[x_2]\mathbb{V}[x_n]}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\operatorname{Cov}[x_n,x_1]}{\sqrt{\mathbb{V}[x_n]\mathbb{V}[x_1]}} & \frac{\operatorname{Cov}[x_n,x_1]}{\sqrt{\mathbb{V}[x_n]\mathbb{V}[x_1]}} & \cdots & 1 \end{bmatrix}, \tag{3.4}$$

where $\operatorname{Cov}\left[x_i, x_j\right] = \mathbb{E}\left[\left(x_i - \mathbb{E}\left[x_i\right]\right)\left(x_j - \mathbb{E}\left[x_j\right]\right)\right]$, $\mathbb{V}\left[x_i\right] = \operatorname{Cov}\left[x_i, x_i\right]$, and $\operatorname{diag}\left(\boldsymbol{\Sigma}\right)$ is the diagonal elements of $\boldsymbol{\Sigma}$. Each element in the correlation matrix $\mathbf{C}$ is essentially a Pearson correlation coefficient. However, this naive approach that directly calculates sample variance and covariance using Eq. (3.3) and Eq. (3.4) fails with the presence of incomplete observations.

Figure 3.2 shows the overall graphical representation of our Bayesian model that can handle missing and ragged values. For a collection of m ragged observations $\mathcal{R} = \{\boldsymbol{r}_i\}_{i=1}^{m}$ over a pre-defined time window, we have $\boldsymbol{r}_i = \mathbf{G}_i \boldsymbol{x}_i$ for $i = 1, \ldots, m$. Next, we assume $\boldsymbol{x}_i$ is a "latent" realization/sample from a multivariate Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ following Eq. (3.2). In a Bayesian setting, we further use a conjugate Gaussian-inverse-

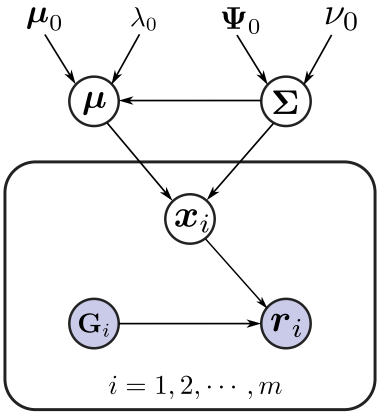


Figure 3.2: The graphical illustration of Bayesian Gaussian model.

Wishart distribution on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Gelman et al., 2013) for efficient inference. The overall data generation process is summarized as:

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi}_0, \nu_0\right), \tag{3.5}$$

$$\boldsymbol{\mu} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda_0}\boldsymbol{\Sigma}\right), \tag{3.6}$$

$$\boldsymbol{x}_i \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \text{ for } i = 1, \ldots, m, \tag{3.7}$$

$$\boldsymbol{r}_i = \mathbf{G}_i \boldsymbol{x}_i, \text{ for } i = 1, \ldots, m, \tag{3.8}$$

where $\mathcal{W}^{-1}\left(\boldsymbol{\Psi}_0, \nu_0\right)$ is the inverse-Wishart distribution with ν_0 degrees of freedom ($\nu_0 \geqslant n$), and an $n \times n$ scale matrix $\boldsymbol{\Psi}_0$; $\boldsymbol{\mu}_0$ is the prior mean. The probability density function of the inverse-Wishart distribution is

$$\mathcal{W}^{-1}\left(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}_0, \nu_0\right) = C\left|\boldsymbol{\Sigma}\right|^{-(\nu_0+n+1)/2} \exp\left[-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1}\right)\right], \tag{3.9}$$

where C is a normalizing constant and $\operatorname{Tr}(\cdot)$ is the trace of a matrix.

Based on the graphical structure presented in Figure 3.2, we can derive an efficient MCMC scheme using Gibbs sampling. For simplicity, we denote by $\mathcal{G} = \{\mathbf{G}_i\}_{i=1}^m$ the set of alignment matrices corresponding to observations $\mathcal{R} = \{\boldsymbol{r}_i\}_{i=1}^m$, by $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^m$ the set of "full" link travel time for the m bus runs, and by $\Theta = \{\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0\}$ the set of hyperparameters for the Gaussian-inverse-Wishart prior distribution in Eqs. (3.5) and (3.6). We start the Gibbs sampling with randomly initialized values for all variables and then iteratively sample each variable from its conditional distribution on other variable:

- Sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from $p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathcal{X}, \Theta\right)$. Because of the conjugate prior distribution, the conditional distribution of the mean vector and the covariance matrix $p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathcal{X}, \Theta\right)$ is a Gaussian-inverse-Wishard distribution:

$$p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathcal{X}, \Theta\right) \sim \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0^*, \frac{1}{\lambda_0^*}\boldsymbol{\Sigma}\right) \mathcal{W}^{-1}\left(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}_0^*, \nu_0^*\right), \tag{3.10}$$

where

$$\begin{aligned} &\boldsymbol{\mu}_0^* = \frac{\lambda_0 \boldsymbol{\mu}_0 + m\bar{\boldsymbol{x}}}{\lambda_0 + m}, \quad \lambda_0^* = \lambda_0 + m, \quad \nu_0^* = \nu_0 + m, \quad \bar{\boldsymbol{x}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i, \\ &\boldsymbol{\Psi}_0^* = \boldsymbol{\Psi}_0 + \mathbf{S} + \frac{\lambda_0 m}{\lambda_0 + m}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0}\right)\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0}\right)^\top, \quad \mathbf{S} = \sum_{i=1}^{m}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^\top. \end{aligned} \tag{3.11}$$

- Sample $\mathcal{X}$ from $p\left(\mathcal{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{R}, \mathcal{G}\right)$. For this step, we no longer have a simple analytical formulation to sample $\mathcal{X}$ due to the linear constraints in Eq. (3.8). We next introduce an effective solution to sample $\boldsymbol{x}_i$ from its conditional distribution in Section 3.5.2.

3.5.2 Sampling Link Travel Time

Assuming the link travel times of different buses are independent. Next, we can factorize the conditional distribution of link travel time as

$$p\left(\mathcal{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{R}, \mathcal{G}\right) = \prod_{i=1}^{m} p\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{r}_i, \mathbf{G}_i\right). \tag{3.12}$$

Therefore, we can draw sample of the bus-specific link travel time vector $\boldsymbol{x}_i$ independently. The conditional distribution of $\boldsymbol{x}_i$ in Eq. (3.12) can be viewed as a multivariate Gaussian distribution truncated on the intersection with a hyperplane, i.e.,

$$\boldsymbol{x}_i \sim \mathcal{N}_{\mathcal{S}_i}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \quad \mathcal{S}_i = \left\{\boldsymbol{x}_i : \mathbf{G}_i \boldsymbol{x}_i = \boldsymbol{r}_i\right\}. \tag{3.13}$$

The probability density function of the hyperplane-truncated multivariate Gaussian is

$$p(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{r}_i, \mathbf{G}_i) = \frac{1}{Z_i} \exp\left[-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right] \delta(\mathbf{G}_i \boldsymbol{x}_i = \boldsymbol{r}_i), \tag{3.14}$$

where Z_i is a normalizing constant; $\delta(*)$ is a function whose value is 1 only if the condition $*$ holds, and 0 otherwise.

There are several available algorithms for efficient sampling over hyperplane-truncated

Gaussian distributions (e.g., Cong et al., 2017; Botev, 2017). We apply a fast sampling algorithm developed by Cong et al. (2017) for this problem. For a given mean vector and covariance matrix, the algorithm for sampling the link travel time vector of the i-th bus is described in Algorithm 1.

Algorithm 1 Sampling from a hyperplane-truncated multivariate Gaussian distribution Cong et al. (2017).

```
1: Sample $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;
2: Return $\boldsymbol{x}_i = \boldsymbol{y} + \boldsymbol{\Sigma}\mathbf{G}_i^\top \left(\mathbf{G}_i\boldsymbol{\Sigma}\mathbf{G}_i^\top\right)^{-1}(\boldsymbol{r}_i - \mathbf{G}_i\boldsymbol{y})$, which can be more efficiently and accurately calculated by
       • Solve $\boldsymbol{\alpha}$ such that $\left(\mathbf{G}_i\boldsymbol{\Sigma}\mathbf{G}_i^\top\right)\boldsymbol{\alpha} = \boldsymbol{r} - \mathbf{G}_i\boldsymbol{y}$;
       • Return $\boldsymbol{x}_i = \boldsymbol{y} + \boldsymbol{\Sigma}\mathbf{G}_i^\top\boldsymbol{\alpha}$.
```

3.5.3 Overall Gibbs Sampling Algorithm

Having obtained the two conditional distributions in Eqs. (3.10) and (3.12), we summarize the overall Gibbs sampling procedure for estimating the correlation matrix in Algorithm 2. We drop the first k_1 iterations as burn-in and estimate the correlation matrix $\hat{\mathbf{C}}$ as the average of samples from the last k_2 iterations. Besides, we store samples of correlation matrices $\{\mathbf{C}^{(i)}\}_{i=1}^{k_2}$ and covariance matrices $\{\boldsymbol{\Sigma}^{(i)}\}_{i=1}^{k_2}$, which are critical ingredients for deriving the Bayesian credible interval for each entry in the correlation matrix and performing probabilistic forecasting of bus travel time. For hyperparameters Θ, we set $\boldsymbol{\mu}_0 = \mathbf{0}_n$, $\lambda_0 = 10$, $\boldsymbol{\Phi}_0 = \boldsymbol{I}_n$, $\nu_0 = n + 2$, where n is number of links.

3.6 Case Study

This section provides three numerical case studies using both synthetic data and real-world data. First, we use a synthetic experiment to test the accuracy of the proposed correlation estimation method and the improvement brought by incorporating missing/ragged values. Next, we apply our model on bus in-out-stop record data to quantify link travel time correlation in a transit corridor in Guangzhou. Finally, we demonstrate an application of using our model in probabilistic forecasting of bus link/trip travel time.

Algorithm 2 Gibbs sampling for correlation estimation.

```
Input: Recording vectors $\mathcal{R}$, alignment matrices $\mathcal{G}$, initial values for hyperparameters $\Theta$,
    the number of iterations $k_1$, $k_2$.
Output: Estimated correlation matrix $\hat{\mathbf{C}}$, a set of samples for correlation matrices $\{\mathbf{C}^{(i)}\}_{i=1}^{k_2}$,
    a set of samples for covariance matrices $\{\boldsymbol{\Sigma}^{(i)}\}_{i=1}^{k_2}$.
for iter $= 1$ to $k_1 + k_2$ do
    Draw $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ according to Eq. (3.5) and Eq. (3.6).
    if iter $> k_1$ then
        Calculate $\mathbf{C}$ by Eq. (3.4), collect $\mathbf{C}$ and $\boldsymbol{\Sigma}$ to the output sets.
    end if
    for $i$ = 1 to $m$ do
        Draw $\boldsymbol{x}_i$ by Algorithm 1.
    end for
    Update the parameters $\Theta = \{\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0\}$ by Eq. (3.11).
end for
Compute $\hat{\mathbf{C}}$ as the average of samples in $\{\mathbf{C}^{(i)}\}_{i=1}^{k_2}$.
return $\hat{\mathbf{C}}$, $\{\mathbf{C}^{(i)}\}_{i=1}^{k_2}$, $\{\boldsymbol{\Sigma}^{(i)}\}_{i=1}^{k_2}$.
```

3.6.1 Case 1: Synthetic Data

We design a simple bus network with 18 links as shown in Figure 3.3 to test the performance of the proposed correlation estimation method. In this bus network, The target Route 1 has 18 links; Route 2 shares 12 links with Route 1 (from link #1 to link #12); Route 3 shares 14 links with Route 1 (from link #5 to link #18). We use a multivariate Gaussian model with pre-defined mean and covariance to synthesize a link-travel-time data set. We set the mean vector to be $\boldsymbol{\mu} = [14, 15, 18, 13, 17, 15, 10, 24, 15, 11, 12, 15, 9, 13, 17, 15, 19, 21]$. We use the Graph kernel to set the covariance matrix by the following steps: (1) The route's structure has local correlations, and we assume the following link pairs are virtually adjacent: (link #4, link #13), (link #5, link #12), (link #7, link #15) to simulate long-range correlations. (2) From the structure, we can obtain the degree matrix $\mathbf{D}$ and the adjacency matrix $\mathbf{A}$, then we can get the symmetrically normalized Laplacian matrix $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\left(\mathbf{D} - \mathbf{A}\right)\mathbf{D}^{-\frac{1}{2}}$. (3) Next, we can compute the kernel matrix $\mathbf{K}$ using the kernel function $\mathbf{K} = \exp\left(\beta\mathbf{L}\right)$. (4) Finally, we can get the correlation matrix $\mathbf{Corr} = \left(\operatorname{diag}\left(\mathbf{K}\right)\right)^{-\frac{1}{2}}\mathbf{K}\left(\operatorname{diag}\left(\mathbf{K}\right)\right)^{-\frac{1}{2}}$, and the covariance matrix $\boldsymbol{\Sigma} = \sigma\mathbf{Corr}$. Here, we set $\beta = 3, \sigma = 10$ to generate the covariance shown in Figure 3.4.

We draw 240 samples of link travel time vectors from the multivariate Gaussian distribution with the above parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Next, we assign 160 samples to Route 1, 80 samples to Route 2, and 80 samples to Route 3. The travel time values for links that do not

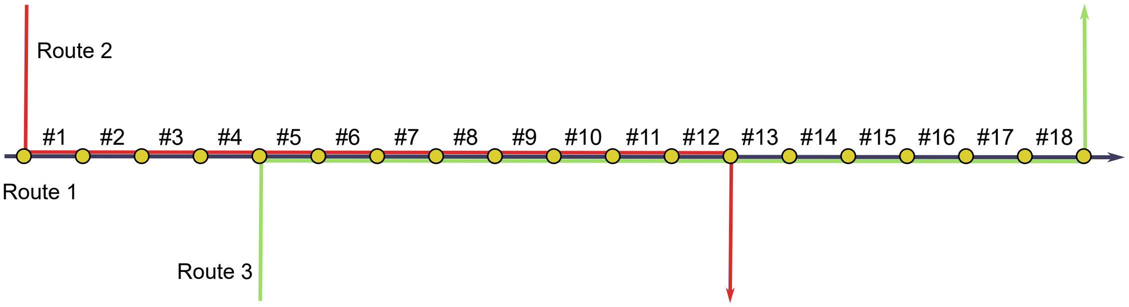
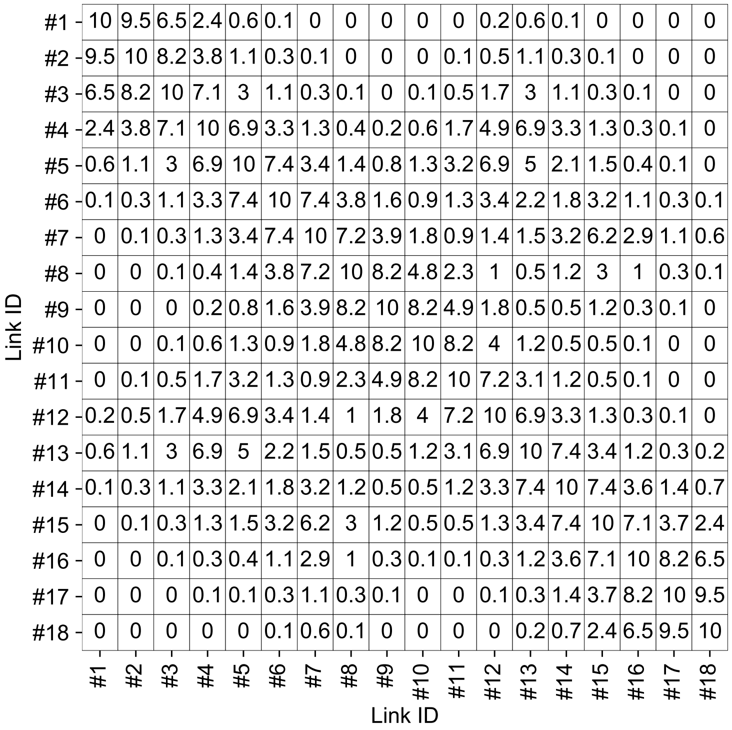


Figure 3.3: The bus network of synthetic data.



Link ID	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
#1	10	9.5	6.5	2.4	0.6	0.1	0	0	0	0	0	0.2	0.6	0.1	0	0	0	0
#2	9.5	10	8.2	3.8	1.1	0.3	0.1	0	0	0	0.1	0.5	1.1	0.3	0.1	0	0	0
#3	6.5	8.2	10	7.1	3	1.1	0.3	0.1	0	0.1	0.5	1.7	3	1.1	0.3	0.1	0	0
#4	2.4	3.8	7.1	10	6.9	3.3	1.3	0.4	0.2	0.6	1.7	4.9	6.9	3.3	1.3	0.3	0.1	0
#5	0.6	1.1	3	6.9	10	7.4	3.4	1.4	0.8	1.3	3.2	6.9	5	2.1	1.5	0.4	0.1	0
#6	0.1	0.3	1.1	3.3	7.4	10	7.4	3.8	1.6	0.9	1.3	3.4	2.2	1.8	3.2	1.1	0.3	0.1
#7	0	0.1	0.3	1.3	3.4	7.4	10	7.2	3.9	1.8	0.9	1.4	1.5	3.2	6.2	2.9	1.1	0.6
#8	0	0	0.1	0.4	1.4	3.8	7.2	10	8.2	4.8	2.3	1	0.5	1.2	3	1	0.3	0.1
#9	0	0	0	0.2	0.8	1.6	3.9	8.2	10	8.2	4.9	1.8	0.5	0.5	1.2	0.3	0.1	0
#10	0	0	0.1	0.6	1.3	0.9	1.8	4.8	8.2	10	8.2	4	1.2	0.5	0.5	0.1	0	0
#11	0	0.1	0.5	1.7	3.2	1.3	0.9	2.3	4.9	8.2	10	7.2	3.1	1.2	0.5	0.1	0	0
#12	0.2	0.5	1.7	4.9	6.9	3.4	1.4	1	1.8	4	7.2	10	6.9	3.3	1.3	0.3	0.1	0
#13	0.6	1.1	3	6.9	5	2.2	1.5	0.5	0.5	1.2	3.1	6.9	10	7.4	3.4	1.2	0.3	0.2
#14	0.1	0.3	1.1	3.3	2.1	1.8	3.2	1.2	0.5	0.5	1.2	3.3	7.4	10	7.4	3.6	1.4	0.7
#15	0	0.1	0.3	1.3	1.5	3.2	6.2	3	1.2	0.5	0.5	1.3	3.4	7.4	10	7.1	3.7	2.4
#16	0	0	0.1	0.3	0.4	1.1	2.9	1	0.3	0.1	0.1	0.3	1.2	3.6	7.1	10	8.2	6.5
#17	0	0	0	0.1	0.1	0.3	1.1	0.3	0.1	0	0	0.1	0.3	1.4	3.7	8.2	10	9.5
#18	0	0	0	0	0	0.1	0.6	0.1	0	0	0	0	0.2	0.7	2.4	6.5	9.5	10

Figure 3.4: Generated covariance matrix **Σ**.

belong to Route 2 or Route 3 are dropped and regarded as missing values. Among the 160 samples for Route 1, we keep 80 samples as full observations, and set ragged values to the rest 80 samples by adding the travel time of link #5 and link #6.

We applied Algorithm 2 to estimate the correlation matrix from the above synthetic data. For convenience, we refer to incomplete observations with missing values (not including ragged values) as "missing observations" and incomplete observations with ragged values as "ragged observation". Three experiments are designed to compare the estimation accuracy with different types of observations: (1) only uses full observations;

(2) uses full and missing observations; (3) uses all observations. The numbers of MCMC iterations are $k_1 = 10000$ and $k_2 = 5000$.

With samples drawn in the last k_2 iterations, the posterior mean of mean vectors estimated with different types of observations are shown in Table 3.1, and the posterior mean of covariance matrices are shown in Figure 3.5.

Next, we use Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) to measure how our estimated distribution $q(\boldsymbol{x}) = \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is different from the true distribution $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Specifically, the KL divergence from $q(\boldsymbol{x})$ to $p(\boldsymbol{x})$ is defined as:

$$D_{KL}\left(p\left(\boldsymbol{x}\right) \parallel q\left(\boldsymbol{x}\right)\right) = \int p\left(\boldsymbol{x}\right) \ln \frac{p\left(\boldsymbol{x}\right)}{q\left(\boldsymbol{x}\right)} d\boldsymbol{x}. \tag{3.15}$$

A smaller KL divergence indicates that the distribution $q(x)$ is closer to the reference distribution $p(x)$. In the experiment, both $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are multivariate Gaussian distributions. We can derive the KL divergence formulation for multivariate Gaussian distributions as:

$$D_{KL}\left(p\left(\boldsymbol{x}\right) \parallel q\left(\boldsymbol{x}\right)\right) = \frac{1}{2}\left[\ln \frac{|\hat{\boldsymbol{\Sigma}}|}{|\boldsymbol{\Sigma}|} - N + \operatorname{Tr}\{\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\top}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right]. \tag{3.16}$$

The KL divergences for the distributions estimated with different observations are shown in Table 3.2. We can see the KL divergence of using all observations is the lowest, indicating that using missing/ragged values can improve the accuracy of the estimated distribution.

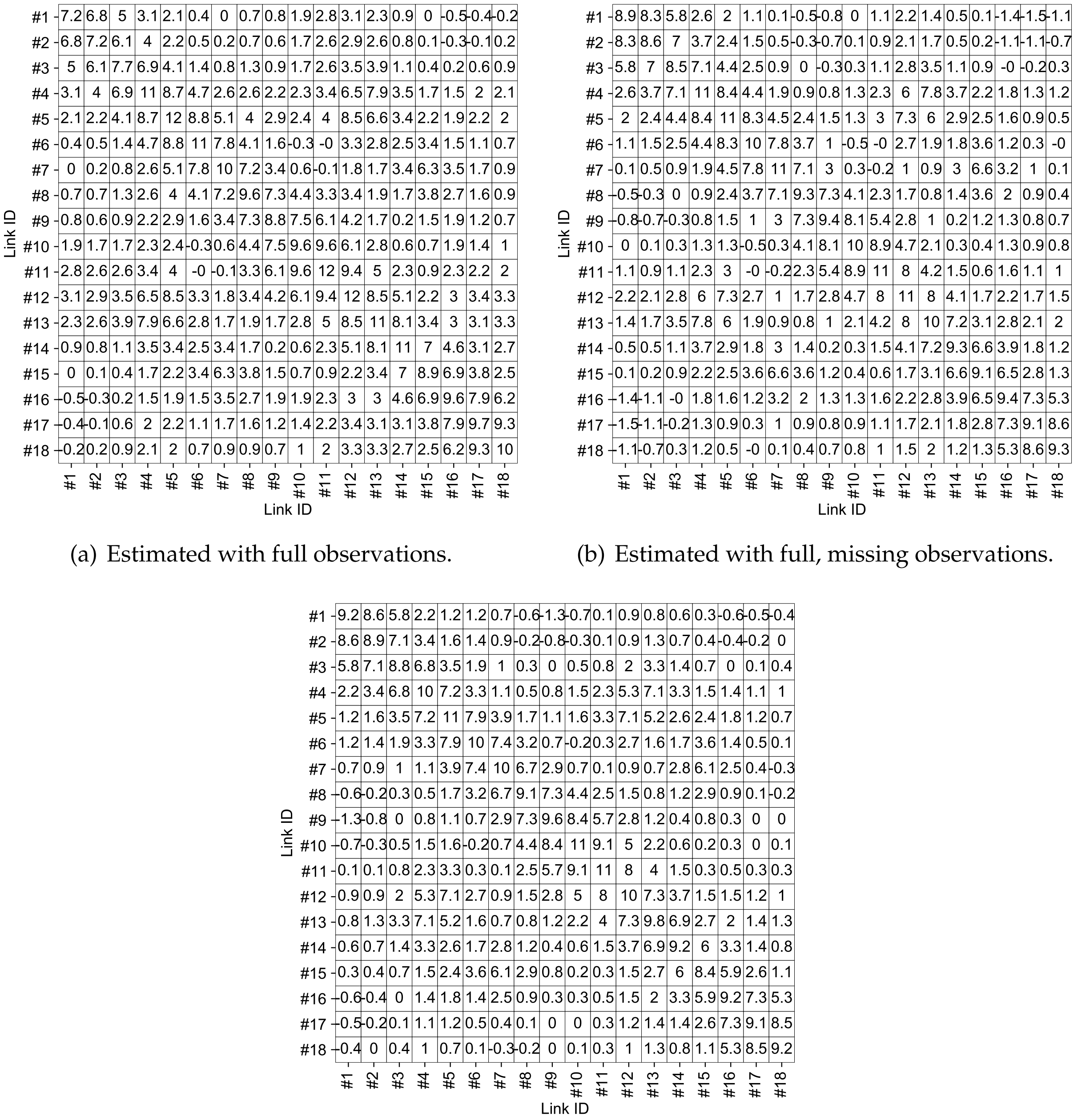
Table 3.1: Posterior mean of mean vectors estimated with different types of observations.

	Mean vector
(1)	[14.1, 15.2, 18.5, 13.6, 17.6, 15.3, 9.9, 23.8, 15.2, 11.5, 12.4, 15.5, 9.6, 13.4, 17.1, 15.0, 19.1, 21.3]
(2)	[14.2, 15.2, 18.3, 13.2, 17.2, 15.2, 10.0, 24.0, 15.1, 11.1, 12.0, 15.0, 9.3, 13.3, 17.1, 14.7, 18.7, 20.8]
(3)	[14.1, 15.1, 18.2, 13.3, 17.1, 15.0, 10.0, 24.1, 15.1, 11.1, 12.0, 15.0, 9.2, 13.2, 17.2, 14.9, 18.8, 20.8]

Table 3.2: KL divergence of distributions estimated with different types of observations.

	(1)	(2)	(3)
D_{KL}	0.2502	0.0748	**0.0565**

Bayesian approach has the advantage that we can estimate the posterior distributions over covariance/correlation matrices. We use credible intervals (CI) to measures the uncertainty of parameters. A CI is an interval with a particular probability to contain an unknown parameter value, and throughout this paper use 95% equal-tailed interval as CI



	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
#1	7.2	6.8	5	3.1	2.1	0.4	0	0.7	0.8	1.9	2.8	3.1	2.3	0.9	0	-0.5	-0.4	-0.2
#2	6.8	7.2	6.1	4	2.2	0.5	0.2	0.7	0.6	1.7	2.6	2.9	2.6	0.8	0.1	-0.3	-0.1	0.2
#3	5	6.1	7.7	6.9	4.1	1.4	0.8	1.3	0.9	1.7	2.6	3.5	3.9	1.1	0.4	0.2	0.6	0.9
#4	3.1	4	6.9	11	8.7	4.7	2.6	2.6	2.2	2.3	3.4	6.5	7.9	3.5	1.7	1.5	2	2.1
#5	2.1	2.2	4.1	8.7	12	8.8	5.1	4	2.9	2.4	4	8.5	6.6	3.4	2.2	1.9	2.2	2
#6	0.4	0.5	1.4	4.7	8.8	11	7.8	4.1	1.6	-0.3	-0	3.3	2.8	2.5	3.4	1.5	1.1	0.7
#7	0	0.2	0.8	2.6	5.1	7.8	10	7.2	3.4	0.6	-0.1	1.8	1.7	3.4	6.3	3.5	1.7	0.9
#8	0.7	0.7	1.3	2.6	4	4.1	7.2	9.6	7.3	4.4	3.3	3.4	1.9	1.7	3.8	2.7	1.6	0.9
#9	0.8	0.6	0.9	2.2	2.9	1.6	3.4	7.3	8.8	7.5	6.1	4.2	1.7	0.2	1.5	1.9	1.2	0.7
#10	1.9	1.7	1.7	2.3	2.4	-0.3	0.6	4.4	7.5	9.6	9.6	6.1	2.8	0.6	0.7	1.9	1.4	1
#11	2.8	2.6	2.6	3.4	4	-0	-0.1	3.3	6.1	9.6	12	9.4	5	2.3	0.9	2.3	2.2	2
#12	3.1	2.9	3.5	6.5	8.5	3.3	1.8	3.4	4.2	6.1	9.4	12	8.5	5.1	2.2	3	3.4	3.3
#13	2.3	2.6	3.9	7.9	6.6	2.8	1.7	1.9	1.7	2.8	5	8.5	11	8.1	3.4	3	3.1	3.3
#14	0.9	0.8	1.1	3.5	3.4	2.5	3.4	1.7	0.2	0.6	2.3	5.1	8.1	11	7	4.6	3.1	2.7
#15	0	0.1	0.4	1.7	2.2	3.4	6.3	3.8	1.5	0.7	0.9	2.2	3.4	7	8.9	6.9	3.8	2.5
#16	-0.5	-0.3	0.2	1.5	1.9	1.5	3.5	2.7	1.9	1.9	2.3	3	3	4.6	6.9	9.6	7.9	6.2
#17	-0.4	-0.1	0.6	2	2.2	1.1	1.7	1.6	1.2	1.4	2.2	3.4	3.1	3.1	3.8	7.9	9.7	9.3
#18	-0.2	0.2	0.9	2.1	2	0.7	0.9	0.9	0.7	1	2	3.3	3.3	2.7	2.5	6.2	9.3	10

(a) Estimated with full observations.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
#1	8.9	8.3	5.8	2.6	2	1.1	0.1	-0.5	-0.8	0	1.1	2.2	1.4	0.5	0.1	-1.4	-1.5	-1.1
#2	8.3	8.6	7	3.7	2.4	1.5	0.5	-0.3	-0.7	0.1	0.9	2.1	1.7	0.5	0.2	-1.1	-1.1	-0.7
#3	5.8	7	8.5	7.1	4.4	2.5	0.9	0	-0.3	0.3	1.1	2.8	3.5	1.1	0.9	-0	-0.2	0.3
#4	2.6	3.7	7.1	11	8.4	4.4	1.9	0.9	0.8	1.3	2.3	6	7.8	3.7	2.2	1.8	1.3	1.2
#5	2	2.4	4.4	8.4	11	8.3	4.5	2.4	1.5	1.3	3	7.3	6	2.9	2.5	1.6	0.9	0.5
#6	1.1	1.5	2.5	4.4	8.3	10	7.8	3.7	1	-0.5	-0	2.7	1.9	1.8	3.6	1.2	0.3	-0
#7	0.1	0.5	0.9	1.9	4.5	7.8	11	7.1	3	0.3	-0.2	1	0.9	3	6.6	3.2	1	0.1
#8	-0.5	-0.3	0	0.9	2.4	3.7	7.1	9.3	7.3	4.1	2.3	1.7	0.8	1.4	3.6	2	0.9	0.4
#9	-0.8	-0.7	-0.3	0.8	1.5	1	3	7.3	9.4	8.1	5.4	2.8	1	0.2	1.2	1.3	0.8	0.7
#10	0	0.1	0.3	1.3	1.3	-0.5	0.3	4.1	8.1	10	8.9	4.7	2.1	0.3	0.4	1.3	0.9	0.8
#11	1.1	0.9	1.1	2.3	3	-0	-0.2	2.3	5.4	8.9	11	8	4.2	1.5	0.6	1.6	1.1	1
#12	2.2	2.1	2.8	6	7.3	2.7	1	1.7	2.8	4.7	8	11	8	4.1	1.7	2.2	1.7	1.5
#13	1.4	1.7	3.5	7.8	6	1.9	0.9	0.8	1	2.1	4.2	8	10	7.2	3.1	2.8	2.1	2
#14	0.5	0.5	1.1	3.7	2.9	1.8	3	1.4	0.2	0.3	1.5	4.1	7.2	9.3	6.6	3.9	1.8	1.2
#15	0.1	0.2	0.9	2.2	2.5	3.6	6.6	3.6	1.2	0.4	0.6	1.7	3.1	6.6	9.1	6.5	2.8	1.3
#16	-1.4	-1.1	-0	1.8	1.6	1.2	3.2	2	1.3	1.3	1.6	2.2	2.8	3.9	6.5	9.4	7.3	5.3
#17	-1.5	-1.1	-0.2	1.3	0.9	0.3	1	0.9	0.8	0.9	1.1	1.7	2.1	1.8	2.8	7.3	9.1	8.6
#18	-1.1	-0.7	0.3	1.2	0.5	-0	0.1	0.4	0.7	0.8	1	1.5	2	1.2	1.3	5.3	8.6	9.3

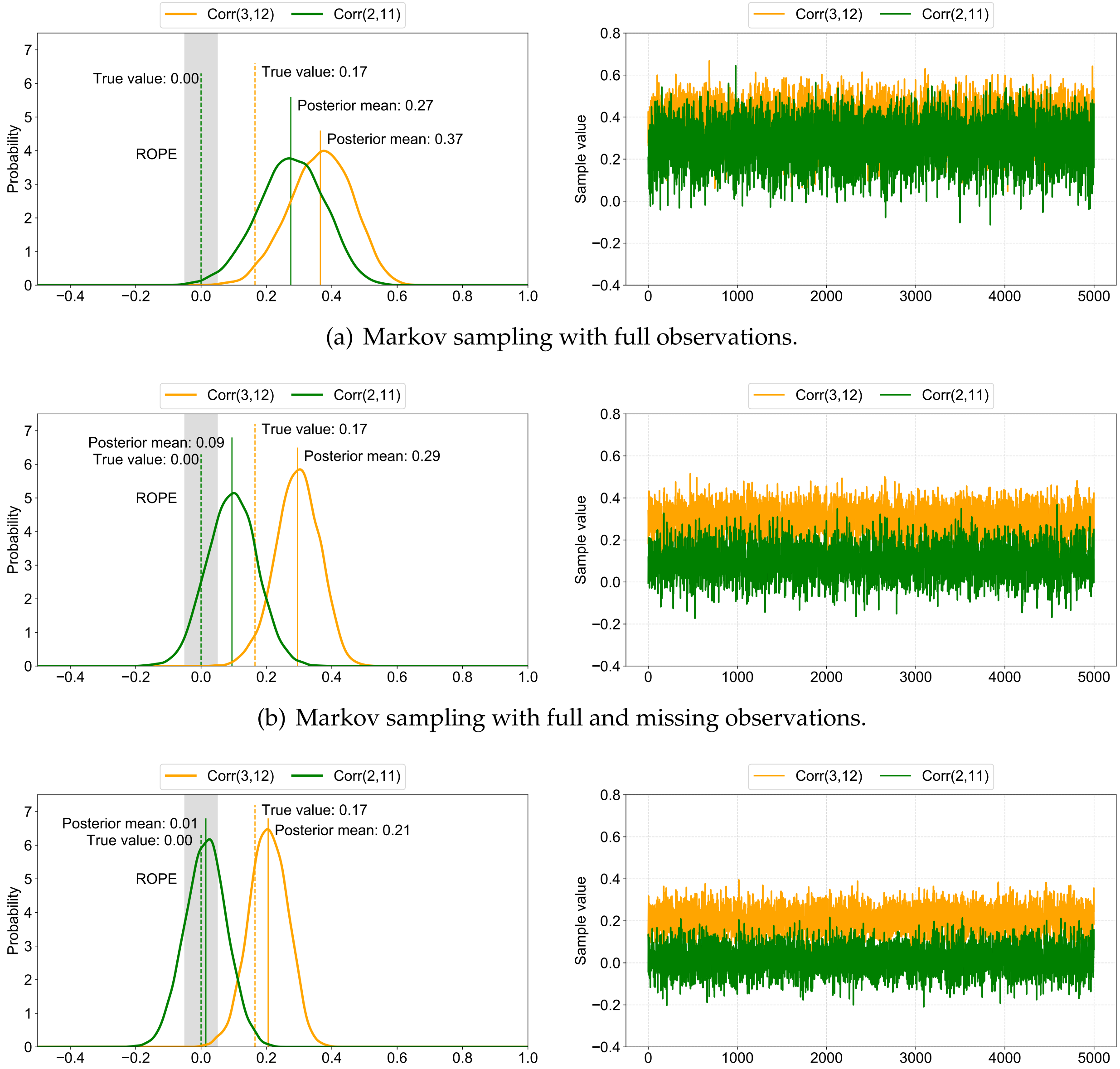
(b) Estimated with full, missing observations.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
#1	9.2	8.6	5.8	2.2	1.2	1.2	0.7	-0.6	-1.3	-0.7	0.1	0.9	0.8	0.6	0.3	-0.6	-0.5	-0.4
#2	8.6	8.9	7.1	3.4	1.6	1.4	0.9	-0.2	-0.8	-0.3	0.1	0.9	1.3	0.7	0.4	-0.4	-0.2	0
#3	5.8	7.1	8.8	6.8	3.5	1.9	1	0.3	0	0.5	0.8	2	3.3	1.4	0.7	0	0.1	0.4
#4	2.2	3.4	6.8	10	7.2	3.3	1.1	0.5	0.8	1.5	2.3	5.3	7.1	3.3	1.5	1.4	1.1	1
#5	1.2	1.6	3.5	7.2	11	7.9	3.9	1.7	1.1	1.6	3.3	7.1	5.2	2.6	2.4	1.8	1.2	0.7
#6	1.2	1.4	1.9	3.3	7.9	10	7.4	3.2	0.7	-0.2	0.3	2.7	1.6	1.7	3.6	1.4	0.5	0.1
#7	0.7	0.9	1	1.1	3.9	7.4	10	6.7	2.9	0.7	0.1	0.9	0.7	2.8	6.1	2.5	0.4	-0.3
#8	-0.6	-0.2	0.3	0.5	1.7	3.2	6.7	9.1	7.3	4.4	2.5	1.5	0.8	1.2	2.9	0.9	0.1	-0.2
#9	-1.3	-0.8	0	0.8	1.1	0.7	2.9	7.3	9.6	8.4	5.7	2.8	1.2	0.4	0.8	0.3	0	0
#10	-0.7	-0.3	0.5	1.5	1.6	-0.2	0.7	4.4	8.4	11	9.1	5	2.2	0.6	0.2	0.3	0	0.1
#11	0.1	0.1	0.8	2.3	3.3	0.3	0.1	2.5	5.7	9.1	11	8	4	1.5	0.3	0.5	0.3	0.3
#12	0.9	0.9	2	5.3	7.1	2.7	0.9	1.5	2.8	5	8	10	7.3	3.7	1.5	1.5	1.2	1
#13	0.8	1.3	3.3	7.1	5.2	1.6	0.7	0.8	1.2	2.2	4	7.3	9.8	6.9	2.7	2	1.4	1.3
#14	0.6	0.7	1.4	3.3	2.6	1.7	2.8	1.2	0.4	0.6	1.5	3.7	6.9	9.2	6	3.3	1.4	0.8
#15	0.3	0.4	0.7	1.5	2.4	3.6	6.1	2.9	0.8	0.2	0.3	1.5	2.7	6	8.4	5.9	2.6	1.1
#16	-0.6	-0.4	0	1.4	1.8	1.4	2.5	0.9	0.3	0.3	0.5	1.5	2	3.3	5.9	9.2	7.3	5.3
#17	-0.5	-0.2	0.1	1.1	1.2	0.5	0.4	0.1	0	0	0.3	1.2	1.4	1.4	2.6	7.3	9.1	8.5
#18	-0.4	0	0.4	1	0.7	0.1	-0.3	-0.2	0	0.1	0.3	1	1.3	0.8	1.1	5.3	8.5	9.2

(c) Estimated with all observations.

Figure 3.5: Posterior mean of covariance matrices estimated with different types of observations.

unless stated otherwise. Moreover, we want to determine whether a particular correlation value is equivalent to a "null" value for practical purposes. For making decisions about the null value, we can use the equivalence test based on the full posterior distribution and region of practical equivalence (ROPE) (Kruschke and Liddell, 2018). The equivalence



(a) Markov sampling with full observations.

(b) Markov sampling with full and missing observations.

(c) Markov sampling with full, missing, and ragged observations.

Figure 3.6: The estimated posterior distributions over two entries of the covariance matrix with different observations.

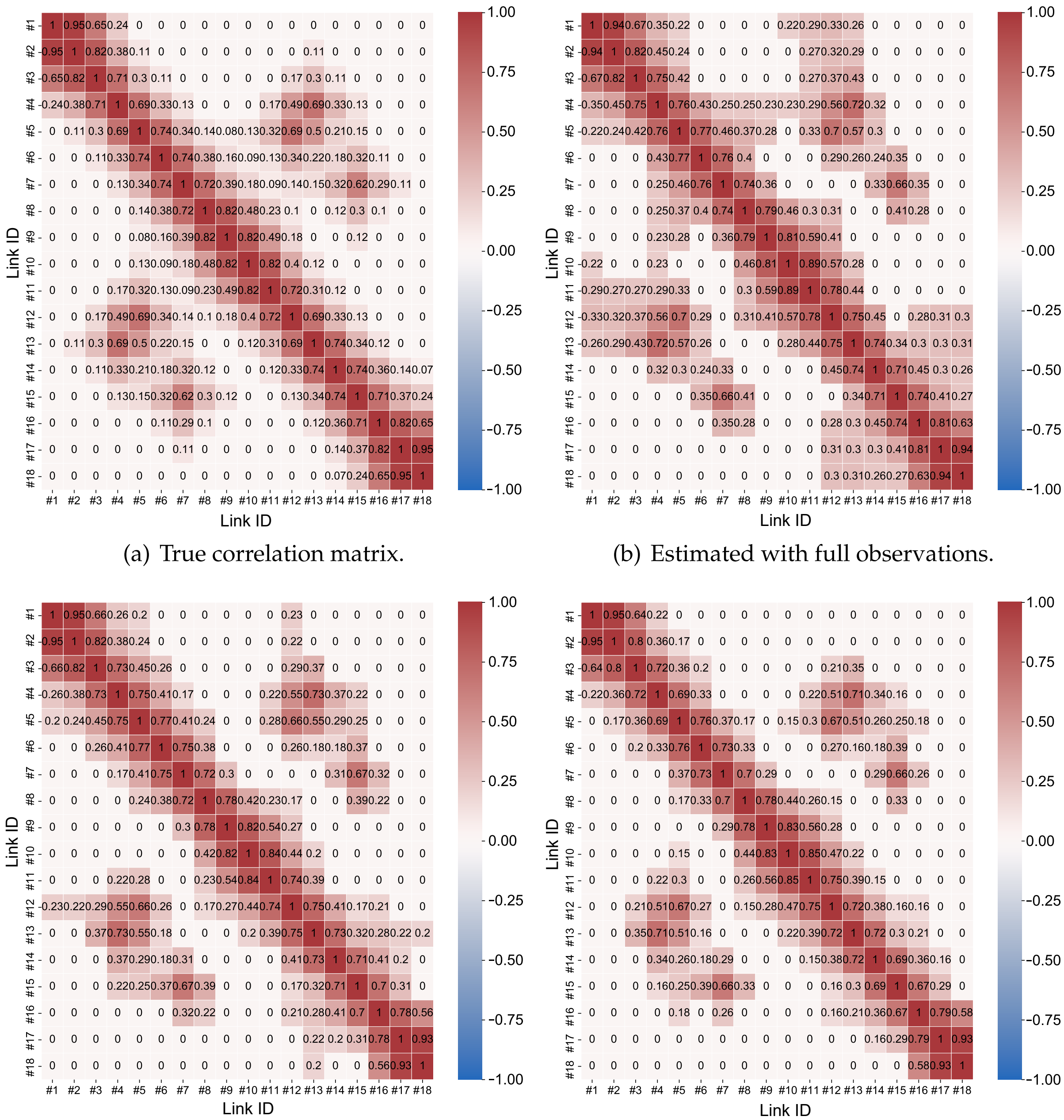
test checks the percentage of full posterior that falls inside the ROPE. The null value is declared to be rejected when the percentage is sufficiently low; the null value is considered to be accepted if the percentage is sufficiently high. Throughout this paper, we set the ROPE range with $(-0.05, 0.05)$ and the rejected-threshold with 5%. Figure 3.6 presents the estimated posterior distributions over two entries ($\mathrm{Corr}(2, 11)$ and $\mathrm{Corr}(3, 12)$) of correlation matrices with different observations. The true value of $\mathrm{Corr}(2, 11)$ is zero. From

Figure 3.6 (a), we can see the posterior mean of $\text{Corr}(2, 11)$ estimated with full observations is 0.27 (CI: $[0.07, 0.46]$), and the posterior distribution shows that the estimated $\text{Corr}(2, 11)$ is much larger than zero, and the Bayesian credible interval is largely outside the ROPE. From Figure 3.6 (b), we can see the posterior mean of $\text{Corr}(2, 11)$ estimated with full and missing observations is 0.09 (CI: $[-0.05, 0.24]$), which is more accurate than only using only full observations and we fail to reject the value zero. In Figure 3.6 (c), the posterior mean of $\text{Corr}(2, 11)$ estimated by using all observations is 0.01 (CI: $[-0.11, 0.14]$), and the percentage of the credible interval that falls in the ROPE is larger than 5%, meaning that we cannot reject the value zero. The true value of $\text{Corr}(3, 12)$ is 0.17, and the posterior mean values estimated with full observations, full and missing observations, and all observations are 0.31 (CI: $[0.17, 0.54]$), 0.24 (CI: $[0.16, 0.42]$), and 0.19 (CI: $[0.09, 0.31]$), respectively. All the credible intervals of $\text{Corr}(3, 12)$ largely fall outside the ROPE, indicating that we can reject the value zero. With the use of missing/ragged observations, the posterior mean of covariance becomes more accurate, and the posterior probability density becomes thinner, indicating a smaller standard deviation.

Finally, we do the equivalence tests for entries of estimated correlation matrices and we set zero for correlations that fail to reject the value zero for better visualization. Moreover, we set zeros for the true correlations which are lower than 0.05 for convenient comparison considering that we use the ROPE with $(-0.05, 0.05)$. The estimated and true correlation matrices are shown in Figure 3.7. Figure 3.7 (a) presents the true correlation matrix. Figure 3.7 (b)-(c) show the correlation matrices estimated with full observations, full and missing observations, all observations, respectively. We can see that using all observations can obtain the most accurate estimated correlation matrix, which agrees with the evaluation using KL divergences.

3.6.2 Case 2: Guangzhou Bus Data

In this section, we apply the proposed Bayesian model to real-world data to empirically quantify the link travel time correlation of a bus route. The data used in this paper are the bus in-out-stop record data collected in Guangzhou, China, during the weekdays from December 8, 2016 to December 15, 2016. The information of the data is outlined in Table 3.3. These data were collected by the bus in-out-stop record system, i.e., the automatic bus announcing system. When a bus enters or exits a bus stop, the system reports the arrival or departure information and records the time stamp accordingly. Thus we can easily obtain the link travel times from the data. We take bus route No. 60 as a case and aim to quantify this bus route's link travel time correlation. First, we select the other three bus routes



(a) True correlation matrix.

(b) Estimated with full observations.

(c) Estimated with full and missing observations.

(d) Estimated correlation with all observations.

Figure 3.7: The true correlation matrix and estimated correlation matrices.

related to route No. 60. All the studied bus routes are displayed in Figure 3.8, and they are in the CBD of Guangzhou. Route No. 60 has 20 links; route No. 257 shares 7 links with route No. 60 (from link #2 to link #8); route No. B18 shares 10 links with route No. 60 (from link #6 to link #15); route No. 210 shares 17 links with route No. 60, but the buses of route No. 210 do not stop after entering link #3 until leaving link #11. As our defined link travel time considers the dwell time, these long-ragged data may have a negative impact on the estimation; we thus only used the 8 shared links (link #2, #3, #12-#17). We divide one day into four periods: morning peak $(7:00-10:00)$, normal period $(10:00-17:00)$, afternoon peak $(17:00-20:00)$, and night period $(20:00-7:00)$. The overview of all used data is shown in Figure 3.9. We can see that all the bus routes have many missing and ragged values.

Table 3.3: Description of Data.

Variable	Description	Example
ID	Identity for bus data record	1612020547101390
OBUID	Identity for bus	911721
TRIP_ID	Identity for bus trip	1612012250030880
ROUTE_ID	Identity for bus route	201
ROUTE_NAME	Bus route name	No. 24
ROUTESUB_ID	Identity of bus route direction	502669
ROUTE_STA_ID	Identity of bus stop	84279
STOP_NAME	Bus stop name	Dunhe Stop
AD_FLAG	Bus state: arrival (1) or departure (0)	1
AD_TIME	The time bus reported arriving at/leaving a bus stop	20161202, 05:47:08

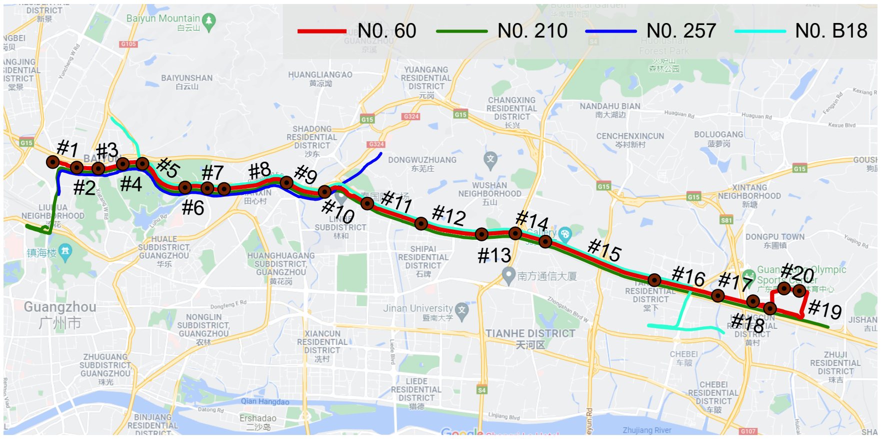


Figure 3.8: Bus route No. 60 and the related routes in Guangzhou bus network.

The numbers of MCMC iterations are $k_1 = 10000$ and $k_2 = 5000$, respectively. Figure 3.10 presents the estimated posterior distributions over two entries ($\mathrm{Corr}(11, 12)$ and $\mathrm{Corr}(11, 16)$) of correlations matrices for different periods. We can find that the distributions over a correlation are distinct for different periods. For example, the posterior mean values of $\mathrm{Corr}(11, 16)$ are 0.05 (CI: $[-0.08, 0.18]$), 0.19 (CI: $[0.07, 0.31]$), 0.59 (CI: $[0.46, 0.7]$), 0.65 (CI: $[0.52, 0.75]$) for morning peak, normal period, afternoon peak, night period, respectively. The equivalence test of $\mathrm{Corr}(11, 16)$ for the morning peak fails to reject the

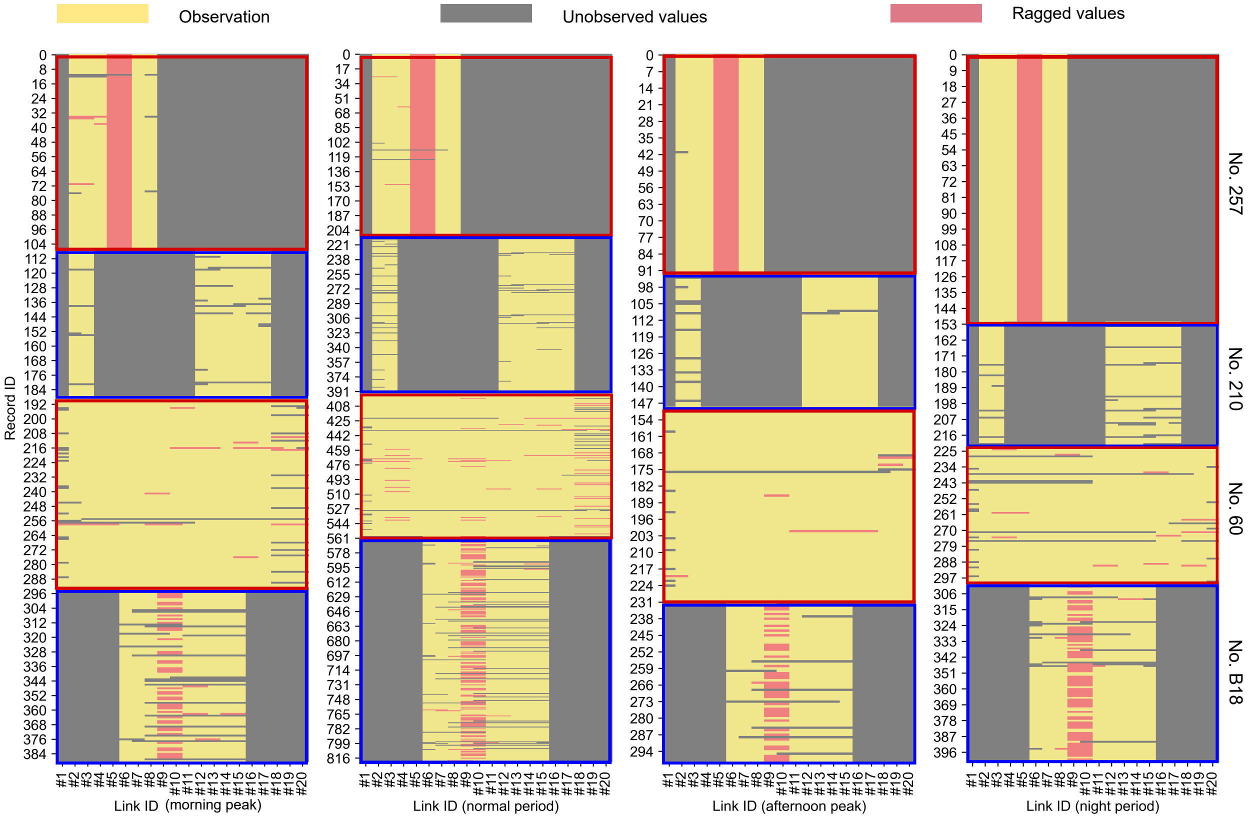
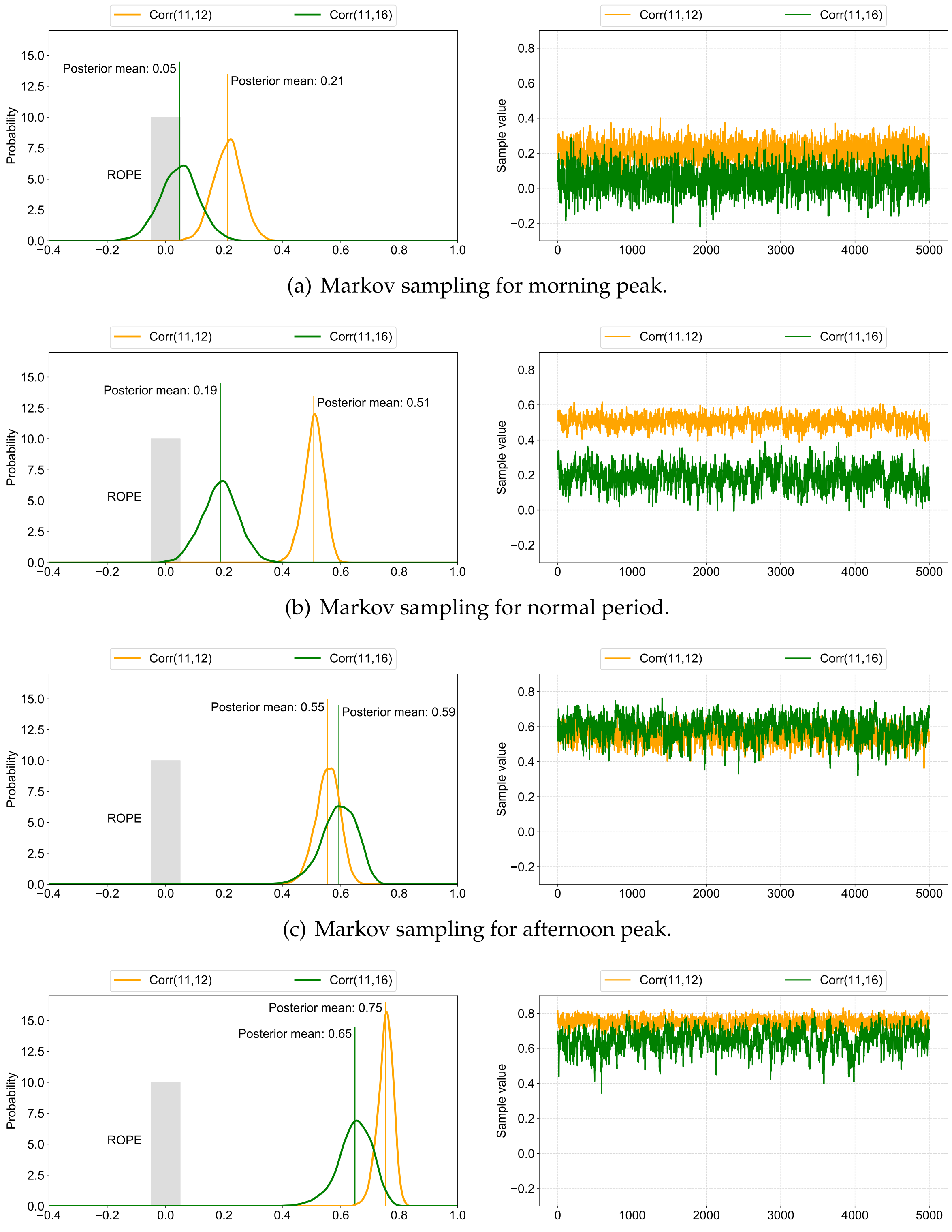


Figure 3.9: Data overview.

value zero, while all the 95% CI of Corr$(11, 16)$ for the other periods largely fall outside ROPE, indicating that travel times on these two links are positively correlated.

We calculate credible intervals for entries of estimated correlation matrices using drawn samples and we set zero for correlations that cannot reject the value zero for better visualization. Figure 3.11 shows the estimated correlation matrices for four different periods. Each cell in the correlation matrix shows the correlation between two variables. Essentially, this kind of correlation matrix is Pearson's Product-Moment Correlation. The cell number can help to understand how strong a relationship is between two variables. The further away the cell value is from zero, the stronger the relationship between the two variables. Generally, when the cell's absolute value of correlation matrix is zero, the relationship between the corresponding variables will be considered as no relationship; when the absolute value is lower than 0.25, the relationship will be considered as a weak correlation; when the absolute value is between 0.25 and 0.5, the relationship will be regarded as a medium relationship; when the absolute value is larger than 0.5, it indicates these two variables are strongly correlated. Furthermore, the sign of the cell value also means a different correlation. A positive value indicates the positive correlation between



(a) Markov sampling for morning peak.

(b) Markov sampling for normal period.

(c) Markov sampling for afternoon peak.

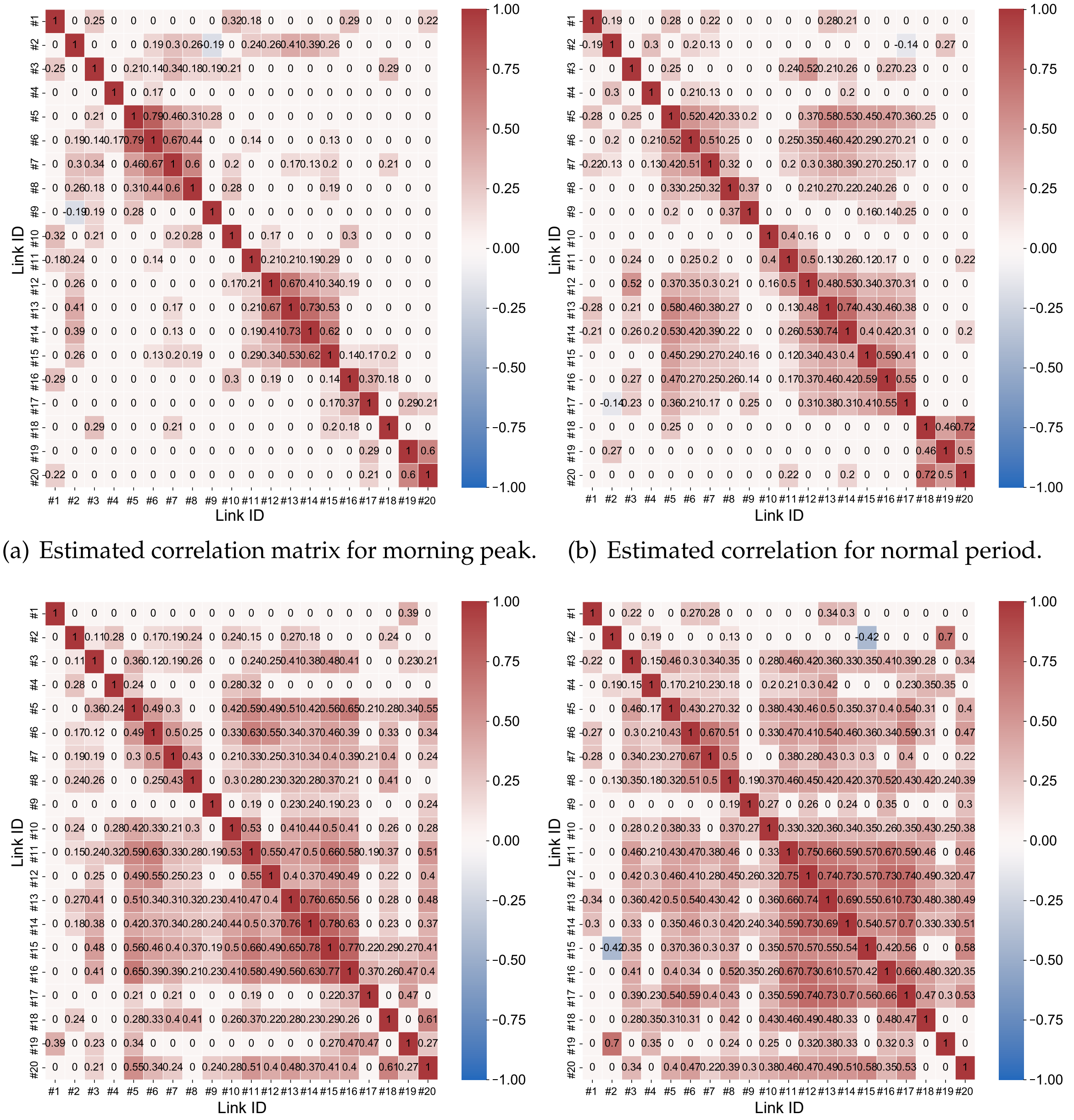
(d) Markov sampling for night period.

Figure 3.10: The estimated posterior distributions over two entries of the correlation matrices.

two variables, while a negative value represents the negative correlation. For the positive correlation, when the value of one variable increases, the value of the other variable increases in a similar way. For the negative correlation, when the value of one variable increases, the value of the other variable tends to decrease.

Figure 3.11 reveals some characteristics of link travel time correlation of the bus route. First, we can find that the estimated correlation matrices vary for four periods, indicating that the link travel time correlation is time-varying. Overall, more correlated link pairs exist during the afternoon peak and the night period, while fewer correlated link pairs exist during the morning peak. The directional bus route stretches from urban business districts to suburban areas; thus, the traffic conditions/passenger flows are different for the morning and afternoon peaks. The better traffic condition and the small passenger flow exist in the morning because few people go to suburban areas on weekday morning. On the contrary, traffic congestions and large passenger flow happen in the afternoon peak as more people go home from urban to suburban areas. Traffic congestion and large passenger flow can cause the bus bunching phenomenon: a lagging bus must collect more passengers and, therefore, needs more travel time; on the other hand, a subsequent bus of the lagging bus will have fewer passengers, and its travel time will be shorter. We conclude that during the afternoon peak, bus bunching can make more link pairs correlated. Second, most link pairs do not have strong correlations as most cell values are lower than 0.5 for these four time periods. Meanwhile, the values of cells with strong correlations are positive, which means link travel time variables of a bus route are more likely to have a positive correlation if they have a strong relationship. A possible reason for these positive correlations is the bus bunching phenomenon. Few negative correlations exist in link pairs though they are weak or medium correlation. Third, both local and long-range correlations exist on the bus route. Many strong correlations exist in local link pairs. For example, adjacent link pairs (link #5, link #6), (link #6, link #7), (link #7, link #8) during the morning peak have strong correlations. Apart from the adjacent link pairs, strong correlations exist in link pairs with long distances. In Figure 3.11 (c), (link #5, link #11), (link #5, link #13), (link #6, link #11) are long-range correlations.

Finally, we clarify that the link travel time correlation in other transportation modes may differ from the bus. Bus bunching is a critical reason that affects link travel time correlation. The link travel time correlation may not be as significant as the bus for modes without the bus bunching phenomenon, such as car and truck.



(a) Estimated correlation matrix for morning peak.
(b) Estimated correlation for normal period.
(c) Estimated correlation for afternoon peak.
(d) Estimated correlation matrix for night period.

Figure 3.11: The estimated correlation matrix for different periods.

3.6.3 Case 3: Link/Trip Travel Time Forecasting

In this section, we show that covariance matrices can be used for probabilistic forecasting of bus link/trip travel time. The proposed Bayesian model can estimate the posterior mean vectors and covariance matrices for different periods; we thus can obtain the conditional posterior distribution over forecasting links conditioned on observed links. Note that here we do not aim to propose a sophisticated forecasting model to compete with state-of-the-art models; instead, the aim is to present a basic model under a simple scenario to illustrate the effectiveness of using covariance matrices in travel time forecasting. The experiment uses the following five weekdays' full observations of route No. 60 from December 16 to December 22, 2016 to test the forecasting performance. As a simple experiment, the task is to forecast the link travel times of the last nine links (from link #12 to link #20) given the link travel times of the first eleven links (from link #1 to link #11).

We select the historical average (HA) as the benchmark model. For the link travel time at a certain period of the day, HA uses the average link travel time at that period in the training set as the forecast value. Then we compare the performance of these two methods, which are evaluated by the root mean square error (RMSE) and the mean absolute percentage error (MAPE):

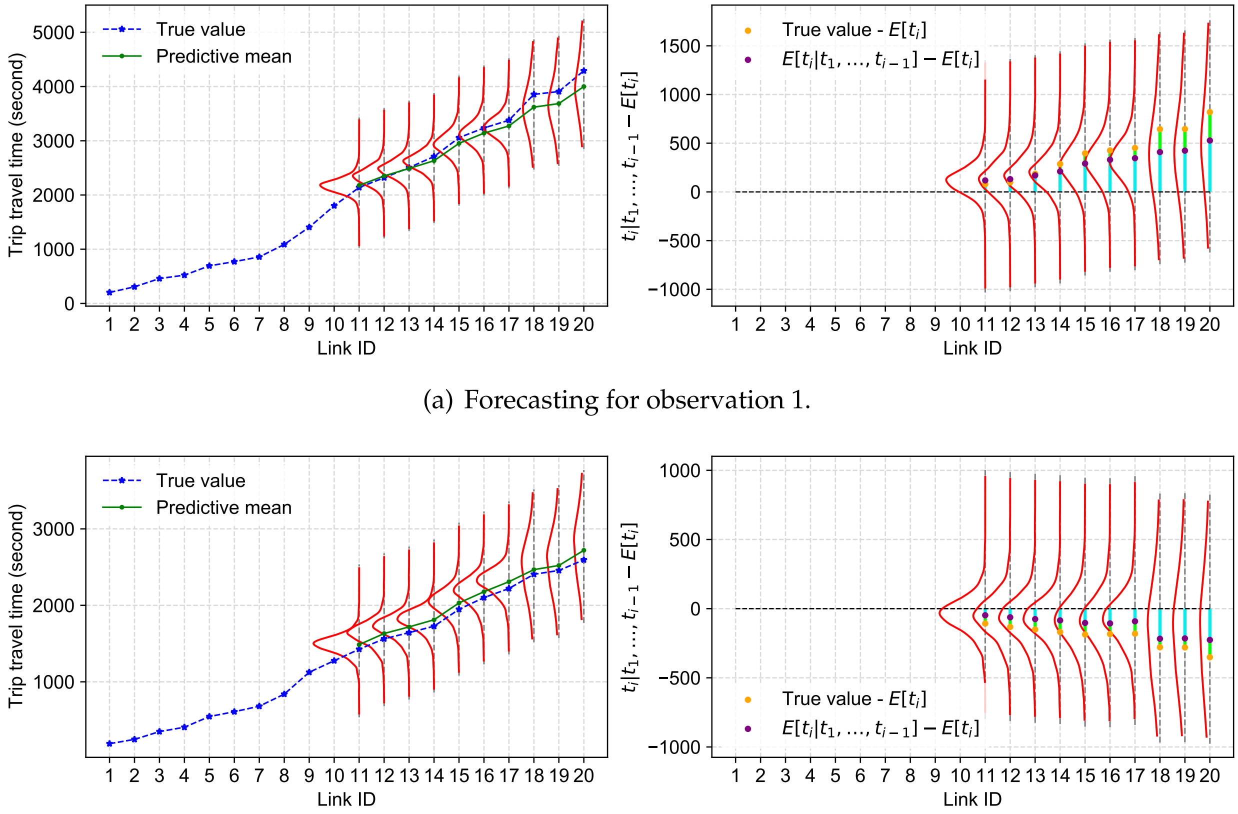
$$\begin{aligned}\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},\\ \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,\end{aligned} \tag{3.17}$$

where $y_i, \hat{y}_i, i = 1, \ldots, n$ are the true values and forecasts, respectively. Table 3.4 presents the forecasting performance using the Bayesian model and historical average. We can find Bayesian model performs better than the historical average method for all periods.

Table 3.4: The forecasting performance of two methods for different periods.

		Morning peak	Normal period	Afternoon peak	Night period
RMSE	Bayesian model	**27.74**	**35.45**	**61.20**	**20.95**
	Historical average	32.54	36.07	67.13	25.86
MAPE	Bayesian model	**0.1193**	**0.1186**	**0.1638**	**0.0885**
	Historical average	0.1563	0.1188	0.1742	0.1116

This model can make probabilistic forecasting for bus trip travel time. As an example, we use two piece of test data in the afternoon peak to show probabilistic forecasting. Assume we have observed the first ten links' travel times, and the goal is to forecast the trip travel time distributions. Figure 3.12 shows the probabilistic forecasting results. In the left panel, the blue points are the true trip travel times, and the green points are the predictive mean values. We can see that the predictive mean values fit the actual values, indicating the Bayesian model can make good forecasting. Moreover, the red bell curves are the trip travel time distributions, and we can see that the red bell curves are fatter with the increasing number of links in a trip, indicating the variance is increasing. The right panels present the mean corrected estimation, and the purple points (we refer to them as corrected mean values) are computed by posterior conditional mean values minus model mean values; the orange points are the difference between true values and model mean values. We can find that the posterior conditional mean can make a more accurate prediction than the model mean. If we do not use the information of the observed link travel times, the forecasting mean vectors should be equivalent to the model mean vectors. As we can see, the corrected mean values for observation 1 shown in Figure 3.12 (a) are larger than zero, while the corrected mean values for observation 2 shown in Figure 3.12 (b) are lower than zero, indicating the link travel time observations from link #1 to #10 indeed help update the forecasting values for the following links.



(a) Forecasting for observation 1.

(b) Forecasting for observation 2.

Figure 3.12: The probabilistic forecasting for trip travel time.

We have to clarify that the posterior predictive distribution of the current model is only a rough reference. As shown in Figure 3.12, the variance of the predictive distribution is too high for a practical application. This could be improved by using a more appropriate probability distribution for link travel time (e.g., the Gaussian mixture model). Besides, measures should be taken to avoid the negative part of link travel time distribution. Despite the above limitation, using the covariance matrices produces probabilistic forecasting of link travel time and the predictive mean is more accurate than the historical average, which verifies the effectiveness of using covariance matrices for link/trip travel time forecasting.

3.7 Conclusion

In this paper, we have proposed a Bayesian Gaussian model to quantify the link travel time correlation of a bus route. The approach overcomes the issue of small sample sizes on a single bus route by incorporating data from other relevant bus routes. The proposed model can also impute those missing and ragged values in an incomplete link travel time vector. Three experiments are conducted in this paper. The first experiment is conducted on synthetic data with known covariance, and our result shows that the proposed Bayesian model can accurately recover the underlying mean and covariance from corrupted link travel time observations. In the second empirical experiment, we used real-world bus in-out-stop record data to quantify link travel time correlation. Our empirical analysis shows that (1) link travel times are clearly not independent on a bus route, and the estimated correlations vary substantially for different time periods of a day; (2) most link pairs are not strongly correlated, and most correlations are positive while negative correlations also exist; (3) both local and long-range correlations could exist on a bus route. Our results also suggest that simplified covariance assumptions (e.g., local spatial correlation) might be inappropriate for modeling travel time on a bus route. Finally, we applied the estimated covariance matrices to forecast link/trip travel time. An additional test data set during five weekdays is used to verify the forecasting performance, and the results show that the proposed model clearly outperforms a historical average baseline.

Our approach has potential implications for both practice and research. First, the proposed Bayesian model can estimate the covariance matrices essential to performing probabilistic forecasting of bus travel time. Second, the imputation method can also handle ragged values in other fields, such as economics, medicine, and social sciences. The ragged definition can be used to model link travel time from origin-destination-based trip travel time observations; in this case, $\mathbf{G}_i$ becomes a row vector encoding the linear transformation to obtain the total travel time for (a single) trip i. Third, this approach can be used in estimating automobile's link travel time correlation in a small network.

Our proposed Bayesian Gaussian model has several limitations. First, it is challenging to infer the high-dimensional covariance structure (e.g., $n > 100$) for automobile road network or bus network. When the dimension of the covariance is large, the computation is highly expensive, and small-size observations constrain the estimation accuracy. Second, this model does not consider the influence of dwell time from multiple bus routes. Our defined link travel time includes dwelling time. However, different bus routes have distinct characteristics of dwell time due to factors including passenger flow/demand, bus schedule, and bus types. For example, bus routes with lower passengers flow will have shorter link travel times, while a larger passenger flow will cause longer travel times. In this case, our assumption that related bus routes share the same link travel time distribution may no longer hold. The influence of dwell time from multiple bus routes could be studied in further research. Third, the way we model the covariance structure of different time periods is by dividing samples into several periods and estimating the proposed model independently. Although simple, this approach ignores the temporal dynamic of the covariance structure—the covariance structure may vary smoothly and continuously over time. Our further research is to develop new models to characterize time-varying link travel time correlation. Last, this model relies on the assumption that the joint link travel times follow multivariate Gaussian distribution. This assumption comes in handy for quantifying correlation while it is too general for link travel time forecasting. Real-world link travel time is non-negative and the distribution of it could be skewed and multimodal. We could try to overcome this limitation by using more accurate distributions (e.g., truncated distribution, log-normal distribution and Gaussian mixture model) in future studies.

Chapter 4

Bayesian Forecasting of Bus Travel Time

This chapter is an article published in *Transportation Science*:

- **Chen, X.**, Cheng, Z., Jin, J. G., Trépanier, M., Sun, L. 2023. Probabilistic forecasting of bus travel time with a Bayesian Gaussian mixture model. *Transportation Science*, 57(6), 1516-1535.

This chapter corresponds to the probabilistic forecasting of bus travel time with a Bayesian Gaussian mixture model. This chapter is based on the important findings of link travel time correlations in the previous Chapter 3.

4.1 Abstract

Accurate forecasting of bus travel time and its uncertainty is critical to service quality and operation of transit systems: it can help passengers make informed decisions on departure time, route choice, and even transport mode choice, and also support transit operators on tasks such as crew/vehicle scheduling and timetabling. However, most existing approaches in bus travel time forecasting are based on deterministic models that provide only point estimation. To this end, we develop in this paper a Bayesian probabilistic model for forecasting bus travel time and estimated time of arrival (ETA). To characterize the strong dependencies/interactions between consecutive buses, we concatenate the link travel time vectors and the headway vector from a pair of two adjacent buses as a new augmented variable and model it with a mixture of constrained multivariate Gaussian distributions. This approach can naturally capture the interactions between adjacent buses (e.g., correlated speed and smooth variation of headway), handle missing values in data, and depict the multimodality in bus travel time distributions. Next, we assume different periods in a day share the same set of Gaussian components, and use time-varying mixing coefficients to characterize the systematic temporal variations in bus operation. For model inference, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm to obtain the posterior distributions of model parameters and make probabilistic forecasting. We test the proposed model using the data from two bus lines in Guangzhou, China. Results show that our approach significantly outperforms baseline models that overlook bus-to-bus interactions, in terms of both predictive means and distributions. Besides forecasting, the parameters of the proposed model contain rich information for understanding/improving the bus service, e.g., analyzing link travel time and headway correlation using covariance matrices and understanding time-varying patterns of bus fleet operation from the mixing coefficients.

4.2 Introduction

Cities are now facing severe traffic congestion and air pollution due to the over-reliance on cars. Promoting public transportation is one of the most effective and strategic ways to achieve sustainable urban transportation. However, there are various factors preventing people from using public transit, such as low reliability of travel time, uncomfortable riding experience, and inaccessible stops far away from home. Survey studies have shown that passengers highly care about the accurate forecasting of bus travel time and the estimated time of arrival (ETA) and its reliability (uncertainty) (Lam and Small, 2001). Probabilistic forecasting of bus travel time provides both the expected value and the uncertainty, which not only helps bus agencies design robust bus management strategies (e.g., bus priority signal control, bus bunching control, dynamic bus holding control (Xuan et al., 2011)) to enhance bus services, but also helps travelers make better travel plans regarding departure time, route choice, and even transport mode choice (Cats and Gkioulou, 2017).

Most existing studies on bus travel time forecasting mainly center on making point estimation (i.e., deterministic forecasting) but ignore the importance of travel time uncertainty (Ricard et al., 2022). There exist many deterministic bus travel time forecasting methods, including historical average (HA) (Farhan et al., 2002), Autoregressive Integrated Moving Average (ARIMA) (Madzlan et al., 2010), Artificial Neural Network (ANN) (Chien et al., 2002; Gurmu and Fan, 2014), Support Vector Machine (SVM) (Bin et al., 2006; Yu et al., 2011; Kumar et al., 2013; Bachu et al., 2021), Kalman Filter (KF) (Cathey and Dailey, 2003), K-nearest neighbors model (KNN) (Liu et al., 2012; Kumar et al., 2019), deep learning models such as Long Short-Term Memory (LSTM) (Osman et al., 2021; Alam et al., 2021), and various hybrid models (Yu et al., 2018; Zhang et al., 2021), to name just a few. Despite its popularity and simplicity, a major limitation of the deterministic approach is that they cannot give the uncertainty of the forecasting. In practice, probabilistic forecasting (i.e., forecast the distribution of bus travel time) is often favored over deterministic forecasting (Yetiskul and Senbil, 2012). For passengers, knowing the distribution of ETA is more useful than a single point estimation, as they may prefer a bus route with the smallest travel time variance among bus routes with similar expected ETA. This "reliability " information of ETA can improve the overall travel experience (Lam and Small, 2001). For operators, forecasting the probabilistic distribution of bus arrival times can be used to enhance schedule reliability. For example, dynamic bus holding strategies (Xuan et al., 2011) have been proposed to prevent bus bunching, where the model requires knowledge of the variance in trip time between stations; forecasting the real-time probabilistic distribution of trip travel time, therefore, allows for more precise control.

A critical step in probabilistic bus travel time forecasting is to construct an appropriate probabilistic distribution for bus travel time, which is very challenging due to the following difficulties: 1) there exit complex correlations among different links within a bus route, 2) there exist strong interactions between two adjacent buses (e.g., bus bunching), and 3) bus travel time distributions are usually not normal and exhibit long-tailed and multimodal characteristics (Ma et al., 2016). A recent study (Chen et al., 2022) demonstrated that link travel times in a bus route exhibit complex local and long-range correlations, and bus travel times on different links are often positively correlated because of factors like bus bunching. However, the very limited existing research on probabilistic bus travel time forecasting often overlooks the complex link travel time correlation. For example, Huang et al. (2021) and Ricard et al. (2022) did not model the link travel time correlation; Ma et al. (2017) and Büchel and Corman (2022a) only considered the (local) correlation between adjacent links; some studies applied a unimodal Gaussian assumption (Taylor, 1982; May et al., 1989; Dai et al., 2019; Chen et al., 2022), which failed to capture the bus travel time realistically. In terms of the interactions among vehicles, only a few works have considered its effects on travel time forecasting with simplified assumptions (e.g., Dai et al., 2019); the potential of leveraging the information from neighboring buses to improve the forecasting remains unknown.

To address the above issues, in this paper we develop a Bayesian probabilistic model for bus travel time and ETA forecasting. Specifically, we concatenate the link travel time vectors and the headway vector of each two adjacent buses (i.e., a pair) as a new augmented random variable. By incorporating the inherent relationship (linear equality constraints) between link travel times and headways, our approach naturally captures the interactions between a bus and its leading bus and handles missing values in the data. To capture the multimodality of bus travel time distribution, we model the augmented random variable with multivariate Gaussian mixture distributions truncated by the hyperplane defined by the headway constraints. Moreover, we borrow the idea of Bayesian hierarchical model to capture temporal differences in bus travel time: different periods in a day share the same set of Gaussian components but different mixing coefficients. Next, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm to obtain the posterior distribution of model parameters. Based on the estimated probabilistic model, we could make conditional probabilistic forecasting for bus travel time in an autoregressive way (the forecasting of a bus relies on the forecasting of the its leading bus). We test the proposed probabilistic forecasting model using a dataset from two bus lines in Guangzhou, China. Results show our approach that considers the dependencies between adjacent buses and the headway relationships significantly outperforms baseline models that overlook these factors, in terms of both predictive means and distributions. Besides forecasting, the parameters of the proposed model contain rich information for understanding/improving the bus service, e.g., analyzing link travel time correlation using correlation matrices and understanding temporal patterns of the bus route from mixing coefficients.

The contributions of this work include three aspects. First, we propose a forecasting model that takes into account and handles various difficulties in modeling bus travel time, including correlations between link travel time on a bus route, interactions between adjacent buses, and missing values in data. These parameters are embedded in tailored augmented random variables. Second, we integrate Gaussian mixture models into a Bayesian hierarchical model framework as the distribution for the augmented random variable. The Gaussian mixture models can depict the multimodality in bus travel time distributions, and the hierarchical structure can reflect different mixing patterns in bus travel time at different periods of a day. Third, we develop a Gibbs sampling algorithm to obtain the posterior distributions of model parameters and enable probabilistic forecasting. Experiments in a real-world dataset show the proposed model can accurately forecast the bus travel time distributions and discover patterns in link travel time correlations.

The remainder of this paper is organized as follows. In Section 4.3, we review previous studies on bus travel time forecasting. In Section 4.4, we describe the problem and present the Bayesian Gaussian mixture model. Next, in Section 4.5, we demonstrate the capability of the proposed model by numerical experiments using real-world data. Finally, we conclude our study, summarize our main findings, and discuss future research directions in Section 4.6.

4.3 Related Work

There exists a large body of studies in the literature on bus travel time forecasting, of which most concentrate on providing point estimation the travel time values instead of forecasting the travel time distribution. In this literature review, we categorize bus travel time forecasting models into deterministic forecasting models and probabilistic forecasting models. After the introduction of deterministic forecasting models, we pay more attention on the studies related to probabilistic forecasting models for bus travel time.

4.3.1 Deterministic Forecasting Model

Deterministic forecasting model for bus travel time includes time series models and machine learning approaches. Time series models often consider bus travel time as a function of its past observations, and one important step is to construct a standard time series of bus travel time from bus data. Farhan et al. (2002) used HA model to forecast the bus travel time. Madzlan et al. (2010) applied ARIMA model to forecast bus route travel time. The forecasting accuracy of time series model highly depends on the characteristics of historical travel time data. If a time series has a stationary pattern or mixture of patterns, time series models can perform well. However, bus travel time is often more complicated due to the factors like road congestion, dynamic passenger flow, traffic accidents, etc. Therefore, the performance of time series models is often limited for bus travel time forecasting.

On the other hand, machine learning approaches can characterize complex patterns and learn non-linear relationships from data. Existing machine learning models to bus travel time forecasting include ANN, SVM, KF, KNN, and hybrid models. Chien et al. (2002) developed two ANN-based models to forecast bus arrival time under link-based and stop-based route constructions; results show that these two ANN-based models have good performance. Similarly, the studies by Gurmu and Fan (2014), Jeong and Rilett (2004) indicate that ANN-based models outperform HA and regression models for bus travel time forecasting. SVM has been applied as a useful method for forecasting bus travel time. Some studies (Bin et al., 2006; Yu et al., 2011; Yang et al., 2016; Xu and Ying, 2017; Ma et al., 2019) have shown SVM outperforms linear regression and time series models. KF is an efficient learning algorithm as it has the ability to update the time-dependent state when new observations become available continuously. Cathey and Dailey (2003) proposed KF model for bus travel time forecasting using real-time and historical data, while this work did not compare KF model with other developed machine learning models. Achar et al. (2019) developed a data-driven method to forecast bus travel time; first, they learned the spatial and temporal correlations/patterns of traffic; then, they used KF to complete the forecasting task. The authors compared their proposed model with HA and Random Forest (RF) and found that their method performed better. Building on this work, Achar et al. (2022) transformed travel time forecasting into a hidden-state estimation problem using a related non-linear dynamical system model. They proposed a solution based on KF and Particle Filter. Liu et al. (2012) adopted KNN to make the bus travel time forecasting using historical bus GPS data, and the results indicated that KNN outperformed ANN in terms of accuracy. Some hybrid methods are developed to combine the advantages of these machine learning models. For example, Yu et al. (2018) proposed an approach combining KNN and RF and found that the hybrid model outperformed KNN, SVM, and RF; however, this method has the problem of low computational efficiency. Zhang et al. (2021) developed a hybrid method to combine SVM with KF, RF and ARIMA, respectively; results show that SVM-KF outperforms other hybrid models. Chen et al. (2004) combined ANN and KF algorithm to forecast the arrival time. Based on this research, Bai et al. (2015) replaced ANN with SVM, and results show that SVM-KF performs better than ANN-KF. Kumar et al. (2018) used KNN algorithm to identify significant input variables, and then combining exponential smoothing technique with recursive estimation scheme based on Kalman Filtering method to forecast bus travel time. Kumar et al. (2017) proposed a method for forecasting bus travel time that considered both temporal and spatial variations. They used traffic stream models to rewrite the conservation of vehicles equation in terms of flow and density, and then transformed it to a partial differential equation in terms of speed. To make the forecasts and updates, they combined the Godunov scheme and Kalman filter. The proposed method outperformed historical average, regression, and ANN methods.

Deep learning models, especially the LSTM module, has been widely used in bus travel time forecasting. Liu et al. (2023) developed a KF-LSTM deep learning method to make bus travel time forecasting, and they found that the KF-LSTM model outperforms the ensemble learning methods to forecast travel time. Recent studies by Osman et al. (2021) and Alam et al. (2021) also used LSTM for bus travel time forecasting, and found that LSTM outperformed ANN, SVR, ARIMA, and HA methods. Similarly, He et al. (2020) applied LSTM for bus travel time prediction but took into consideration the impact of heterogeneous traffic patterns. Petersen et al. (2019) proposed a deep neural network with the combination of convolutional and LSTM layers for bus travel time prediction. Li et al. (2023a) proposed a novel deep-learning model based on sequence and graph embedding to forecast bus travel time. First, they used the sequence embedding part to extract the complex and potential sequence patterns from many trajectory sequences. Then, the network embedding part is to capture the spatial correlation among bus stops. Finally, a fusion prediction part is to combine sequence and network embedding vectors to make the forecasting. Li et al. (2023b) utilized the Interaction Networks to model the interactions between transit speed, dwell time, and traffic speed for arrival time prediction. Results showed that the proposed model outperformed LSTM, RNN, RF, and ensemble methods.

4.3.2 Probabilistic Forecasting Model

There are only a few studies on probabilistic forecasting for bus travel time. Dai et al. (2019) proposed a probabilistic model to estimate bus travel time considering the link running time and dwell time. The authors assumed that: the link travel time is composed of the link running time and the dwell time, and they are independent; the link running time follows shifted log-normal distribution; the stop dwell time is the sum of the queueing time, the passengers boarding/alighting time, and the merging time (the bus merges into the main road traffic). They did not model the correlations between link running time and dwell time, despite their importance for bus travel time forecasting. Yu et al. (2017) proposed an accelerated failure time survival model which could simultaneously estimate expected travel times and travel time uncertainty. In this model, the arrival of a bus to a downstream stop could be regarded as the event; they have considered using the independent variables including headway deviation, scheduled headway, onboard passenger, weather, travel time of the previous bus, and day/time period. However, this model ignores using the local and long-range correlations in bus link travel time. Ma et al. (2017) proposed a generalized Markov chain approach for estimating the probability distribution of bus trip travel times from link travel time distributions, taking the correlations in time and space into consideration. This approach first uses clustering methods to cluster the link travel time observations, and then uses a logit model to estimate the transition probability; finally, using a Markov chain procedure, the probability distribution of the trip travel time can be estimated. However, the Markov chain in this framework only models the correlation between adjacent link travel times but ignores the long-range correlation. Huang et al. (2021) proposed two data-driven methods based on Functional Data Analysis (FDA) and Bayesian Support Vector Regression (BSVR) to forecast the distribution of bus travel time. Both FDA and BSVR are essentially kernel methods. FDA approach is a well-proven mathematical way to describe the stochastic process of link travel time and can provide continuous-time link travel time forecasting, while BSVR can provide discrete-time link travel time forecasting with a prescribed discretization interval. The authors utilize the FDA and BSVR for each specific link and then add up relevant link travel times. Again, they ignore the link travel time correlation. Büchel and Corman (2022a) proposed a hidden Markov chain framework to estimate bus travel time distribution, which can capture the dependency structure of consecutive links and include conditional correlations. Moreover, it also captures the dependency of consecutive link dwell times. However, long-range correlations cannot be easily modeled in this framework. Ricard et al. (2022) proposed two types of probabilistic models: similarity-based density estimation models and a smoothed logistic regression for probabilistic classification. Similarity-based density estimation models first find the set of similar trips and then estimate the density of the particular set by fitting a parametric, semi-parametric, or non-parametric model. Multinomial logistic regression is used for probabilistic classification and its generated probability mass function can be smoothed into a probability density function. The authors developed these two methods in order to make a long-term forecast for bus travel time, thus they did not consider using the feature of link travel time. Büchel and Corman (2022b) proposed a Bayesian network approach to forecast bus travel time distribution. They assumed that: the dwell time of a given bus at a given stop depends on the dwell time of the same bus at the previous stop, dwell time of the previous bus at the same stop, and the headway from the previous bus; the running time depends on the running time of the same bus in the previous link and the running time of the previous bus in the same link. This Bayesian framework provides a nice solution to characterize the dependency between adjacent links; however, it also ignores the long-range correlations in bus link travel time. Rodriguez-Deniz and Villani (2022) developed a probabilistic real-time forecasting model for bus delay. The core of this model is the feature construction; they considered the short-term and long-term component: the short-term effect is modeled as a linear combination of several previous delays of recent buses; the long-term effect is modeled by using Gaussian process prior on parameters. This model has considered the interactions between buses as well as local correlations, but it also fails to model the long-range correlations of bus delays.

Although work on probabilistic forecasting for bus travel time is scarce, there are some studies on modeling bus travel time distributions, mostly with the objective of quantifying the bus travel time reliability (Büchel and Corman, 2020). Bus travel time distribution modeling can provide a foundation for probabilistic forecasting. Taylor (1982) collected bus data and pointed out that bus link travel time follows a normal distribution. Mazloumi et al. (2010) explored the travel time distributions for different departure time windows at different times of the day; results show that in narrower departure time windows, bus travel time distributions are best characterized by normal distributions. For wider departure time windows, peak-hour travel times follow normal distributions, while off-peak travel times follow log-normal distributions. Similarly, Rahman et al. (2018) analyzed the bus travel time distributions for different spatial horizons; results show that log-normal distribution is more appropriate for short-term horizon, while normal distribution is more suitable for long-term horizon. Uno et al. (2009) revealed that bus link travel time on arterial roadway is positively skewed and generally follows a log-normal distribution. Kieu et al. (2015) also recommended a log-normal distribution as the best fit for bus travel time on urban roads. Dhivya Bharathi et al. (2020) assumed bus travel time follows log-normal distribution and they proposed a log-normal autoregressive model to make the forecasting for bus travel time. Büchel and Corman (2018) compared the unimodal distributions including normal, Weibull, log-normal, Gamma, cauchy, and logistic distribution; results show that the log-normal probability distribution is a good fit for bus travel times at peak and off-peak conditions. Chepuri et al. (2020) compared the Burr, generalized extreme value (GEV), and log-normal distributions for bus travel time, and the results show that GEV is supervior over others. Similarly, Harsha and Mulangi (2021) considered seven travel time distributions (including Burr, GEV, Gamma, log-logistic, log-normal, normal and Weibull distributions) and evaluated their performance; results show that GEV distribution performs the best. Ma et al. (2016) compared both unimodal distributions and multimodal distributions for bus link travel time, and found that the normal, log-normal, logistic, log-logistic, and Gamma distributions have a relatively similar performance, and the Gaussian mixture model performs much better in terms of accuracy, robustness, and interpretability.

In summary, existing studies are mainly concerned with deterministic forecasting instead of probabilistic forecasting for bus travel time. Although a few studies have developed probabilistic forecasting for bus travel time, there are some limitations: (1) they ignored the complex link travel time correlation/covariance (Dai et al., 2019; Yu et al., 2017; Ma et al., 2017; Huang et al., 2021; Büchel and Corman, 2022a,b; Rodriguez-Deniz and Villani, 2022); (2) they overlooked the strong interactions between consecutive buses (Dai et al., 2019; Yu et al., 2017; Ma et al., 2017; Huang et al., 2021; Büchel and Corman, 2022a).

4.4 Methodology

4.4.1 Problem Description

A bus link (or simply a link) is the directional road segment connecting two adjacent bus stops on a bus route. In this paper, the bus travel time on the m-th link is defined as the time difference between the arrival of a bus at the m-th and the $(m+1)$-th bus stop, including the dwell time at the m-th bus stop. Link travel time of buses can be obtained from various types of data sources, such as smart card data, automatic vehicle location (AVL) data, and automatic bus announcing systems. We denote by $\ell_{i,m}$ the link travel time of the i-th bus on the m-th link. With these definitions, the trip travel time of the i-th bus from stop m_1 to stop m_2 can be readily calculated by $\sum_{m=m_1}^{m_2} \ell_{i,m}$.

This paper focuses on forecasting the travel time of a bus on its upcoming links and trips and also providing ETA distribution. A previous work (Chen et al., 2022) has shown that link travel time within a single bus trip can be significantly correlated, and using such correlation can improve bus travel time forecasting. A limitation of this work is that the dependencies of the travel time among different buses are ignored. Considering the close spatiotemporal distance and similar traffic conditions of two adjacent buses, it is tempting to use the travel time information of a leading bus to forecast the travel time of the next (following) bus. In bus systems, buses on a given route often exhibit strong interactions due to effects like bus bunching, leading to strong correlations between them. When forecasting the travel time of a single bus, we can only use information from the observed links along its path. However, by modeling adjacent buses on the same route, we can also incorporate information about upcoming links based on the leading bus. This additional information can be highly valuable for improving the accuracy of travel time forecasts. Inspired by this and on top of the previous work, in this paper we further leverage the travel time correlation between a pair of buses to improve bus travel time forecasting.

4.4.2 Augmented Random Variable

The link travel time of bus i on a bus route with n links ($n+1$ bus stops) can be aggregated into a vector $\boldsymbol{\ell}_i = [\ell_{i,1}, \ell_{i,2}, \cdots, \ell_{i,n}]^\top$. We define an augmented random variable $\boldsymbol{x}$ to capture the link travel time between two adjacent buses from the same bus line. The link travel time and the headway of each two adjacent buses (the subject bus i and its leading bus $i-1$) produce a sample of $\boldsymbol{x}$:

$$\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{\ell}_i \\ \boldsymbol{\ell}_{i-1} \\ \boldsymbol{h}_i \end{bmatrix} = [\ell_{i,1}, \ell_{i,2}, \cdots, \ell_{i,n}, \ell_{i-1,1}, \ell_{i-1,2}, \cdots, \ell_{i-1,n}, h_{i,1}, h_{i,2}, \cdots, h_{i,n}]^\top \in \mathbb{R}^{3n}, \quad (4.1)$$

where the headway $h_{i,m}$ is the time interval between the arrival of the $(i-1)$-th bus and the i-th bus at the m-th bus stop (we do not count the headway at the last bus stop—stop $(n+1)$). Note that $h_{i,\cdot}$ will become negative if overtaking happens. and there is an inherent relationship between the link travel time and the headway:

$$h_{i,m+1} - h_{i,m} + \ell_{i-1,m} - \ell_{i,m} = 0, \;\; m = 1, \ldots, n-1. \quad (4.2)$$

Therefore, the $3n$ dimensional random variable $\boldsymbol{x}$ has only $2n+1$ degrees of freedom. The inclusion of headway explicitly bonds $\boldsymbol{\ell}_i$ with $\boldsymbol{\ell}_{i-1}$.

Missing and ragged values are unavoidable in real-world link travel time data. Using the relationship between adjacent buses can enhance the accuracy of the missing value imputation. We use the method proposed by Chen et al. (2022) to jointly formulate the missing/ragged values and the headway constraints from Eq. (4.2). Consider the example in Figure 4.1 where the arrival time of bus i at stop #3 is not observed, headway $h_{i,3}$ becomes a missing value. Although individual values of $\ell_{i,2}$, $\ell_{i,3}$ are missing, the sum of the link travel time $(\ell_{i,2} + \ell_{i,3})$ can be inferred from the bus arrival time at the upstream and the downstream stops, which is a case of ragged value. Overall, the missing/ragged

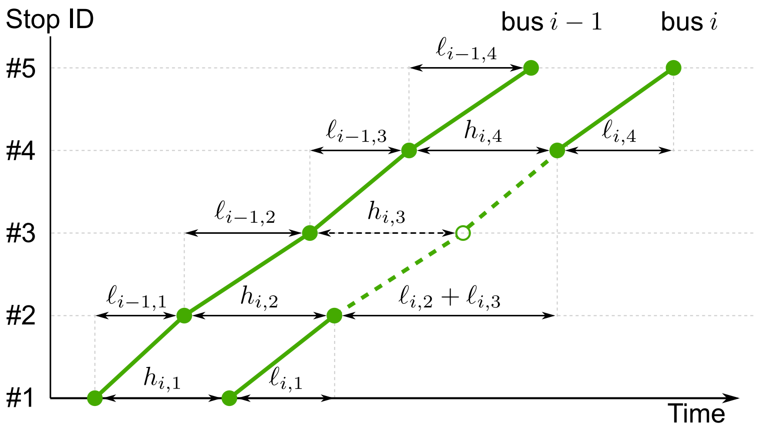


Figure 4.1: Trajectories of two adjacent buses. Solid dots: observed arrival time at a bus stop. Circle: unknown arrival time at the bus stop.

values and headway constraints can be summarized into a linear equation:

$$\underbrace{\begin{bmatrix} 1&0&0&0&0&0&0&0&0&0&0&0\\ 0&1&1&0&0&0&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&0&0&0\\ 1&0&0&0&-1&0&0&0&-1&1&0&0\\ 0&1&0&0&0&-1&0&0&0&-1&1&0\\ 0&0&1&0&0&0&-1&0&0&0&-1&1 \end{bmatrix}}_{\mathbf{G}_i} \boldsymbol{x}_i = \begin{bmatrix} \ell_{i,1}\\ \ell_{i,2}+\ell_{i,3}\\ \ell_{i,4}\\ \ell_{i-1,1}\\ \ell_{i-1,2}\\ \ell_{i-1,3}\\ \ell_{i-1,4}\\ h_{i,1}\\ h_{i,2}-h_{i,1}+\ell_{i,1}-\ell_{i-1,1}\\ h_{i,3}-h_{i,2}+\ell_{i,2}-\ell_{i-1,2}\\ h_{i,4}-h_{i,3}+\ell_{i,3}-\ell_{i-1,3} \end{bmatrix} = \underbrace{\begin{bmatrix} r_{i,1}\\ r_{i,2}\\ r_{i,3}\\ r_{i,4}\\ r_{i,5}\\ r_{i,6}\\ r_{i,7}\\ r_{i,8}\\ 0\\ 0\\ 0 \end{bmatrix}}_{\boldsymbol{r}_i}, \tag{4.3}$$

where we call $\mathbf{G}_i$ the *alignment matrix* and $\boldsymbol{r}_i$ the *recording vector* for bus i. An alignment matrix is a matrix with elements in $\{-1, 0, 1\}$ that encodes missing/ragged positions and headway constraints. The recording vector $\boldsymbol{r}_i$ records all observed information attached with $\boldsymbol{x}_i$. The hyperplane defined by Eq. (4.3) is the support of random variable $\boldsymbol{x}$.

Alignment matrices and recording vectors can be directly accessed from the source data, but the values of $\boldsymbol{x}_i$ are not always available because of the missing data problem. Next, the main task is to estimate the probability distribution (a Gaussian mixture model) of the augmented random variable $\boldsymbol{x}$ using historical alignment matrices $\{\mathbf{G}_i\}$ and recording vectors $\{\boldsymbol{r}_i\}$. Once obtaining the probabilistic distribution of $\boldsymbol{x}$, we can forecast the bus travel time on upcoming links by calculating the conditional probability given the travel time of the leading buses and upstream links. The detailed forecasting procedure will be described in Section 4.4.5.

4.4.3 Bayesian Multivariate Gaussian Mixture Model

The distributions of bus link travel times are often positively skewed, heavy-tailed, and sometimes multimodal. Ma et al. (2016) compared several unimodal distributions (including normal, log-normal, logistic, log-logistic, and Gamma distributions) and multimodal distributions for bus travel time, and suggested using Gaussian mixture models for bus link travel times. In this paper, we also use multivariate Gaussian mixture distributions to model the augmented random variable $\boldsymbol{x}$. There are four advantages of the Bayesian Gaussian mixture model: 1) the Gaussian mixture distribution is well-suited for modeling multi-modality and can approximate complex distributions. Figure 4.12 shows the empirical distribution of link travel time. We observed that while many links exhibit positively skewed and unimodal distributions, such as link #11 and link #20, have bimodal distributions. This justifies the use of the Gaussian mixture model to approximate the skewed and multi-modal distributions. 2) The Gaussian distribution has conjugate prior distributions, which is convenient for us to derive the posterior distribution analytically. 3) The bus travel time problem has many missing/ragged values. However, by using the sampling scheme from hyper-plane truncated Gaussian distribution, we can convert the data imputation problem to a Gaussian distribution problem, which has a nice property that enables us to handle missing/ragged values in the data. 4) We want to make conditional forecasting for bus travel time, and Gaussian distribution has an excellent property that the unobserved part conditional on the observed part is also Gaussian distribution. Moreover, we divide a day into T periods and assume the mixing coefficients are different for each period and use a hierarchical framework to capture the temporal differences.

When not considering the headway constraints in Eq. (4.3), the augmented random variable at the t-th period follows a multivariate Gaussian mixture model:

$$p^t\left(\boldsymbol{x}^t\right)=\sum_{k=1}^{K}\pi_k^t\mathcal{N}\left(\boldsymbol{x}^t\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right), \tag{4.4}$$

where we use a superscript $(\cdot)^t$ to denote the time period, K is the number of components, $\pi_k^t>0$ is a mixing coefficient with $\sum_{k=1}^{K}\pi_k^t=1$, and each of the K components follows a multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}_k\in\mathbb{R}^{3n}$ and a $3n\times 3n$ covariance matrix $\mathbf{\Sigma}_k$. When considering the linear constraints in Eq. (4.3), the distribution of $\boldsymbol{x}^t$ in Eq. (4.4) becomes a hyperplane-truncated multivariate Gaussian mixture (Cong et al., 2017; Chen et al., 2022).

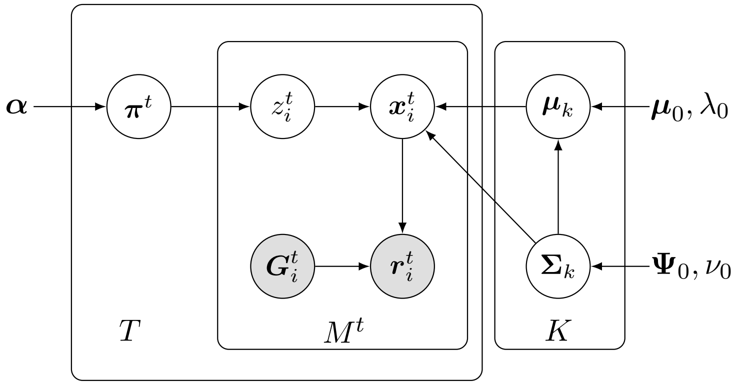


Figure 4.2: The graphical illustration of Bayesian Gaussian mixture model.

The augmented link travel time of each period is characterized by a mixture of several shared Gaussian distributions. Figure 4.2 shows the overall graphical representation of our hierarchical Bayesian multivariate Gaussian mixture model. Assume there are M^t buses at period t, we have $\{\boldsymbol{x}_i^t\}_{t=1,i=1}^{T,M^t}$ to be a set of "latent" realizations/samples drawn from the multivariate Gaussian mixture distributions truncated on the hyperplanes from Eq. (4.3). We need to estimate parameters $\{\boldsymbol{\pi}^t = [\pi_1^t, \pi_2^t, \cdots, \pi_K^t]^\top\}_{t=1}^T$, $\{\boldsymbol{\mu}_k\}_{k=1}^K$, and $\{\mathbf{\Sigma}_k\}_{k=1}^K$ using the alignment matrices $\mathcal{G} = \{\mathbf{G}_i^t\}_{t=1,i=1}^{T,M_t}$ and the recording vectors $\mathcal{R} = \{\boldsymbol{r}_i^t\}_{t=1,i=1}^{T,M_t}$. In the graphical model, z_i^t is a component label, indicating the component $\boldsymbol{x}_i^t$ belongs to. In a Bayesian setting, we use a conjugate Gaussian-inverse-Wishart prior on $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ and a Dirichlet prior on $\boldsymbol{\pi}^t$ for efficient inference (Gelman et al., 2013). The overall data generation process is summarized as:

$$\boldsymbol{\pi}^t \sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right), \tag{4.5}$$

$$\mathbf{\Sigma}_k \sim \mathcal{W}^{-1}\left(\mathbf{\Psi}_0, \nu_0\right), \tag{4.6}$$

$$\boldsymbol{\mu}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda_0}\mathbf{\Sigma}_k\right), \tag{4.7}$$

$$z_i^t \sim \text{Categorical}\left(\boldsymbol{\pi}^t\right), \tag{4.8}$$

$$\boldsymbol{x}_i^t \mid z_i^t = k \sim \mathcal{N}\left(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right), \tag{4.9}$$

$$\boldsymbol{r}_i^t = \mathbf{G}_i^t \boldsymbol{x}_i^t, \tag{4.10}$$

where $\boldsymbol{\alpha}$ is the concentration parameter of the Dirichlet distribution; $\mathcal{W}^{-1}\left(\mathbf{\Psi}_0, \nu_0\right)$ is the inverse-Wishart distribution with a scale matrix $\mathbf{\Psi}_0$ and ν_0 degrees of freedom; $\boldsymbol{\mu}_0$ and λ_0 are parameters for the Gaussian prior.

4.4.4 Model Inference

Based on the graphical model illustrated in Figure 4.2, we can derive an efficient MCMC scheme using Gibbs sampling. For simplicity, we let $\mathcal{X} = \{\boldsymbol{x}_i^t\}_{t=1,i=1}^{T,M^t}$ denote the full set of the "latent" augmented variable for the $\sum_{t=1}^{T} M^t$ bus pairs, and let $\Theta = \{\boldsymbol{\mu}_0, \lambda_0, \mathbf{\Psi}_0, \nu_0\}$ denote the set of hyperparameters for the Gaussian-inverse-Wishart prior distribution in Eqs. (4.6) and (4.7). In addition, we denote by $\mathcal{X}_k^t$ the set of data vectors at period t belonging to mixture component k and by $\mathcal{X}_k$ the set of all data vectors belonging to mixture k. We start the Gibbs sampling with random initialization for all variables and then iteratively sample each variable from its conditional distribution on other variables:

- Sample $\boldsymbol{\pi}_t$ from $p\left(\boldsymbol{\pi}^t \mid \boldsymbol{z}^t, \boldsymbol{\alpha}\right)$. The conditional distribution is

$$p\left(\boldsymbol{\pi}^t \mid \boldsymbol{z}^t, \boldsymbol{\alpha}\right) \propto p\left(\boldsymbol{\pi}^t \mid \boldsymbol{\alpha}\right) p\left(\boldsymbol{z}^t \mid \boldsymbol{\pi}^t\right). \tag{4.11}$$

  The prior distribution $p\left(\boldsymbol{\pi}^t \mid \boldsymbol{\alpha}\right) = \text{Dirichlet}\left(\boldsymbol{\pi}^t \mid \boldsymbol{\alpha}\right) \propto \prod_{k=1}^{K} {\pi_k^t}^{\alpha_k - 1}$, and $p\left(\boldsymbol{z}^t \mid \boldsymbol{\pi}^t\right)$ can be seen as a multinomial distribution

$$p\left(\boldsymbol{z}^t \mid \boldsymbol{\pi}^t\right) = \text{Multinomial}_K\left(\boldsymbol{z}^t \mid N, \boldsymbol{\pi}^t\right) \propto \prod_{k=1}^{K} {\pi_k^t}^{M_k^t} \tag{4.12}$$

  , where M_k^t is the number of $\{z_i^t\}_{i=1}^{M^t}$ assigned to class k. Therefore, the conditional posterior distribution is a Dirichlet distribution:

$$p\left(\boldsymbol{\pi}^t \mid \boldsymbol{z}^t, \boldsymbol{\alpha}\right) \sim \text{Dirichlet}\left(M_1^t + \alpha_1, M_2^t + \alpha_2, \cdots, M_K^t + \alpha_K\right). \tag{4.13}$$

- Sample z_i^t from $p\left(z_i^t \mid \boldsymbol{\pi}_i^t, \boldsymbol{\mu}, \mathbf{\Sigma}, \boldsymbol{x}_i^t\right)$. The conditional distribution of assignment for each observation is given by $p\left(z_i^t = k \mid \boldsymbol{\pi}_i^t, \boldsymbol{\mu}, \mathbf{\Sigma}, \boldsymbol{x}_i^t\right) \propto p\left(z_i^t = k \mid \boldsymbol{\pi}^t\right) p\left(\boldsymbol{x}_i^t \mid \boldsymbol{\mu}, \mathbf{\Sigma}\right) = \pi_k^t \mathcal{N}\left(\boldsymbol{x}_i^t \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right)$. Then we can do the normalization to obtain the categorical conditional distribution over z_i^t by

$$p\left(z_i^t = k \mid \boldsymbol{\pi}^t, \boldsymbol{\mu}, \mathbf{\Sigma}, \boldsymbol{x}_i^t\right) = \frac{\pi_k^t \mathcal{N}\left(\boldsymbol{x}_i^t \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right)}{\sum_{m=1}^{K} \pi_m^t \mathcal{N}\left(\boldsymbol{x}_i^t \mid \boldsymbol{\mu}_m, \mathbf{\Sigma}_m\right)}. \tag{4.14}$$

- Sample $(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ from $p\left(\boldsymbol{\mu}, \mathbf{\Sigma} \mid \mathcal{X}_k, \Theta\right)$. Thanks to the conjugate prior distribution, the conditional distribution of the mean vector and the covariance matrix $p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathcal{X}_k, \Theta\right)$ is a Gaussian-inverse-Wishart distribution:

$$p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathcal{X}_k, \Theta\right) \sim \mathcal{N}\left(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0^*, \frac{1}{\lambda_0^*} \boldsymbol{\Sigma}_k\right) \mathcal{W}^{-1}\left(\boldsymbol{\Sigma}_k \mid \boldsymbol{\Psi}_0^*, \nu_0^*\right), \tag{4.15}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_0^* &= \frac{\lambda_0 \boldsymbol{\mu}_0 + M_k \bar{\boldsymbol{x}}}{\lambda_0 + M_k}, \qquad \lambda_0^* = \lambda_0 + M_k, \qquad \nu_0^* = \nu_0 + M_k, \\
\bar{\boldsymbol{x}} &= \frac{1}{M_k} \sum_{t=1}^{T} \sum_{i=1}^{M_k^t} \boldsymbol{x}_i^t, \quad M_k = \sum_{t=1}^{T} M_k^t, \quad \mathbf{S} = \sum_{t=1}^{T} \sum_{i=1}^{M_k^t} \left(\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}\right)^{\top}, \\
\boldsymbol{\Psi}_0^* &= \boldsymbol{\Psi}_0 + \mathbf{S} + \frac{\lambda_0 M_k}{\lambda_0 + M_k} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_{\mathbf{0}}\right)\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_{\mathbf{0}}\right)^{\top}.
\end{aligned} \tag{4.16}$$

- Sample $\mathcal{X}$ from $p\left(\mathcal{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{z}, \mathcal{R}, \mathcal{G}\right)$. At this step, we no longer have a simple analytical formulation to sample $\mathcal{X}$ due to the linear constraints in Eq. (4.10). Here, the conditional distribution can be factorized as

  $$p\left(\mathcal{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{z}, \mathcal{R}, \mathcal{G}\right) = \prod_{t=1}^{T} \prod_{i=1}^{M^t} p\left(\boldsymbol{x}_i^t \mid \boldsymbol{\mu}_{z_i^t}, \boldsymbol{\Sigma}_{z_i^t}, \boldsymbol{r}_i^t, \mathbf{G}_i^t\right). \tag{4.17}$$

  Consequently, we can draw sample of the bus-pair vector $\boldsymbol{x}_i^t$ independently. The conditional distribution of $\boldsymbol{x}_i^t$ in Eq. (4.17) can be regarded as a multivariate Gaussian distribution truncated on the intersection with a hyperplane, i.e.,

  $$\boldsymbol{x}_i^t \,|\, z_i^t = k \sim \mathcal{N}_{\mathcal{S}_i^t}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \quad \mathcal{S}_i^t = \left\{\boldsymbol{x}_i^t \mid \mathbf{G}_i^t \boldsymbol{x}_i^t = \boldsymbol{r}_i^t\right\}. \tag{4.18}$$

  The probability density function of the hyperplane-truncated multivariate Gaussian is

  $$p(\boldsymbol{x}_i^t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{r}_i^t, \mathbf{G}_i^t) = \frac{1}{Z_i^t} \exp\left[-\frac{1}{2}(\boldsymbol{x}_i^t - \boldsymbol{\mu}_{z_i^t})^T \boldsymbol{\Sigma}_{z_i^t}^{-1} (\boldsymbol{x}_i^t - \boldsymbol{\mu}_{z_i^t})\right] \delta(\mathbf{G}_i^t \boldsymbol{x}_i^t = \boldsymbol{r}_i^t), \tag{4.19}$$

  where Z_i^t is a normalizing constant; $\delta(*)$ is a function whose value is 1 if the condition $*$ holds, and 0 otherwise. Similar to Chen et al. (2022), we adopt a fast sampling scheme (Algorithm 3) developed by Cong et al. (2017) for this problem.

Finally, we summarize the Gibbs sampling procedure for estimating the parameters in Algorithm 4. We drop the first d_1 iterations as burn-in and then store samples of parameters

Algorithm 3 Sampling from a hyperplane-truncated multivariate Gaussian distribution (Cong et al., 2017).

```
1: Sample $\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{\mu}_{z_i^t}, \boldsymbol{\Sigma}_{z_i^t}\right)$;
2: Return $\boldsymbol{x}_i^t = \boldsymbol{u} + \boldsymbol{\Sigma}_{z_i^t}\mathbf{G}_i^{t\top}\left(\mathbf{G}_i^t\boldsymbol{\Sigma}_{z_i^t}\mathbf{G}_i^t\right)^{-1}\left(\boldsymbol{r}_i^t - \mathbf{G}_i^t\boldsymbol{u}\right)$, which can be more efficiently and accurately calculated by
    • Solve $\boldsymbol{\beta}$ such that $\left(\mathbf{G}_i^t\boldsymbol{\Sigma}_{z_i^t}\mathbf{G}_i^{t\top}\right)\boldsymbol{\beta} = \boldsymbol{r}_i^t - \mathbf{G}_i^t\boldsymbol{u}$;
    • Return $\boldsymbol{x}_i^t = \boldsymbol{u} + \boldsymbol{\Sigma}_{z_i^t}\mathbf{G}_i^{t\top}\boldsymbol{\beta}$.
```

$\boldsymbol{\pi}^t$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ from the following d_2 iterations. In particular, these stored samples mixing coefficients $\left\{\boldsymbol{\pi}^{t(\rho)}\right\}_{\rho=1}^{d_2}$, mean vectors $\left\{\boldsymbol{\mu}_k^{(\rho)}\right\}_{\rho=1}^{d_2}$, and covariance matrices $\left\{\boldsymbol{\Sigma}_k^{(\rho)}\right\}_{\rho=1}^{d_2}$ are critical ingredients for deriving the posterior distribution of the parameters and performing probabilistic forecasting of bus travel time. For hyperparameters, we set $\boldsymbol{\mu}_0 = \mathbf{0}_{3n}$, $\lambda_0 = 10$, $\boldsymbol{\Phi}_0 = \boldsymbol{I}_{3n}$, $\nu_0 = 3n + 2$, and $\boldsymbol{\alpha} = \mathbf{0.2}_K$, where n is the number of bus links. It should be noted that model training is in fact offline based on historical data, and only the Markov chains (i.e., samples) of the parameters are used in the forecasting task.

4.4.5 Probabilistic Forecasting

We divide the links of each bus into observed and upcoming links at the time of making forecasting. Observed links are passed links with known travel times, and upcoming links are the links that the bus is yet to pass and for which we need to forecast travel times. Because of the Gaussian assumption in each mixture component, we can easily forecast the bus travel time on upcoming links by calculating the conditional distribution given observed link travel times and headways. Generally, there are upcoming links for both the following bus j and the leading bus $(j-1)$; although it is possible to make forecasting conditioning on only the observed links, we adopt an autoregressive approach that also uses the forecasting of the leading bus's upcoming links to forecast the bus travel time of the following bus, because this autoregressive approach uses the information of bus $(j-2)$—the leading bus of $(j-1)$—to reinforce the forecasting. For the first bus (without a leading bus) of a day, we only use observed links to forecast the upcoming links (like the method proposed by Chen et al. (2022)).

We use Figure 4.3 to illustrate the forecasting process. In Figure 4.3(a), bus $j-1$ has just finished the run; bus j has passed the first two links and arrived at stop #3; bus $j+1$

Algorithm 4 Gibbs sampling for parameter estimation.

```
Input: Recording vectors $\mathcal{R}$, alignment matrices $\mathcal{G}$, hyperparameters $\Theta$ and $\boldsymbol{\alpha}$, iterations $d_1, d_2$.
Output: Samples of mixture weights $\left\{\boldsymbol{\pi}^{t\,(\rho)}\right\}_{\rho=1}^{d_2}$, samples of mean vectors $\left\{\boldsymbol{\mu}_k^{(\rho)}\right\}_{\rho=1}^{d_2}$, and samples of covariance matrices $\left\{\boldsymbol{\Sigma}_k^{(\rho)}\right\}_{\rho=1}^{d_2}$.
for iter $= 1$ to $d_1 + d_2$ do
    for $t = 1$ to $T$ do
        Draw $\boldsymbol{\pi}^t$ according to Eq. (4.5).
        if iter $> d_1$ then
            Collect $\boldsymbol{\pi}^t$ to the output set.
        end if
    end for
    for $k = 1$ to $K$ do
        Draw $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ according to Eq. (4.6) and Eq. (4.7).
        if iter $> d_1$ then
            Collect $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ to the output sets.
        end if
    end for
    for $t = 1$ to $T$ do
        for $s = 1$ to $M_t$ do
            Calculate p($z_s^t$) according to Eq. (4.14).
            Draw $z_s^t$ according to Eq. (4.8).
            Draw $\boldsymbol{x}_s^t$ by Algorithm 3.
        end for
    end for
    for $k = 1$ to $K$ do
        Update the parameters $\Theta = \{\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0\}$ by Eq. (4.16).
    end for
    Update the parameters $\boldsymbol{\alpha}$ by Eq. (4.13).
end for
return $\left\{\boldsymbol{\pi}^{t\,(\rho)}\right\}_{\rho=1}^{d_2}, \left\{\boldsymbol{\mu}_k^{(\rho)}\right\}_{\rho=1}^{d_2}, \left\{\boldsymbol{\Sigma}_k^{(\rho)}\right\}_{\rho=1}^{d_2}$.
```

has departed from the origin stop but does not arrive at stop #2. We would like to use the observed links/headways to make forecasting for the upcoming links of bus j and bus $j+1$. For bus j, we can make forecasting by using its upstream links (the first two links), all the observed link travel times of the leading bus $j-1$, and the corresponding observed headways (the first three headways). In terms of bus $j+1$, we could use the observed upstream link travel times of bus j, the forecasts of the upcoming link travel times of bus j, and the observed headway. As time passes by, related buses could update the forecasting once having new observed links. At the time point shown in Figure 4.3(b), bus $j+1$ gets a new observed link; thus we could update the forecasting of bus $j+1$. At the next time point shown in Figure 4.3(c), bus j gets a new observed link, allowing itself to update the forecasting. Note that in this case, although bus $j+1$ does not have any additionally observed links, the updated information of its leading bus (i.e., bus j) could reinforce/enhance its forecasting. In addition, we can see a new bus $j+2$ begins to run on the route; with the observed headway, we could make forecasting for bus $j+2$. In Figure 4.3(d), bus j has finished the run; bus $j+1$ and $j+2$ get a new observed link, respectively. In this case, we could make forecasting for bus $j+1$ and $j+2$ following the same aforementioned procedure. In this paper, our model will concentrate on providing probabilistic forecasting for the upcoming travel times of links for buses that have already commenced their journeys. While our model does possess the capability to forecast the travel times of upcoming links for buses that have not yet embarked on their journeys, we will not be evaluating its performance for such trips in this study since those trips have no observed information.

For the j-th bus pair with a following bus started at the t-th period in a day, assume at current time point t^* we can observe the first p links of the leading bus, the first q links of the following bus, and the corresponding headways, then we aim to forecast $\hat{\boldsymbol{x}}_j^{t,t^*}$ for the upcoming links of the following bus. Apart from the observed information, we also combine the forecasts $\hat{\boldsymbol{x}}_{j-1}^{t_l,t^*}$ of the leading bus to construct the recording vector $\boldsymbol{r}_j^{t,t*}$ and alignment matrix $\mathbf{G}_j^{t,t^*}$. The posterior predictive distribution of the augmented random variable $\boldsymbol{x}_j^{t,t^*}$ is

$$\begin{aligned} &p(\boldsymbol{x}_j^{t,t^*} \mid \boldsymbol{r}_j^{t,t^*}, \mathbf{G}_j^{t,t^*}) = \\ &\iiiint p(\boldsymbol{x}_j^{t,t^*} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{z}^t, \boldsymbol{r}_j^{t,t^*}, \mathbf{G}_j^{t,t^*}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \Theta) p(\boldsymbol{z}^t \mid \boldsymbol{\pi}^t) p(\boldsymbol{\pi}^t \mid \boldsymbol{\alpha}) \, d\boldsymbol{\mu} \, d\boldsymbol{\Sigma} \, d\boldsymbol{z}^t \, d\boldsymbol{\pi}^t. \end{aligned} \tag{4.20}$$

From the joint distribution, we can easily obtain the posterior predictive distribution $p\left(\hat{\boldsymbol{x}}_j^{t,t^*}\right)$ over the upcoming $(n-q)$ links by conditional sampling. Since we have gathered

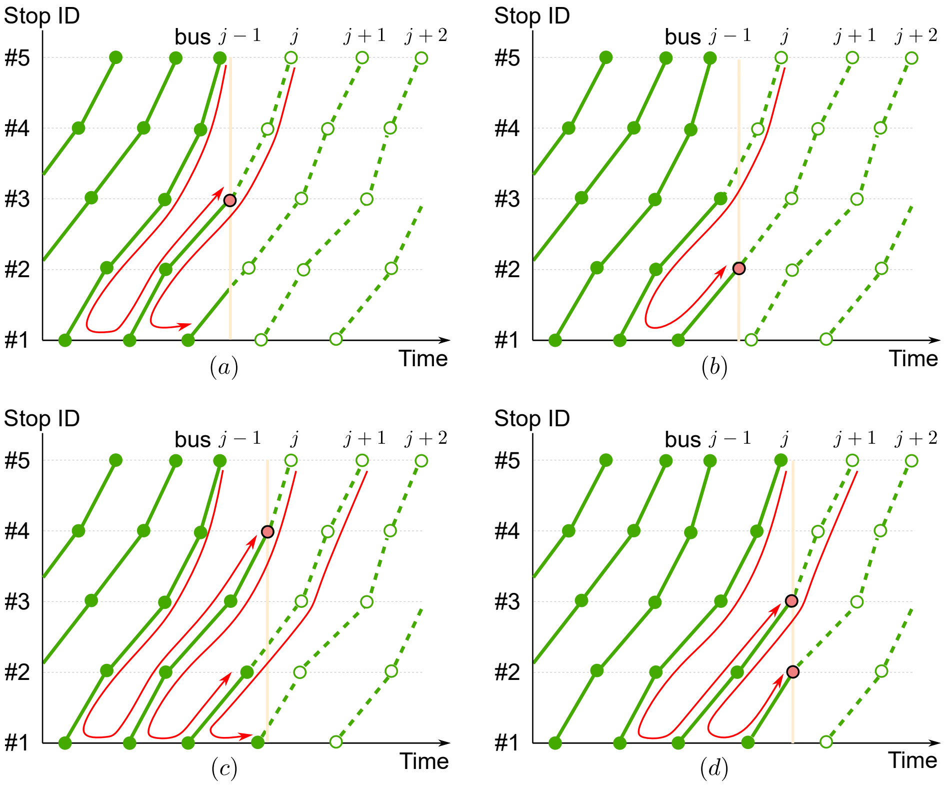


Figure 4.3: Illustration of travel time forecasting. Solid dots: observed arrival time at bus stops. The hollow dot: unknown arrival time at the bus stop. Solid red dots: new observed arrival time at bus stops. Each sub-figure is a new round of forecasting triggered by a new observation of bus arrival. For each adjacent bus pair, we forecast the following bus's upcoming link travel time (dashed green lines) based on observed and previously forecast link travel times (green lines covered by a red arrow area).

a large number of parameter samples in model inference/training, we could directly use the stored samples to make probabilistic forecasting without retraining. We summarize the procedure of probabilistic forecasting in Algorithm 5.

4.5 Experiments

4.5.1 Data and Experimental Settings

In this section, we evaluate the proposed probabilistic forecasting model on real-world data. The data used in this paper are the bus in-out-stop record data collected by the automatic bus announcing system in Guangzhou, China, during the weekdays from December 1st, 2016 to December 31st, 2016. When a bus enters or exits a bus stop, the announcing system reports the arrival or departure information and records the timestamp accordingly. Thus

Algorithm 5 Gibbs sampling for probabilistic forecasting.

```
Input: Observed vector $\boldsymbol{r}_j^{t,t*}$, alignment matrices $\mathbf{G}_j^{t,t*}$, samples of mixture weights $\left\{\boldsymbol{\pi}^{t(\rho)}\right\}_{\rho=1}^{d_2}$, samples of covariance matrices $\left\{\boldsymbol{\Sigma}_k^{(\rho)}\right\}_{\rho=1}^{d_2}$, samples of mean vectors $\left\{\boldsymbol{\mu}_k^{(\rho)}\right\}_{\rho=1}^{d_2}$.
Output: A set of samples for the forecast $\hat{\boldsymbol{x}}_j^{t,t*}$.
for $\rho = 1$ to $d_2$ do
    Compute $p(z_\rho^t)$ according to Eq. (4.14).
    Draw $z_\rho^t$ according to Eq. (4.8).
    Draw $\boldsymbol{x}_j^{t,t*}$ by Algorithm 3.
    Collect $\hat{\boldsymbol{x}}_j^{t,t*}$ to the output set.
end for
return $\left\{\hat{\boldsymbol{x}}_j^{t,t*(\rho)}\right\}_{\rho=1}^{d_2}$.
Get the posterior predictive distributions from samples $\left\{\hat{\boldsymbol{x}}_j^{t,t*(\rho)}\right\}_{\rho=1}^{d_2}$.
```

we can easily obtain the link travel times/headways from the data. We take the bus route No. 60 as a case study and aim to make probabilistic forecasting for travel time of this route. This bus route is in the urban area of Guangzhou and it has 21 stops and 20 links, as shown in Figure 4.4. The overview of data is shown in Figure 4.5. We can see that the bus route has many missing and ragged values. Moreover, Figure 4.12 shows the empirical distributions of link travel times of route No. 60. We can see that many link travel times exhibit positively skewed and unimodal distributions while some links such as link #11 and link #20 have bimodal distributions, which further justifies the use the Gaussian mixture model to approximate the skewed and multimodal distributions.

As the measurements have different units, we first perform data standardization (z-score normalization) so that all variables are centered around 0 with a standard deviation of 1. By doing so, we can better model and learn the covariance matrix. For example, $\ell_{i,m}$ (the m-th link travel time of the i-th bus) can be rescaled/standardized with

$$\tilde{\ell}_{i,m} = \frac{\ell_{i,m} - \mu_{\ell_m}}{\sigma_{\ell_m}}, \tag{4.21}$$

where μ_{ℓ_m} is the mean of travel time at the m-th link; σ_{ℓ_m} is the standard deviation of travel time at the m-th link. Note that the constraints in Eq. (4.2) no longer hold after the

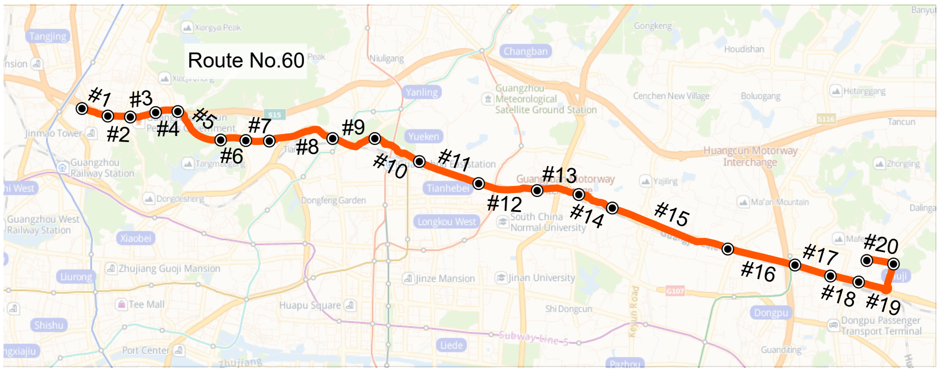


Figure 4.4: Bus route No. 60 in Guangzhou bus network.

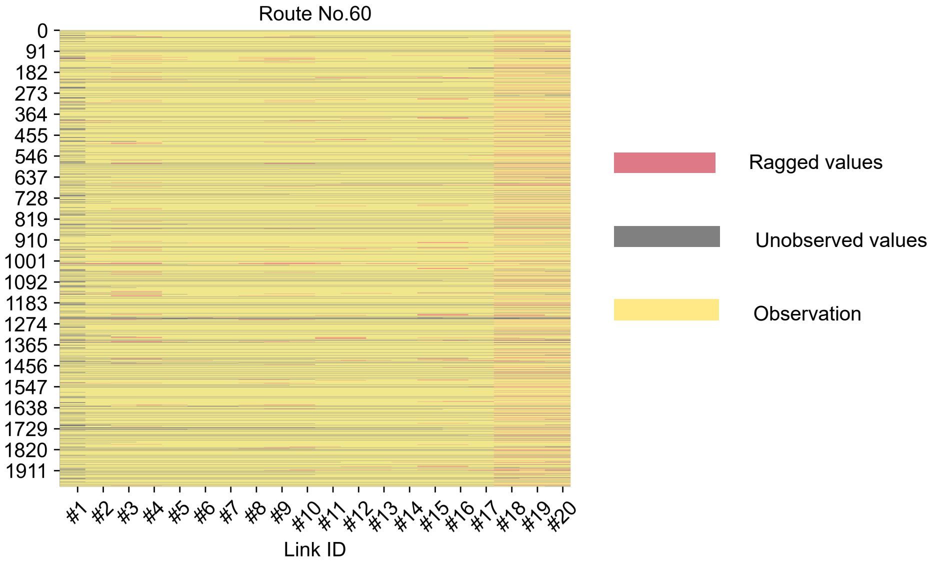


Figure 4.5: Data overview.

standardization (e.g., $\tilde{\ell}_{i,2} + \tilde{\ell}_{i,3} \neq r_{i,2}, \tilde{h}_{i,3} - \tilde{h}_{i,2} + \tilde{\ell}_{i,2} - \tilde{\ell}_{i-1,2} \neq 0$). Therefore, we need to perform a linear transform on the constraints to ensure the equality.

As we estimate the model parameters by using standardized data, the forecasts could be recovered by an inverse transformation $\hat{\ell}_{i,m} = \hat{\tilde{\ell}}_{i,m}\sigma_{\ell_m} + \mu_{\ell_m}$, where $\hat{\tilde{\ell}}_{i,m}$ is the forecast of standardizing data. We use the first sixteen weekdays (December 1st to December 23rd) for model estimation/training, and the following five weekdays (December 26th to December 30th) to test the forecasting model. In this experiment, we make probabilistic forecasting for the link/trip travel time to evaluate the proposed model. The numbers of MCMC iterations are $d_1 = 9000$ (burn-in iterations) and $d_2 = 1000$ (sampling iterations), respectively. In our experiments, the offline training process (Algorithm 4) takes approximately thirty minutes on a personal computer, while the forecasting time (Algorithm 5) for a single bus trip is less than one second, which is acceptable for real-time applications.

4.5.2 Performance Metrics

We use the root mean squared error (RMSE) and the mean absolute percentage error (MAPE) to evaluate the performance for point estimation based on the mean. We use the logarithmic score (LogS) and the continuous rank probability score (CRPS) as the metrics to evaluate the performance of probabilistic forecasting.

- RMSE and MAPE are defined as:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \\ \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \end{aligned} \tag{4.22}$$

  where $y_i, \hat{y}_i, i = 1, \ldots, n$ are the true values and forecasts, respectively.

- The logarithmic score refers to likelihood and is formally defined as

$$\text{LogS}(f_X, y) = -\log f_X(y), \tag{4.23}$$

  where f_X is the forecasting probability density function (PDF), and y is the observation. LogS is equivalent to the log-likelihood of the forecasting probability distribution and it captures all possible information about the observations related the model. Here, we use the average LogS of all observations to evaluate the model.

- The continuous rank probability score is often used as a quantitative measure of probabilistic forecasting; it is defined as the quadratic measure of discrepancy between the forecasting cumulative distribution function (CDF), noted F_X, and $\mathbb{I}(x \geqslant y)$, the empirical CDF of the observation y

$$\text{CRPS}(F_X, y) = \int_{-\infty}^{\infty} \left(F_X(x) - \mathbb{I}(x \geqslant y)\right)^2 dx, \tag{4.24}$$

  where $\mathbb{I}(\cdot)$ is the indicator function. We use the average CRPS of all observations as one metric.

4.5.3 Models in Comparison

We compare the following models to demonstrate the effect of the leading bus and headway on probabilistic forecasting for bus travel time.

- Model A: all buses are independent. In this case, variable $\boldsymbol{x}_i$ only contains link travel time of bus i, i.e., $\boldsymbol{x}_i = [\boldsymbol{l}_i]$. For a particular bus, the model only uses observed links of itself to make conditional probabilistic forecasting.
- Model B: adjacent buses are considered but without explicitly modeling headway variation, i.e., $\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{\ell}_i \\ \boldsymbol{\ell}_{i-1} \end{bmatrix}$.
- Model C: the proposed model that characterizes the joint behavior of a pair of adjacent buses. Variable $\boldsymbol{x}_i$ is in its full version as in Eq. (4.1).

In addition, we compare our proposed method with two benchmark methods: Bayesian neural networks and Kalman Filter. Given the presence of missing or ragged data, we perform data imputation using linear interpolation before training the models and making forecasts.

- Bayesian neural networks (BNN): To make probabilistic forecasting, we use a Bayesian version of a neural network, as proposed by Liang (2005). We consider the interactions between adjacent buses to make a fair comparison. The input features include the link travel times of the leading and following bus up to the current time, as well as the observed headway. The output is a vector that corresponds to the travel times of the following bus on all future links. Note that the input and output dimensions of a neural network are fixed. Thus, for each bus line, we train three neural networks, one for each scenario, to achieve the best performance. The input and output dimensions of the neural network in each scenario depend on the number of observed and future links, which are detailed in Section 4.5.4.
- Kalman Filter (KF): To account for the influence of link length on travel time, we begin by detrending the data, i.e., subtracting the mean travel time for each link from the raw data. We then apply a KF to the detrended data, which models the residuals and provides a way to make short-term trip travel time forecasts using an autoregressive approach. However, the long-range forecasting with KF is not well-suited to our needs. Here, the term "short-term trip travel time forecasts" refers to the ability of the KF to provide relatively accurate travel time predictions for

the next a few upcoming links. In other words, it excels at forecasting travel times for trips in the immediate future. On the other hand, the “long-range forecasting" refers to predicting the trip travel times when many upcoming links are involved. This type of forecasting becomes more difficult as the number of upcoming links increases. Additionally, the KF approach is limited in that it does not fully account for interactions between adjacent buses, which are important factors affecting travel time.

4.5.4 Forecasting Results

We apply Algorithm 4 to estimate model parameters and Algorithm 5 to make probabilistic forecasting for bus travel time. In addition to bus route No. 60, we also consider forecasting bus travel time for route No. 527, which has more stops/links (34 stops). For bus route No. 60, we test models with 1, 2, and 5 classes, and for bus route No. 527, we test models with 1, 2, 5, and 8 classes. To evaluate the impact of the number of observed links on travel time forecasting, we test route No. 60 with 5, 10, and 15 observed links, and route No. 527 with 5, 15, and 25 observed links.

The tables displaying the forecasting performance of different models are divided into two parts. The first part shows the performance for forecasting link travel time; the results for No. 60 and No. 527 are shown in Table 4.1 and Table 4.3, respectively. The second part shows the performance for forecasting trip travel time (i.e., estimated time of arrival at the final stop $n+1$), which are shown in Table 4.2 and Table 4.4. We observe that the forecasting performance of each model significantly improves with an increase in the number of observed links. In particular, the proposed model (Model C) outperforms both Model A and Model B, which demonstrates the importance of information on the leading bus and headways (i.e., interactions). This finding shows that travel time and headways of the bus pair can reinforce the probabilistic forecasting for bus travel time. We demonstrate the importance of the mixture model in characterizing bus travel time by observing the models with different numbers of classes in the Gaussian mixture model. Furthermore, our proposed Model C outperforms BNN and KF, showing its superior performance over these baseline models.

4.5.5 Interpreting Mixture Components/Classes

In this part, we use route No. 60 to show the interpretability of the model. After the above analysis, we use $K = 2$ to show the practical implication of the probabilistic forecasting

Table 4.1: Performance of different models for link travel time forecasting of route No. 60.

		Observed links											
		5 links				10 links				15 links			
		RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS
BNN		35.2	0.1491	16.28	-4.548	31.8	0.1278	14.33	-4.411	27.7	0.1157	11.93	-4.252
KF		41.5	0.1772	19.74	-4.781	35.9	0.1397	16.60	-4.575	32.0	0.1275	14.43	-4.427
Model A	$K = 1$	33.9	0.1439	15.48	-4.495	31.8	0.1274	14.54	-4.413	27.9	0.1151	13.03	-4.367
	$K = 2$	33.8	0.1436	14.90	-4.553	32.1	0.1275	14.13	-4.698	27.9	0.1175	12.55	-4.418
	$K = 5$	34.1	0.1430	14.51	-4.456	32.6	0.1252	13.53	-4.855	29.6	0.1200	11.79	-4.288
Model B	$K = 1$	33.5	0.1369	15.02	-4.451	29.7	0.1142	13.26	-4.342	30.3	0.1179	13.32	-4.344
	$K = 2$	33.7	0.1442	14.86	-4.434	29.3	0.1171	12.89	-4.303	31.1	0.1233	13.07	-4.297
	$K = 5$	34.5	0.1387	14.51	-4.411	29.7	0.1148	12.34	-4.261	31.9	0.1245	12.12	-4.220
Model C	$K = 1$	33.0	0.1341	14.49	-4.422	29.3	0.1139	12.62	-4.306	31.9	0.1187	12.78	-4.273
	$K = 2$	**29.7**	**0.1252**	**13.11**	**-4.334**	**22.0**	0.0989	10.26	**-4.164**	**17.0**	0.0918	**7.93**	**-3.970**
	$K = 5$	30.3	0.1253	13.19	-4.341	22.1	**0.0986**	**10.22**	-4.171	17.1	**0.0874**	7.97	-3.990

Best results are highlighted in bold fonts.

Table 4.2: Performance of different models for trip travel time forecasting of route No. 60.

		Observed links											
		5 links				10 links				15 links			
		RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS
BNN		198.4	0.0814	117.57	-7.056	145.3	0.0926	85.85	-6.578	75.3	0.0881	46.50	-5.942
KF		233.1	0.0910	141.52	-7.483	177.3	0.1076	106.39	-6.954	91.4	0.0992	57.51	-6.358
Model A	$K = 1$	187.8	0.0789	110.23	-6.921	132.2	0.0860	77.42	-6.411	71.5	0.0863	43.87	-5.845
	$K = 2$	188.4	0.0790	105.97	-6.751	136.3	0.0877	76.14	-6.492	70.4	0.0855	41.70	-5.764
	$K = 5$	190.1	0.0790	106.96	-6.712	137.2	0.0876	73.86	-6.320	72.9	0.0878	37.24	-5.544
Model B	$K = 1$	177.4	0.0760	102.13	-6.696	119.9	0.0762	68.51	-6.272	70.9	0.0884	42.10	-5.780
	$K = 2$	182.1	0.0801	105.01	-6.709	117.7	0.0770	67.24	-6.254	73.6	0.0911	41.71	-5.720
	$K = 5$	185.1	0.0786	102.37	-6.704	117.0	0.0740	63.98	-6.180	74.0	0.0908	36.45	-5.584
Model C	$K = 1$	171.6	0.0713	96.51	-6.594	115.9	0.0729	64.17	-6.151	75.6	0.0909	40.41	-5.647
	$K = 2$	**149.5**	**0.0686**	**83.79**	**-6.443**	87.2	0.0651	48.46	-5.865	36.0	**0.0619**	19.42	-4.943
	$K = 5$	151.6	0.0694	84.77	-6.502	**86.1**	**0.0641**	**47.83**	**-5.850**	**35.8**	0.0625	**19.38**	**-4.931**

Best results are highlighted in bold fonts.

Table 4.3: Performance of different models for link travel time forecasting of route No. 527.

		Observed links											
		5 links				15 links				25 links			
		RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS
BNN		53.4	0.1988	20.83	-10.970	49.7	0.1852	16.94	-9.477	49.1	0.1938	15.65	-9.137
KF		58.1	0.2155	24.26	-12.702	52.5	0.1985	18.02	-10.331	52.9	0.2158	16.77	-10.240
Model A	$K = 1$	52.4	0.1960	20.08	-10.617	50.7	0.1881	18.82	-9.977	49.3	0.1832	17.80	-9.441
	$K = 2$	50.5	0.1882	18.64	-9.893	48.9	0.1826	17.51	-9.289	47.8	0.1781	16.61	-8.874
	$K = 5$	51.3	0.1911	19.25	-10.191	49.1	0.1832	17.67	-9.361	48.6	0.1811	17.27	-9.178
	$K = 8$	51.6	0.1927	19.44	-10.317	50.7	0.1881	18.86	-9.974	49.1	0.1837	17.65	-9.361
Model B	$K = 1$	51.6	0.1921	19.43	-10.315	48.4	0.1801	17.14	-9.103	47.0	0.1753	16.11	-8.572
	$K = 2$	49.3	0.1839	17.73	-9.444	43.1	0.1617	13.27	-7.095	44.5	0.1664	14.24	-7.627
	$K = 5$	48.2	0.1791	16.92	-9.029	44.5	0.1668	14.24	-7.621	43.7	0.1635	13.69	-7.323
	$K = 8$	48.8	0.1827	17.46	-9.251	45.7	0.1703	15.11	-8.072	44.3	0.1658	14.12	-7.545
Model C	$K = 1$	49.1	0.1832	17.65	-9.365	46.6	0.1171	15.76	-8.530	43.3	0.1604	13.97	-7.468
	$K = 2$	46.9	0.1711	15.87	-8.534	40.3	0.1432	13.00	-6.563	40.7	0.1487	13.24	-6.835
	$K = 5$	**44.6**	**0.1648**	**14.35**	**-7.684**	38.2	0.1354	12.67	-6.026	**37.8**	**0.1303**	**12.43**	**-5.924**
	$K = 8$	44.7	0.1674	14.41	-7.700	**37.8**	**0.1287**	**12.42**	**-5.843**	38.1	0.1305	12.45	-5.970

Best results are highlighted in bold fonts.

Table 4.4: Performance of different models for trip travel time forecasting of route No. 527.

		Observed links											
		5 links				15 links				25 links			
		RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS	RMSE	MAPE	CRPS	LogS
BNN		385.7	0.1254	236.82	-11.765	204.0	0.1161	108.15	-9.323	116.7	0.1042	46.57	-8.512
KF		401.5	0.1321	241.58	-11.781	216.6	0.1277	115.25	-9.970	130.3	0.1197	52.66	-8.984
Model A	$K = 1$	383.5	0.1205	233.92	-11.029	205.4	0.1187	109.21	-9.339	121.0	0.1107	50.14	-8.733
	$K = 2$	380.1	0.1200	231.55	-10.991	203.7	0.1164	108.03	-9.317	118.3	0.1101	48.25	–8.515
	$K = 5$	381.7	0.1202	232.69	-11.003	204.2	0.1167	108.39	-9.328	120.7	0.1106	49.93	-8.537
	$K = 8$	381.8	0.1204	232.70	-11.001	204.6	0.1169	108.67	-9.325	120.5	0.1106	49.76	-8.532
Model B	$K = 1$	372.4	0.1198	226.16	-10.910	197.6	0.1062	103.71	-9.263	113.8	0.1001	45.06	-8.471
	$K = 2$	368.2	0.1190	223.28	-10.872	192.5	0.106	100.15	-9.210	109.9	0.0998	42.38	-8.436
	$K = 5$	363.4	0.1188	219.91	-10.837	193.8	0.1061	101.01	-9.229	113.4	0.1000	44.71	-8.469
	$K = 8$	367.9	0.1195	223.00	-10.868	193.9	0.1061	101.19	-9.222	112.7	0.0999	44.25	-8.452
Model C	$K = 1$	364.3	0.1189	220.41	-10.849	195.1	0.1065	101.91	-9.235	111.5	0.0997	43.43	-8.441
	$K = 2$	359.1	0.1181	216.88	-10.795	191.5	0.1058	99.42	-9.213	109.6	**0.0994**	42.14	-8.404
	$K = 5$	**352.7**	**0.1180**	**212.37**	**-10.732**	188.7	0.1056	97.58	-9.174	**108.3**	0.0995	**41.21**	**-8.316**
	$K = 8$	353.4	0.1183	212.83	-10.741	**186.3**	**0.1055**	**95.80**	**-9.156**	109.2	0.0995	41.85	-8.423

Best results are highlighted in bold fonts.

for bus travel time. Figure 4.13 demonstrates that Markov chains are close to their steady state distribution. Figure 4.6 shows the estimated mean vectors (standardization) for both classes/patterns. We can see that the mean vectors demonstrate significant differences in some link travel times (e.g., link #11, #12, #13, #14, #15) and many headways. To better find the difference of classes in terms of link travel time/headway, Figure 4.7 visualize the trajectory plots by using the sampling link travel times and headways. By comparing the estimated trajectories of class 1 and class 2, we can find that: class 1 has longer link/trip travel times than class 2; class 1 has shorter headways than class 2; class 1 has larger variances for link travel times. Overall, the mixture components implicitly capture the operational patterns (e.g., demand/frequency/traffic state) of the bus route: class 1 seems more like the operation in congested traffic and rush hours—higher frequency but slower speed, while class 2 may represents the operation in a free traffic state and off-peak hours.

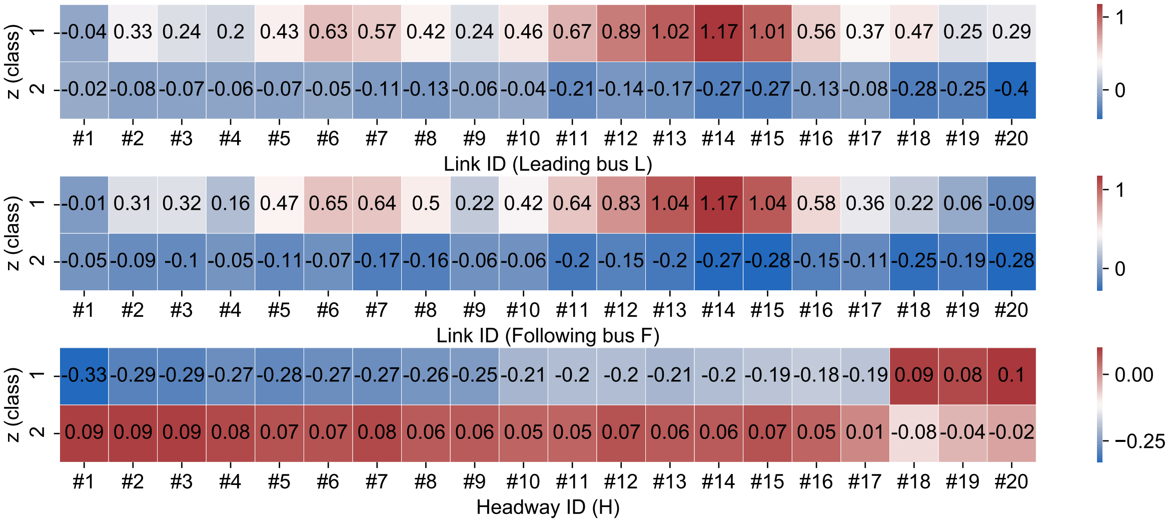
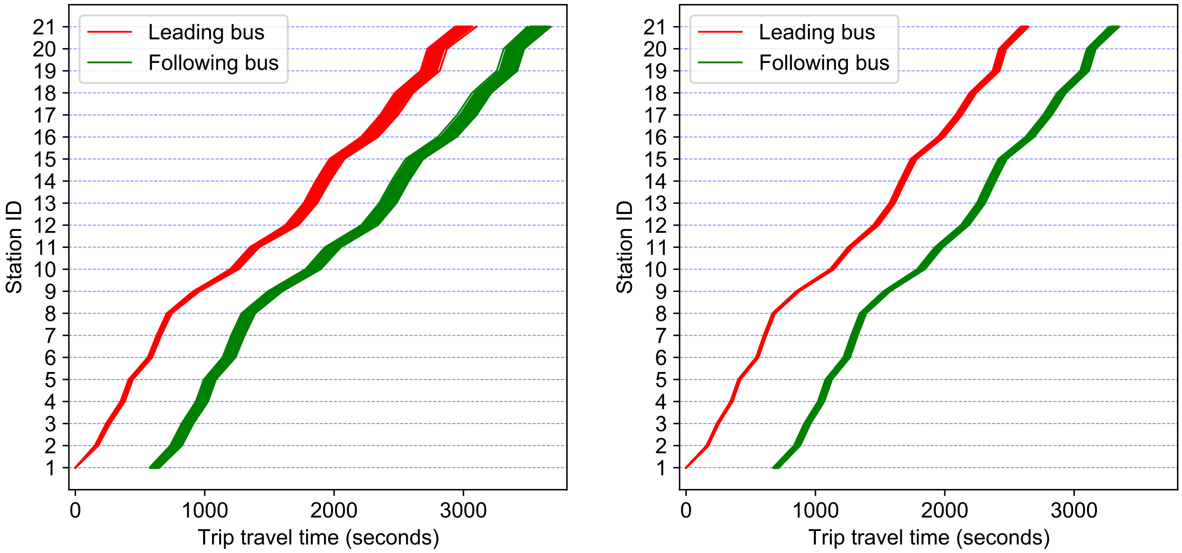


Figure 4.6: Mean vectors for different classes.

To interpret the classes, Figure 4.8 depicts the clear time-evolving patterns of each component. This further confirms our interpretation on classes 1 and 2. We can see that class 1 is dominant for afternoon peak hours while it is inferior to class 2 for off-peak hours and morning peak hours. Generally, morning peak and afternoon peak have similar traffic characteristics while in this case morning peak has similar characteristics to off-peak hours. This is because the studied directional bus route stretches from urban business districts to suburban areas. The better traffic condition and the small passenger flow exist in the morning because few people go to suburban areas on weekday mornings. On the contrary, traffic congestion and large passenger flow happen in the afternoon peak as more people



(a) Trajectory samples for class 1.

(b) Trajectory samples for class 2.

Figure 4.7: Distribution of the estimated trajectory for different classes.

go home from urban to suburban areas. Therefore, class 1 (afternoon peak) exhibits a longer travel time, shorter headway (higher frequency), and larger uncertainty than class 2 (morning peak, off-peak hours).

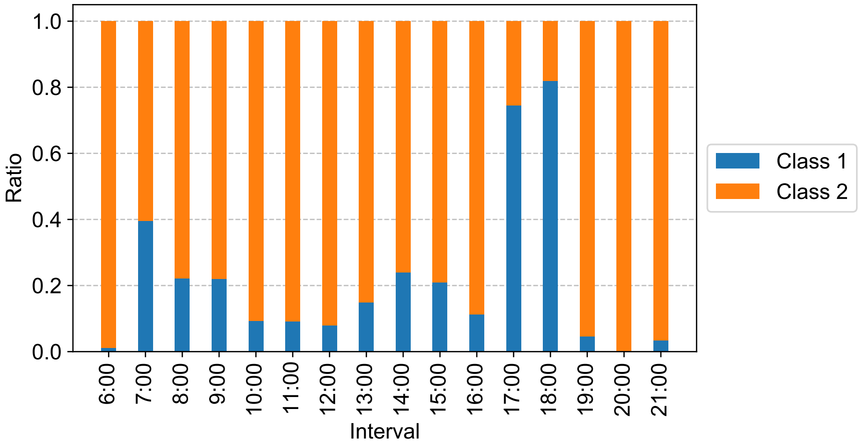


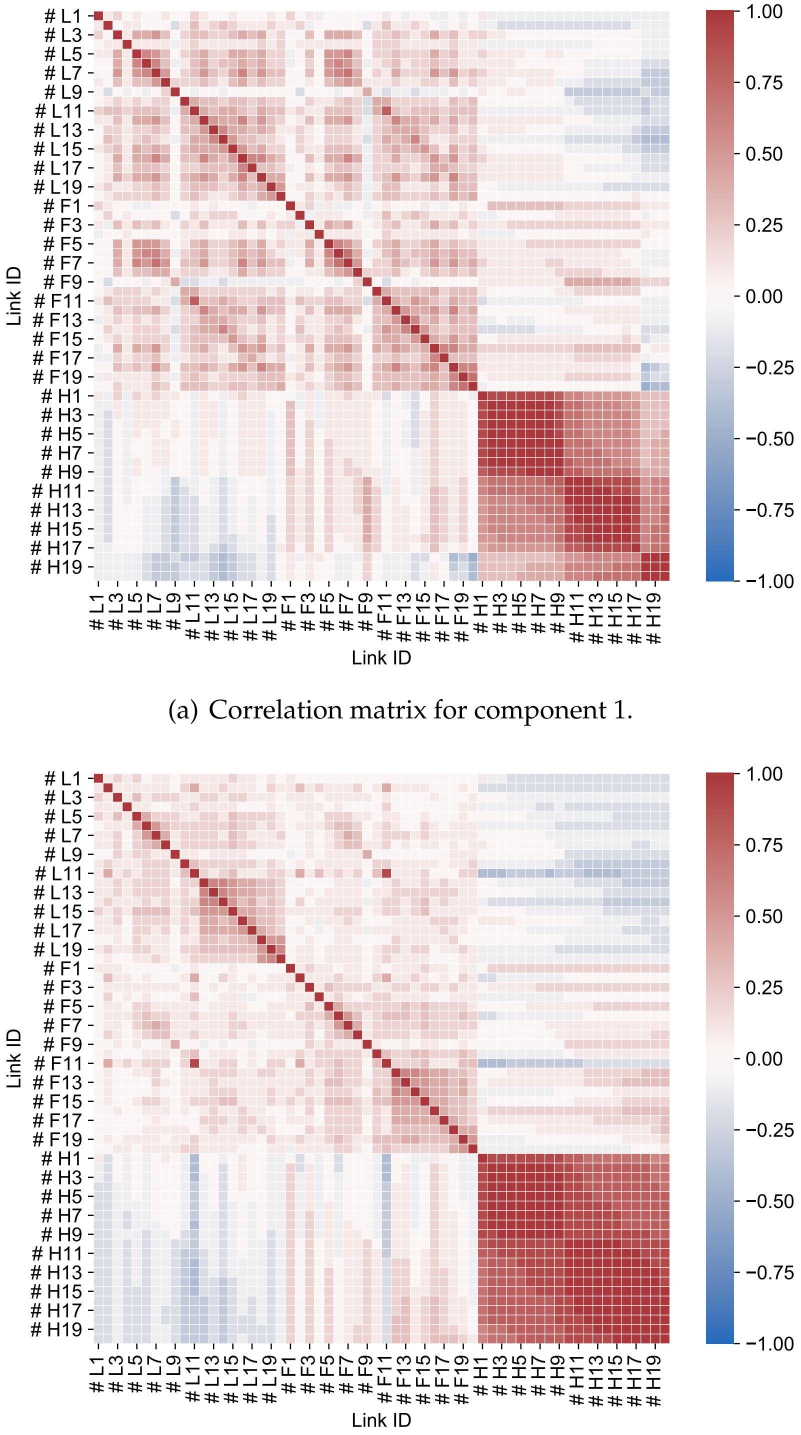
Figure 4.8: Component distribution for different intervals.

Figure 4.9 depicts the correlation matrices for different classes/patterns. Both of the correlation matrices show the complex characteristics of the correlations between link travel times/headways: 1) long-range correlations, 2) negative correlations, and 3) different patterns. On the other hand, the two components also show substantial differences. The most significant difference is that the leading bus and the following bus for class 1 could be more correlated than that for class 2, which is reasonable since class 1 represents the scenarios in which the services are slower but more frequent.

4.5.6 Distribution of the forecast Bus Travel Time

We show the application of the probabilistic forecasting model for bus route No. 60 of a selected day in Figure 4.10. The left panel intuitively illustrates the probabilistic forecasting for bus travel time at 4:50 p.m., while the right panel corresponds to the forecasting at 5:10 p.m. All actual bus operations are shown by green lines in the time-space diagram. Assume that past operation of buses (i.e., before 4:50 p.m. and before 5:10 p.m.) are observed and we use the observed information to forecast the future time-space position of buses. The variability of the probabilistic forecasting is shown with the 10th, 25th, 40th, 60th, 75th, and 90th percentile values. Note that determining the best percentile to reflect reliability for decision-making by operations and riders is not a simple task and requires further investigation, such as studying passengers' risk aversion towards travel time reliability. Our model can output any percentile according to practical needs, and we use the 10th to 90th percentiles to visually examine the prediction performance. We can see that the proposed model can make a good probabilistic forecasting for bus arrival time. By comparing the two figures, we can also find that for a bus trip, more observed links can help improve forecasting accuracy and reduce forecasting variability.

Finally, we use the two bus trips to show the forecasting distributions. Figure 4.11 shows the probabilistic forecasting results. In this experiment, we assume that we have observed the first eleven link travel times of the following bus and the corresponding link travel times/headways of the leading bus, then we intend to make probabilistic forecasting for the trip travel times of the last several links. In the left panel, the brown points are the trip travel times of the leading bus; the blue points are the true trip travel times of the following bus, and the green points are the predictive mean values. We can see that the predictive mean values fit the actual values quite well, demonstrating the proposed probabilistic model can achieve accurate forecasts. Moreover, as the number of links in a trip increases, we observe that the red bell curves become more spread out, indicating an increased variance of the trip travel time. The right panels show the mean corrected estimation, and the purple points (we refer to them as corrected mean values) are computed by posterior conditional mean values minus model mean values; the orange points are computed by the difference between true values and model mean values. We can find that the posterior conditional mean can make a more accurate forecasting than the

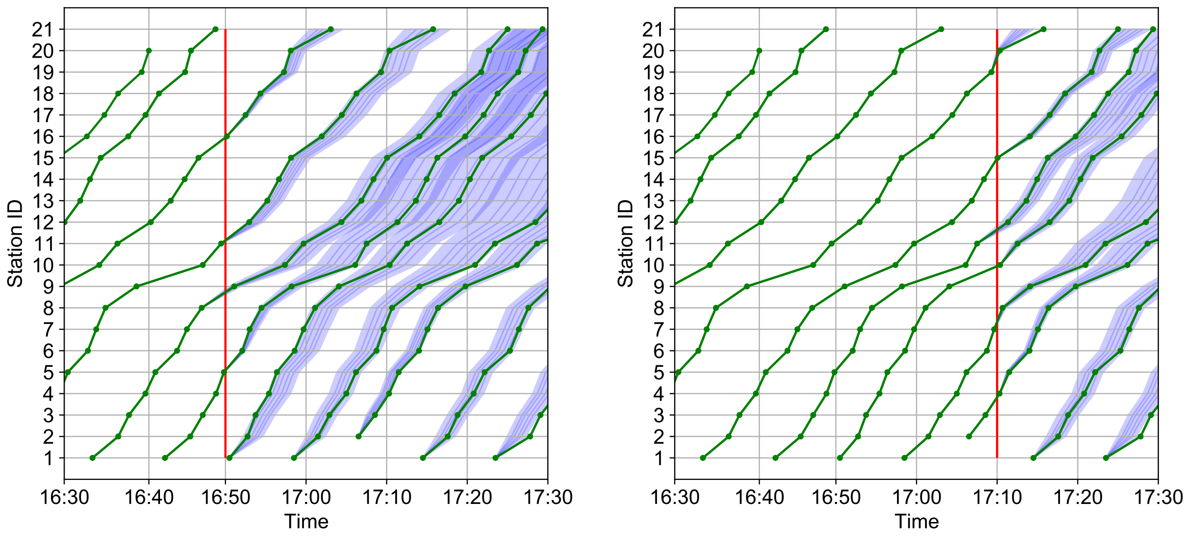


(a) Correlation matrix for component 1.

(b) Correlation matrix for component 2.

Figure 4.9: Correlation matrices for different components.

model mean. If we do not use the information of the observed link travel times/headways,



(a) Probabilistic forecasting for 16:50.

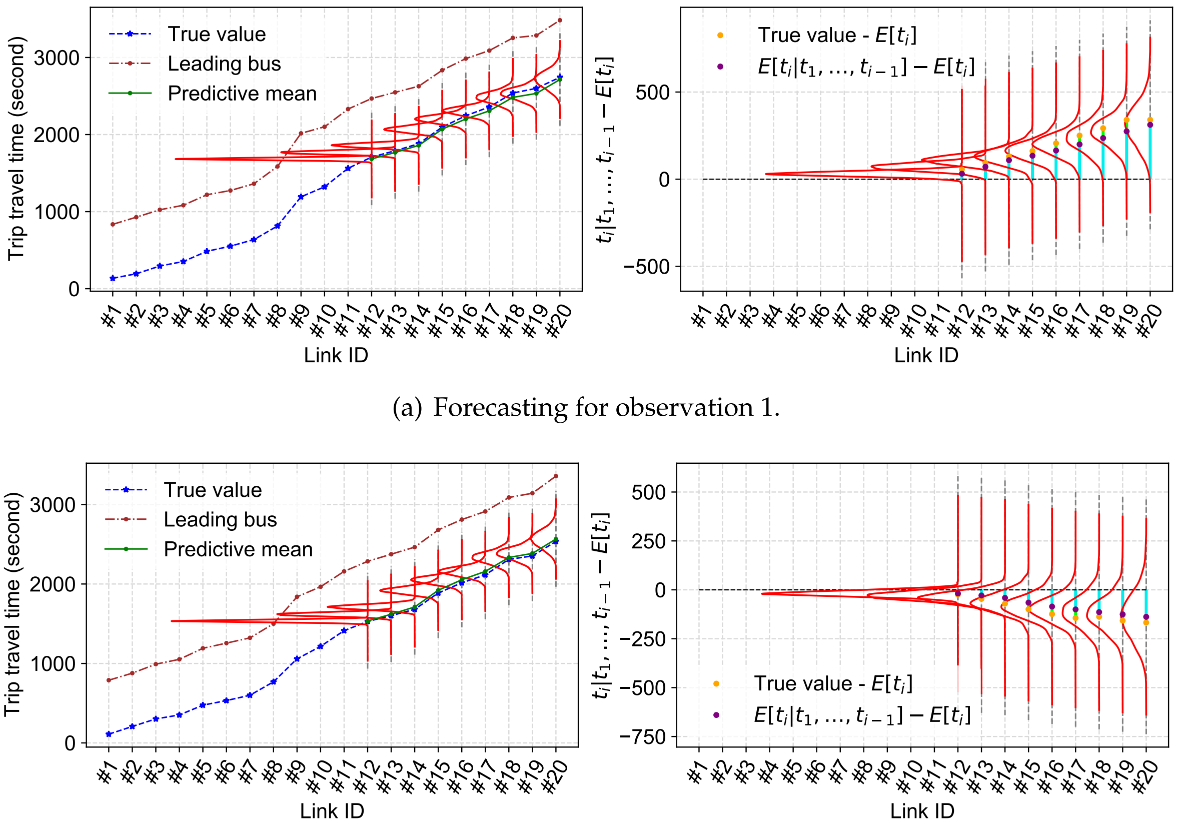
(b) Probabilistic forecasting for 17:10.

Figure 4.10: Samples of probabilistic forecasting of route No. 60.

the forecasting mean vectors should be the model mean vectors. As can be seen, the corrected mean values for observation 1 as shown in Figure 4.11 (a) are larger than zero, while the corrected mean values for observation 2 as shown in Figure 4.11 (b) are lower than zero, demonstrating the observed information indeed reinforces the forecasting for the upcoming links.

4.6 Conclusion

In this paper, we propose a new representation that combines bus link travel time and headway from a pair of adjacent buses, and model it with Multivariate Gaussian mixture distributions for probabilistic bus travel time forecasting. Our approach naturally captures/handles the link travel time correlations of a bus route, the interactions between adjacent buses, the multimodality of bus travel time distribution, and missing values in data. We also integrate the Gaussian mixture model with a Bayesian hierarchical framework to capture bus travel time patterns in different periods of a day. We conduct numerical experiments on a real-world dataset to evaluate the proposed model and our results confirms the superiority of the proposed model. It should be also noted that the training of this model is actually performed in an offline manner based on historical data, and conditional forecasting is very efficient based on stored Markov chains of model parameters. The model can be also retrained every a few days/weeks to accommodate



(a) Forecasting for observation 1.

(b) Forecasting for observation 2.

Figure 4.11: The probabilistic forecasting for trip travel time.

temporal variation in the system.

Besides making probabilistic forecasting, our approach has other potential practice and research implications. First, the parameters of the proposed model contain rich information for bus agencies to build better timetables and schedules and to evaluate the resilience and reliability of timetables/schedules. For example, analyzing link travel time correlations and delay propagation using correlation matrices, and understanding temporal patterns of the bus route from mixing coefficients. Second, with sufficient training data, we can also use the proposed model to make probabilistic simulations about train/bus operation and delay propagation. Third, our proposed framework can incorporate further information, such as passenger demand, onboard passengers/occupancy, to reinforce the probabilistic forecasting for bus travel time. This will allow us to characterize how travel time, headway/delay, boarding demand and onboard occupancy interact with each other in a data-driven way and help operators achieve additional real-time occupancy/load forecasting along the bus route, which is important to travelers under the pandemic (see e.g., Pasini et al. (2019); Jenelius (2019, 2020)).

Our proposed probabilistic forecasting model for bus travel time only considers using one leading bus. In practice, a subject bus of interest could be correlated with more than one leading buses; therefore, we can incorporate more leading buses to reinforce the probabilistic travel time forecasting for the following bus. In the extreme case, we can take into account all the buses in operation along the route (e.g., 6 buses at 5:10 pm in Figure 4.10(b)) and make conditional forecasting for them simultaneously without using the autoregressive approach. However, the increasing number of leading buses will result in a significant increase in dimensionality (e.g., vector size > 100). This will not only lead to increased computational cost as the covariance matrix will be large but also require a much larger set of training data. To address this issue, we will consider using a mixture of probabilistic principal component analysis (PCA) to reduce dimensionality. Our further research will utilize the mixture of probabilistic PCA to model the following bus with more leading buses. With this idea, it is also possible to build a much larger model than can generate probabilistic forecasting for all buses (i.e., the whole fleet) on the road when we have sufficient amount of training data. Furthermore, investigating the links between bus travel time and physical operation conditions would be a compelling future direction. For instance, from Figure 4.12, we observe that link #8, which has two intersections, has a larger travel time with greater variance than link #6, which has no intersections. Although our current focus is on developing a data-driven model for bus travel time forecasting, we are keen on exploring the relationship between bus travel time and various physical operation conditions in future studies. Another potential direction is to employ probabilistic forecasting for bus travel time across the bus network. In this study, we assume that all bus lines are independent and only examine interactions among buses on the same bus line. However, buses from different bus lines may have correlations or interactions, especially when they share many common road segments. Hence, future research could explore the correlations among buses on the bus network and leverage them to improve travel time forecasting.

4.7 Appendix

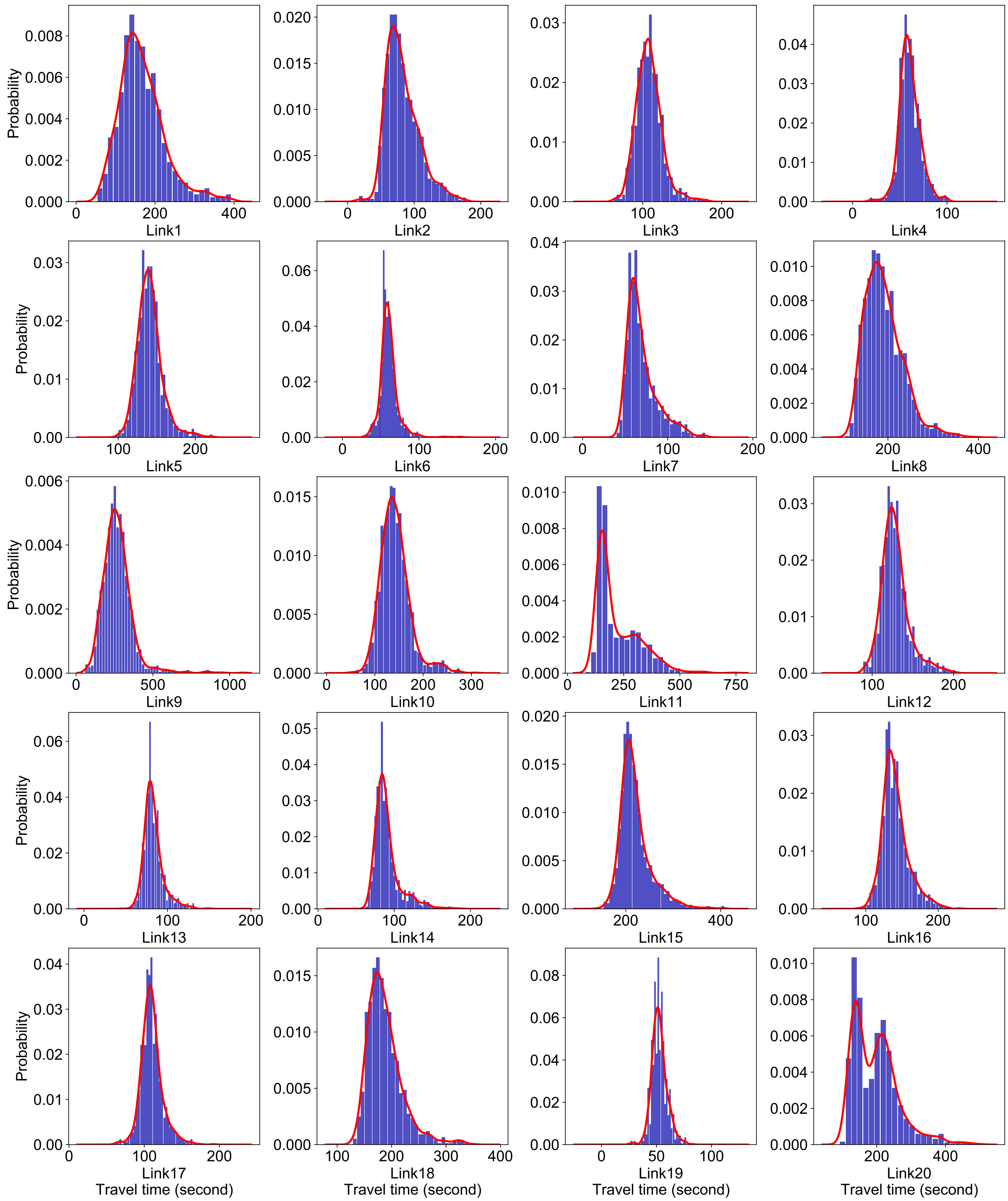
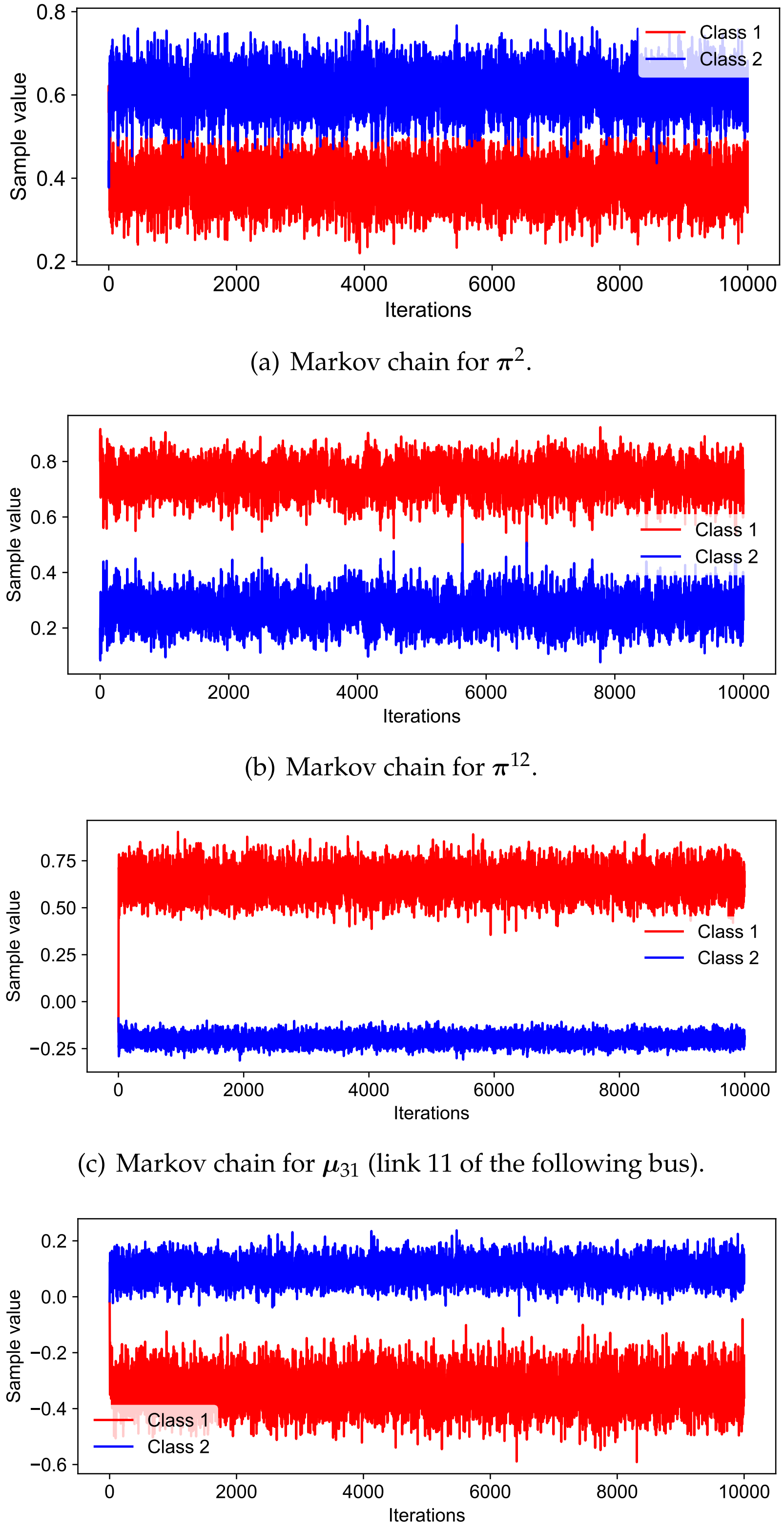


Figure 4.12: Empirical distribution of link travel time of bus route No. 60.



(a) Markov chain for π^2.

(b) Markov chain for π^{12}.

(c) Markov chain for μ_{31} (link 11 of the following bus).

(d) Markov chain for μ_{41} (headway at the first stop).

Figure 4.13: Markov chains for some parameters.

Chapter 5

Bayesian Forecasting of Bus Passenger Occupancy

This chapter is a research article submitted to *Transportation Research Part B: Methodological*:

- **Chen, X.**, Cheng, Z., Schmidt, A. M., Sun, L. 2024. Conditional forecasting of bus travel time and passenger occupancy with Bayesian Markov regime-switching vector autoregression. arXiv preprint arXiv:2401.17387.

This chapter corresponds to the probabilistic forecasting of bus travel time and passenger occupancy with a Bayesian Markov regime-switching vector autoregression model. This chapter focuses on the probabilistic forecasting of passenger occupancy while it is the extension of the previous Chapter 4.

5.1 Abstract

Accurately forecasting bus travel time and passenger occupancy with uncertainty is essential for both travelers and transit agencies/operators. However, existing approaches to forecasting bus travel time and passenger occupancy mainly rely on deterministic models, providing only point estimates. In this paper, we develop a Bayesian Markov regime-switching vector autoregressive model to jointly forecast both bus travel time and passenger occupancy with uncertainty. The proposed approach naturally captures the intricate interactions among adjacent buses and adapts to the multimodality and skewness of real-world bus travel time and passenger occupancy observations. We develop an efficient Markov chain Monte Carlo (MCMC) sampling algorithm to approximate the resultant joint posterior distribution of the parameter vector. With this framework, the estimation of downstream bus travel time and passenger occupancy is transformed into a multivariate time series forecasting problem conditional on partially observed outcomes. Experimental validation using real-world data demonstrates the superiority of our proposed model in terms of both predictive means and uncertainty quantification compared to the Bayesian Gaussian mixture model.

5.2 Introduction

The rapid progress of urbanization has brought increasing population and economic agglomeration in large cities. Urban transportation problems such as increased traffic congestion and pollution, high energy consumption and greenhouse gas emissions, and growing safety and accessibility concerns, have been persistently challenging the development of sustainable cities and communities. In the *2030 Agenda for Sustainable Development*, the United Nations has emphasized the critical role of public transportation in shaping a sustainable society (United Nations, 2015). However, despite the growing investment in infrastructure, North American cities have not seen rapid growth and even observed a decline in ridership in recent years, even before the COVID-19 pandemic (Erhardt et al., 2022). One of the key reasons is that the operation of transit services suffers from reliability issues. Bus operation is a highly challenging problem due to the inherent instability of the system—a slightly delayed bus will be further delayed as it will encounter more waiting passengers and experience longer dwell time. Unreliability in travel time and overcrowdedness resulting from unstable operations (e.g., “bus bunching”, see Daganzo, 2009; Bartholdi III and Eisenstein, 2012) have been the main factors preventing travelers from using public transportation (Carrel et al., 2013). Having access to accurate travel time and occupancy forecasting along with uncertainty becomes important to travelers to make informed travel planning in terms of mode choice, route choice, and even vehicle choice (e.g., waiting for a less crowded bus or boarding a full vehicle) (Yu et al., 2017). For transit agencies/operators, probabilistic forecasting could benefit the design of robust bus management strategies, such as bus route design (e.g., Zheng et al., 2016), bus crowding control (e.g., Wang et al., 2021b), timetable design (e.g., Jiang et al., 2021), and bus bunching control (e.g., Xuan et al., 2011; Bartholdi III and Eisenstein, 2012).

In general, a bus link refers to the one-way segment that connects two adjacent bus stops along a bus route. Link travel time of a bus is defined as the time difference between the arrivals at two adjacent bus stops associated with the bus link. Therefore, the trip travel time of the bus from one stop to another can be calculated by summing up all link travel times between these two stops. The passenger occupancy of a bus on the link is defined as the total number of passengers onboard while the bus is traversing that particular link. The main goal of this paper is to provide real-time predictions of downstream travel time and passenger occupancy of a bus along a given route.

Previous studies on forecasting bus travel time have predominantly employed deterministic approaches, based on techniques such as Artificial Neural Network (Gurmu and Fan, 2014), Support Vector Machine (Yu et al., 2011; Kumar et al., 2013), K-nearest neighbors model (Kumar et al., 2019), Long-Short-Term Memory neural network (He et al., 2018), and various hybrid models (Yu et al., 2018). In addition, deterministic passenger occupancy forecasting models have also been developed, including Lasso regularized linear regression model (Jenelius, 2019), partial least squares regression (Jenelius, 2020), Random Forest (Wood et al., 2023), and deep learning model (Bapaume et al., 2023). Despite the widespread use and simplicity of these deterministic models, a significant drawback is that they only provide point estimates, overlooking the randomness and uncertainty associated with the prediction.

There are only a few studies on probabilistic forecasting for bus travel time and passenger occupancy (e.g., Ma et al., 2017; Dai et al., 2019; Büchel and Corman, 2022a,b; Chen et al., 2023). However, these studies often adopt oversimplified assumptions and ignore important operational characteristics of bus systems, including:

- Strong interactions/correlations between travel time and passenger occupancy. Previous studies on probabilistic prediction for bus travel time (Ma et al., 2017; Dai et al., 2019; Büchel and Corman, 2022a,b) and passenger occupancy (Wang et al., 2021a) have typically treated these variables independently. However, the intricate interactions and correlations between bus travel time and passenger occupancy are not adequately considered. For example, the boarding and alighting processes take longer for a crowded bus compared to an empty one (Sun et al., 2014b). Incorporating these interactions through a joint modeling approach could significantly enhance the accuracy of probabilistic forecasting.

- Complex link correlations of both travel time and passenger occupancy. Chen et al. (2022) demonstrated that link travel times on a bus route exhibit complex local and long-range correlations. In Section 5.3, we also find that link passenger occupancy on a bus route exhibits local and long-range correlations. However, existing studies often focus only on local correlations while neglecting long-range correlations in probabilistic forecasting for both bus travel time (Ma et al., 2017; Büchel and Corman, 2022a,b) and passenger occupancy (Wang et al., 2021a).
- Interactions/correlations between adjacent buses along a bus route. Adjacent buses often have strong interactions that lead to system instability and bus bunching (Daganzo, 2009; Bartholdi III and Eisenstein, 2012); for instance, due to the increase in headway (i.e., duration between two arrivals), a delayed bus could see more passengers waiting at the bus stop and get further delayed. To our knowledge, only Büchel and Corman (2022b) and Chen et al. (2023) considered the interactions/correlations between adjacent buses along a bus route in their proposed models for bus travel time forecasting, while most studies (Ma et al., 2017; Büchel and Corman, 2022a) focused on modeling the correlations between adjacent links along a bus route.

The recent contribution by Chen et al. (2023) develops a time-dependent Gaussian mixture model, which treats the concatenation of link travel time data from two consecutive buses as a random variable. The estimated model can then be used to perform conditional forecasting for the travel times of all downstream links. The study demonstrates that incorporating information from neighboring buses significantly improves forecasting accuracy. However, a notable limitation of this model is that two consecutive observations are assumed to be conditionally independent given their latent states, thus neglecting the temporal/dynamic relationships among multiple buses. This paper addresses the aforementioned challenges by developing a *joint* Bayesian model for bus travel time and passenger occupancy, building upon the foundation laid by Chen et al. (2023). To model the correlations between travel time and passenger occupancy, we construct a variable that combines the link travel time vector, the passenger occupancy vector, and the departure headway. More importantly, we employ a Bayesian Markov regime-switching vector autoregressive model to characterize the dynamic relationship among multiple buses. This new approach effectively captures essential interactions between adjacent buses, along with the multimodality and skewness of bus travel time and passenger occupancy distributions. Furthermore, it adeptly models intricate state transitions, particularly crucial when forecasting bus travel time and passenger occupancy with limited observations for the following bus. For model estimation, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm to draw samples from the resulting posterior distribution of the model parameters. As we follow the Bayesian paradigm to estimate the parameters of the model, predictions are obtained by approximating the posterior predictive distribution. We fit the proposed model to the smart card data of one bus route in an anonymous city. The experimental results confirm that the proposed Markov regime-switching vector autoregressive model outperforms existing methods in terms of both point estimates and uncertainty quantification. This holistic approach contributes to a more robust and nuanced understanding of bus travel time and passenger occupancy dynamics, offering improved forecasting capabilities in real-world scenarios.

The remainder of this paper is organized as follows. In Section 5.3, we perform an empirical analysis of bus travel time and passenger occupancy using real-world data. In Section 5.4, we present the proposed Bayesian Markov regime-switching vector autoregressive model. The forecasting process and the probabilistic forecasting model are described in Section 5.5. We then showcase the capabilities of our proposed model by analyzing real-world data in Section 5.6. Finally, we conclude the study and summarize our key findings in Section 5.7.

5.3 Data Description and Analysis

Smart card data, a prevalent data source in public transit studies, is collected from electronic fare payment systems—transactions are created when passengers use smart cards (e.g., contactless cards or mobile payment apps) to pay for their bus trips (Pelletier et al., 2011). These data sets capture information about boarding and alighting, including time, location, fare paid, and card ID. In this paper, we use the smart card data for one bus route (32 stops and 31 links) in an anonymous city to prepare the bus travel time and passenger occupancy data.

Figure 5.1 shows bus trajectories with passenger occupancy in one day. We can see that some adjacent buses have strong interactions like bus bunching, especially during morning peak hours. Bus bunching typically arises due to various factors, including traffic

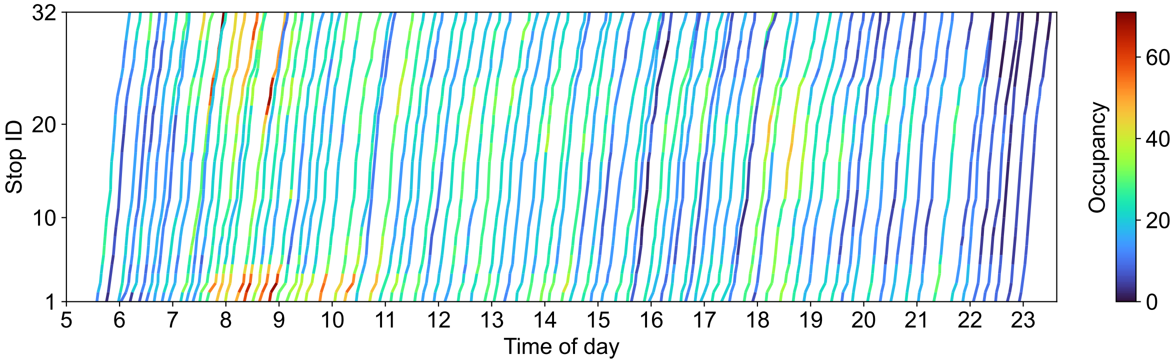


Figure 5.1: Bus time-space trajectories with passenger occupancy. The horizontal axis represents the "Time of day" (from 5:00 to 24:00) and the vertical axis represents "Stop ID" (from the starting stop to the final stop, 32 stops in total). Each curve contains 31 segments, and it depicts the trajectory of a bus traveling from the departure stop to the final stop. The color of a segment shows the passenger occupancy of the bus.

congestion, unpredictable passenger boarding and alighting times, variations in travel speeds, and delays caused by external factors such as road conditions, traffic signals, and accidents. When a bus falls behind schedule, it tends to experience increased passenger occupancy at subsequent stops and becomes further delayed. As a result, bus travel time and passenger occupancy have strong correlations, and adjacent buses on the same route often demonstrate complex interactions.

Figure 5.2 shows some example empirical distributions of travel time, passenger occupancy, and headway on three links. We observe clear characteristics such as positive skewness, heavy tails, and multimodality. Therefore, it is essential to develop a model that can effectively characterize such complex distributions to ensure the quality of probabilistic forecasting.

We empirically compute the correlation and cross-correlation matrices among the associate variables, which are shown in Figure 5.3. The first panel presents the link travel time correlation matrix, which reveals that link travel times have both local and long-range correlations. It is worth noting that the label h in this matrix refers to the headway at the departure/originating stop, which enables us to capture the relationship between the link travel times and departure headway. Therefore, we could use observed link travel times and the headway to forecast downstream link travel times. The second panel presents the passenger occupancy correlation matrix, which indicates that passenger occupancy also has strong local and long-range correlations. Similarly, passenger occupancy is also correlated with departure headway. In the third panel, we observe the cross-correlation matrix between link travel time and passenger occupancy, and we can

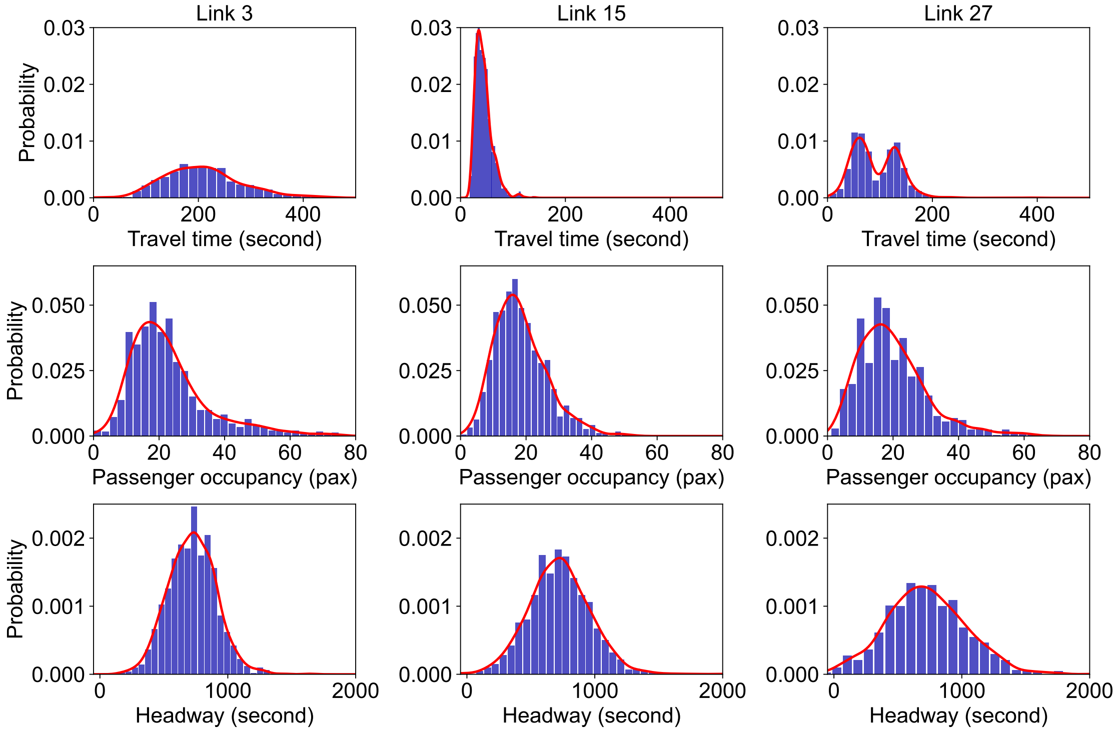


Figure 5.2: Examples of empirical distributions of link travel time (first row), passenger occupancy (second row), and headway (third row). The red lines represent kernel density estimates of the empirical distributions across the links.

see that they have strong correlations. Consequently, jointly modeling link travel time and passenger occupancy could make more accurate forecasting. The last two panels show the cross-correlation matrix between the leading bus and the following bus in terms of link travel time and passenger occupancy, respectively. We can see that two adjacent buses are strongly correlated in link travel time and passenger occupancy. Therefore, it is also crucial to take into account the interaction between two adjacent buses to improve prediction performance. In summary, our empirical findings demonstrate three key observations: i) strong interactions/correlations between travel time and passenger occupancy, ii) complex link correlations in travel time and passenger occupancy, and iii) strong interactions/correlations between adjacent buses along a bus route. Thus, we contend that, in constructing a forecasting model, bus travel time and occupancy should be jointly modeled, with explicit consideration given to the interactions between two adjacent buses. By doing so, the model will provide a more accurate representation of the dynamic and interconnected nature of bus travel time and passenger occupancy, thus improving the predictive performance of the model.

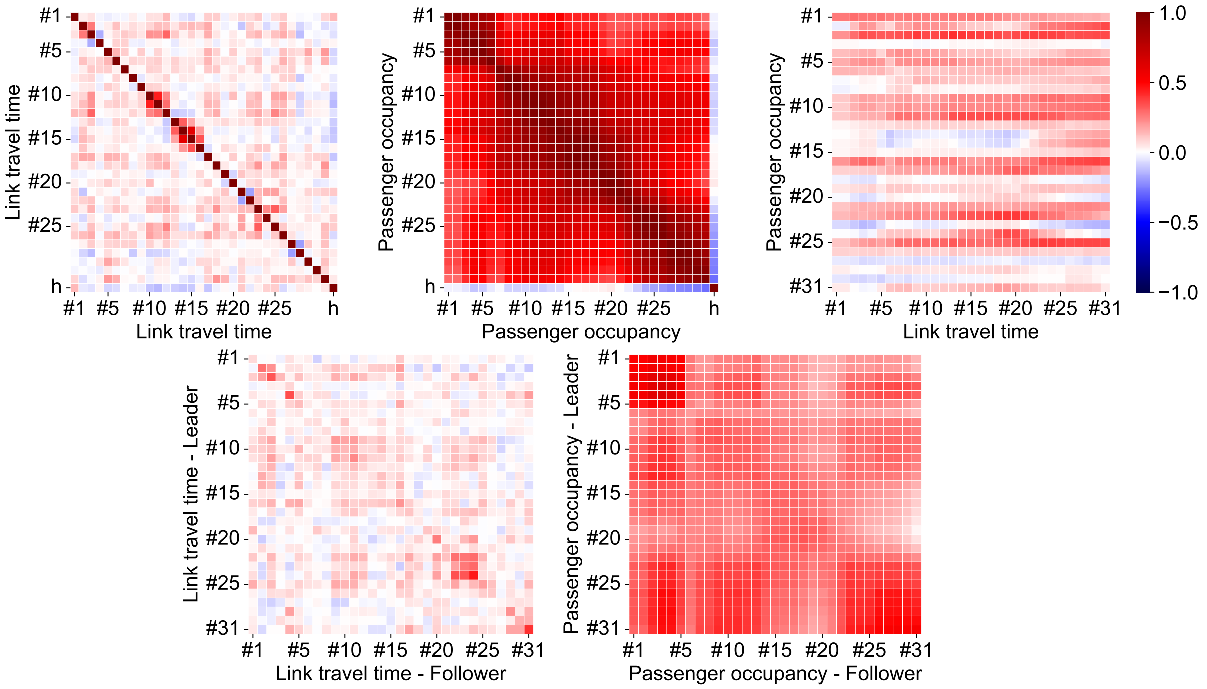


Figure 5.3: Correlation and cross-correlation matrices of associated variables.

5.4 Proposed Model and Bayesian Inference

5.4.1 Notations of Variables

Let $\ell_{i,m}^{(d)}$ represent the travel time of the i-th bus on the m-th link on the d-th day. The trip travel time of the i-th bus from stop m_1 to stop m_2 is given by $\sum_{m=m_1}^{m_2-1} \ell_{i,m}^{(d)}$. We denote by $f_{i,m}^{(d)}$ the passenger occupancy of the i-th bus on the m-th link on the d-th day. The link travel time and occupancy of bus i on a bus route with n links (i.e., $n+1$ bus stops) can be stacked into $\boldsymbol{\ell}_i^{(d)} = \left[\ell_{i,1}^{(d)}, \ell_{i,2}^{(d)}, \cdots, \ell_{i,n}^{(d)}\right]^\top \in \mathbb{R}^n$ and $\boldsymbol{f}_i^{(d)} = \left[f_{i,1}^{(d)}, f_{i,2}^{(d)}, \cdots, f_{i,n}^{(d)}\right]^\top \in \mathbb{R}^n$, respectively. Note that here we define both travel time and occupancy to be real numbers for simplicity, whereas real-world travel time data should be strictly positive, and occupancy data should be in the form of counts. We define the departure headway $h_i^{(d)}$ as the time interval between the arrival of the $(i-1)$-th bus and the i-th bus at the originating bus stop on the d-th day. Figure 5.4 shows the representation of the variables in bus trajectories. We introduce a multivariate random variable, $\boldsymbol{y}_i^{(d)}$, as the concatenation of link travel time, the occupancy of the passengers and the departure progress for the bus i on the d-th day:

$$\boldsymbol{y}_i^{(d)} = \left[\boldsymbol{\ell}_i^{(d)\top}, \boldsymbol{f}_i^{(d)\top}, h_i^{(d)}\right]^\top \in \mathbb{R}^{2n+1}. \tag{5.1}$$

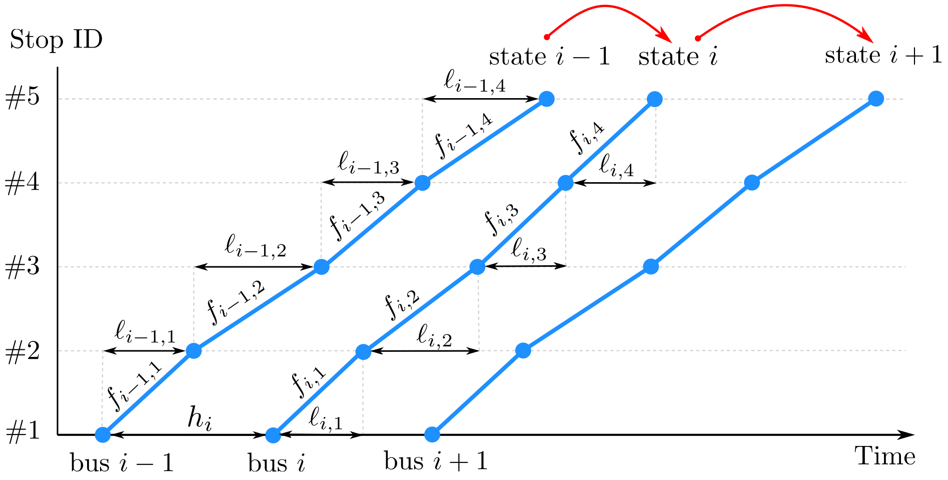


Figure 5.4: Representation of bus trajectories with passenger occupancy.

In doing so, we create a concise representation that allows us to analyze and model the relationship among these variables more effectively. Therefore, from historical smart card data, we could collect the observations $\left\{\boldsymbol{y}_i^{(d)}\right\}_{i=1,d=1}^{I_d,D}$ denoted by $\mathcal{Y}$, where I_d is the total number of bus runs we have in the training/historical data on the d-th day and D represents the number of days in the data set.

5.4.2 Bayesian Markov Regime-Switching Vector Autoregressive Model

To model the random variable, Chen et al. (2023) developed a Bayesian time-dependent Gaussian mixture model for probabilistic forecasting. This model assumes that each pair of buses has a latent class (i.e., hidden state), and the core idea is to use the observed information to infer the hidden state. In Figure 5.4, we can observe two successive bus pairs: the first pair i (with bus $i-1$ and bus i) and the second pair $i+1$ (with bus i and $i+1$). In the Bayesian Gaussian mixture model, the relationship between hidden states of adjacent bus pairs (e.g., pair i and pair $i+1$) is not modeled directly. In other words, the state of bus pair $i+1$ has no direct relationship with the state of bus pair i except that they likely come from the same time window. Consequently, when estimating the hidden state of a bus, the accuracy of the estimation heavily relies on the amount of observed information available for that bus. However, as mentioned in Section 5.3, the interactions between adjacent buses, particularly in scenarios where bus bunching occurs, reveal a clear interdependence between adjacent states. By accurately modeling the relationships between the states of adjacent buses, we can leverage more information to infer the state of the current bus and improve the estimation. For instance, in Figure 5.4, when we want to estimate the hidden state of bus $i+1$, even if we have limited observed links for the current bus $i+1$, we can accurately estimate the state of bus i because this bus could have more observed information. Using the modeled relationship between adjacent states and the observed links of bus $i+1$, we can enhance the estimation of the state of bus $i+1$. Therefore, considering the modeling of relationships between adjacent hidden states becomes crucial for accurate probabilistic forecasting. To address this issue, we propose employing a Bayesian regime-switching Markov model to capture the dependency between adjacent hidden states and to capture the relationship between two adjacent buses.

Markov regime-switching models are designed to analyze and forecast time series data that may shift between different states or regimes over time. These models have gained significant popularity in econometrics and time series analysis since the work of Hamilton (1989). Regime-switching models have been applied to many tasks such as speech recognition (Kim and Nelson, 2017) and motion recognition (Fox et al., 2014).

In our Bayesian Markov regime-switching vector autoregressive model, we assume that the conditional distribution of the vector $\boldsymbol{y}_i^{(d)}$, given the preceding observed vector $\boldsymbol{y}_{i-1}^{(d)}$ and the current hidden state $z_i^{(d)}$, is described by

$$\boldsymbol{y}_i^{(d)} \mid \boldsymbol{y}_{i-1}^{(d)}, z_i^{(d)} = k \sim \mathcal{N}\left(\boldsymbol{A}_k \boldsymbol{y}_{i-1}^{(d)} + \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{5.2}$$

where $\boldsymbol{y}_i^{(d)} \mid \boldsymbol{y}_{i-1}^{(d)}, z_i^{(d)} = k$ follows a multivariate Gaussian distribution. This model is commonly used to capture complex and heterogeneous relationships among variables in multivariate time series analysis (Krolzig, 2013). In this equation, $\boldsymbol{A}_k$ is a coefficient matrix that characterizes the influence of the preceding bus on the current bus under the regime of the hidden state k. The mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$ are parameters specific to the hidden state k, allowing the model to capture the varying dynamics of the bus system under different operational conditions. The hidden state $z_i^{(d)}$ represents the latent regime or condition of the i-th bus on the d-th day. These hidden states encapsulate unobserved factors that influence the bus's operational characteristics, such as traffic conditions, weather, or other variables. By modeling $\boldsymbol{y}_i^{(d)}$ as a function of both its preceding observation $\boldsymbol{y}_{i-1}^{(d)}$ and the hidden state $z_i^{(d)}$, the model can capture the complex and dynamic interactions within the bus route, thus improving the predictive accuracy of bus travel time and passenger occupancy over nonstationary operational conditions.

Another critical aspect of our proposed model is the transition of hidden states, represented by probability $p\left(z_i^{(d)} \mid z_{i-1}^{(d)}\right)$. This probability is modeled as a categorical distribution dependent on the state transition matrix $\boldsymbol{\pi}$. We use $\boldsymbol{\pi}_k = (\pi_1, \ldots, \pi_K)^\top$ to denote a vector representing the transition probabilities from state k to other states, i.e., $0 \leqslant \pi_k \leqslant 1$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\pi} = \left[\boldsymbol{\pi}_1^\top, \ldots, \boldsymbol{\pi}_K^\top\right]^\top$ represents the state transition matrix. Specifically, the probability of transition from state $z_{i-1}^{(d)}$ to state $z_i^{(d)}$ is given by $\pi_{z_{i-1}^{(d)}, z_i^{(d)}}$, which follows a categorical distribution parameterized by $\boldsymbol{\pi}_{z_{i-1}^{(d)}}$

$$z_i^{(d)} \mid z_{i-1}^{(d)} \sim \text{Categorical}\left(\boldsymbol{\pi}_{z_{i-1}^{(d)}}\right). \tag{5.3}$$

This formulation captures the Markov property of the model, where the probability of bus i being in a particular state is dependent solely on the state of its preceding bus $i-1$. This structure is critical in modeling how the state of each bus is influenced by its immediate predecessor. For the initial state, we assume

$$z_1^{(d)} \sim \text{Categorical}\left(\boldsymbol{\pi}^*\right), \tag{5.4}$$

where $\boldsymbol{\pi}^*$ is the probability distribution of the initial state, which is the marginal distribution calculated from the state transition matrix $\boldsymbol{\pi}$.

5.4.3 Prior Specification

In general, we assume the following conjugate prior distributions for $\boldsymbol{\Sigma}_k$, $\boldsymbol{\mu}_k$, and $\boldsymbol{A}_k$,

$$\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi}_0, \nu_0\right), \tag{5.5}$$

$$\boldsymbol{\mu}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda_0}\boldsymbol{\Sigma}_k\right), \tag{5.6}$$

$$\boldsymbol{A}_k \sim \mathcal{MN}\left(\boldsymbol{M}_0, \boldsymbol{\Sigma}_k, \boldsymbol{V}_0\right), \tag{5.7}$$

where $\mathcal{W}^{-1}\left(\boldsymbol{\Psi}_0, \nu_0\right)$ is the inverse-Wishart distribution with a scale matrix $\boldsymbol{\Psi}_0$ and ν_0 degrees of freedom; $\boldsymbol{\mu}_0$ and λ_0 are parameters for the Gaussian prior; $\mathcal{MN}\left(\boldsymbol{M}_0, \boldsymbol{\Sigma}_k, \boldsymbol{V}_0\right)$ is the matrix Gaussian distribution with parameters $\boldsymbol{M}_0$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{V}_0$. These prior distributions are fundamental to our Bayesian framework, allowing the model to incorporate prior knowledge and uncertainty effectively. In addition, we use a Dirichlet prior distribution for each transition probability:

$$\boldsymbol{\pi}_k \mid \boldsymbol{\alpha} \sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right), \tag{5.8}$$

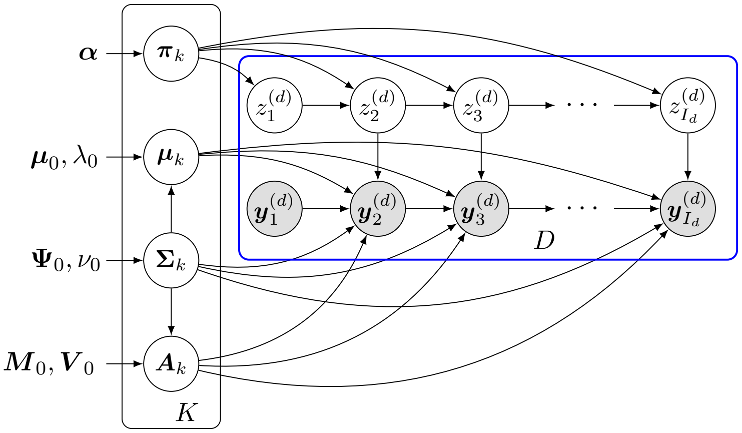


Figure 5.5: Graphical representation of a Bayesian Markov regime-switching vector autoregressive model.

where $\boldsymbol{\alpha}$ is the concentration parameter of the Dirichlet distribution. The Dirichlet distribution is a natural choice for modeling probability vectors $\boldsymbol{\pi}_k$ because it ensures that the probabilities are non-negative and sum up to one, which are essential properties for any set of transition probabilities.

5.4.4 Model Overview

Figure 5.5 provides the overall graphical representation of the proposed model. Considering the vector autoregressive model and the Markov regime-switching model as separate dimensions, we can think of the proposed combination as a spatial-temporal model. The overall model specification is summarized as follows:

(1) Draw model parameters of state k from 1 to K:

(a) Draw the state transition probability $\boldsymbol{\pi}_k \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

(b) Draw the covariance matrix $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_0, \nu_0)$.

(c) Draw the mean vector $\boldsymbol{\mu}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda_0}\boldsymbol{\Sigma}_k\right)$.

(d) Draw the coefficient matrix $\boldsymbol{A}_k \sim \mathcal{MN}(\boldsymbol{M}_0, \boldsymbol{\Sigma}_k, \boldsymbol{V}_0)$.

(2) For each sequence of d-th day from 1 to D:

(a) Draw the initial state $z_1^{(d)} \sim \text{Categorical}(\boldsymbol{\pi}^*)$.

(b) For each bus i from 2 to I_d:

(i) Draw state sequence $z_i^{(d)} \mid z_{i-1}^{(d)} \sim \text{Categorical}\left(\boldsymbol{\pi}_{z_{i-1}^{(d)}}\right)$.

(ii) Draw observations $\boldsymbol{y}_i^{(d)} \mid \boldsymbol{y}_{i-1}^{(d)}, z_i^{(d)} = k \sim \mathcal{N}\left(\boldsymbol{A}_k\boldsymbol{y}_{i-1}^{(d)} + \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$.

5.4.5 MCMC Algorithm to Approximate Posterior Distribution

For any $k = 1, \ldots, K$, we define $M_k = \left\{(i,d) \mid z_i^{(d)} = k, d = 1, \ldots, D, i = 1, \ldots, I_d\right\}$ as the set of bus indices with state $z_i^{(d)} = k$; we use $|M_k|$ to denote the number of elements in M_k. Similarly, for any $k = 1, \ldots, K$ and $k' = 1, \ldots, K$, we define $M_{k,k'} = \left\{(i,d) \mid z_i^{(d)} = k, z_{i+1}^{(d)} = k', d = 1, \ldots, D, i = 1, \ldots, I_d - 1\right\}$ as the set of indices of buses that have $z_i^{(d)} = k$ with its follower having $z_{i+1}^{(d)} = k'$. We define $\boldsymbol{A} = \{\boldsymbol{A}_k \mid k = 1, \ldots, K\}$ as the set of coefficient matrices. Define $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k \mid k = 1, \ldots, K\}$ as the set of mean vectors and $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k \mid k = 1, \ldots, K\}$ as the set of covariance matrices. For simplicity, we let $\Theta = \{\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0\}$ denote the set of hyperparameters for the Gaussian-inverse-Wishart prior distribution in Eq. (5.5) and (5.6). In addition, we let $\boldsymbol{y}_{1:I_d}^{(d)}$ and $\boldsymbol{z}_{1:I_d}^{(d)}$ denote the observations and states of the bus sequence on the d-th day; we denote by $\mathcal{Y}_k = \left\{\boldsymbol{y}_i^{(d)} \mid z_i^{(d)} = k, d = 1, \ldots, D, i = 1, \ldots, I_d\right\}$ the set of data vectors belonging to state k. Let $\boldsymbol{Z} = \left\{z_i^{(d)} \mid d = 1, \ldots, D, i = 1, \ldots, I_d\right\}$ to be the set of states of observations. We use $\boldsymbol{\Gamma} = \{\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{Z}\}$ to denote a parameter set. The likelihood of $\boldsymbol{A}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{Z}$ given observations $\mathcal{Y}$ is given by

$$
\begin{aligned}
L(\mathcal{Y}, \boldsymbol{\Gamma}) = & \prod_{d=1}^{D} \prod_{i=2}^{I_d} p\left(\boldsymbol{y}_i^{(d)} \mid \boldsymbol{y}_{i-1}^{(d)}, z_i^{(d)}, \boldsymbol{A}_{z_i^{(d)}}, \boldsymbol{\mu}_{z_i^{(d)}}, \boldsymbol{\Sigma}_{z_i^{(d)}}\right) p\left(\boldsymbol{A}_{z_i^{(d)}} \mid \boldsymbol{M}_0, \boldsymbol{\Sigma}_{z_i^{(d)}}, \boldsymbol{V}_0\right) \\
& p\left(\boldsymbol{\mu}_{z_i^{(d)}} \mid \boldsymbol{\mu}_0, \frac{1}{\lambda_0} \boldsymbol{\Sigma}_{z_i^{(d)}}\right) p\left(\boldsymbol{\Sigma}_{z_i^{(d)}} \mid \boldsymbol{\Phi}_0, \nu_0\right) \prod_{d=1}^{D} p\left(z_1^{(d)} \mid \boldsymbol{\pi}^*\right) \prod_{i=2}^{I_d} p\left(z_{i+1}^{(d)} \mid z_i^{(d)}, \boldsymbol{\pi}\right) p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}).
\end{aligned}
\tag{5.9}
$$

Due to the large number of observations, it is impossible to marginalize out $\boldsymbol{Z}$ from Eq. (5.9). Therefore, based on the graphical model illustrated in Figure 5.5, we derive an efficient MCMC scheme using Gibbs sampling. We start the Gibbs sampling with random initialization for all parameters, and then iteratively sample each parameter from its conditional distribution on other parameters. Posterior full conditional distributions of the parameters are derived as follows:

- **Sampling state transition probability $\boldsymbol{\pi}_k$ from $p\left(\boldsymbol{\pi}_k \mid \boldsymbol{z}^k, \boldsymbol{\alpha}\right)$.** We use $\boldsymbol{z}^k$ to denote the vector containing latent variables whose values are k. The conditional distribution is $p\left(\boldsymbol{\pi}_k \mid \boldsymbol{z}^k, \boldsymbol{\alpha}\right) \propto p\left(\boldsymbol{\pi}_k \mid \boldsymbol{\alpha}\right) p\left(\boldsymbol{z}^k \mid \boldsymbol{\pi}_k\right)$. The prior distribution $p\left(\boldsymbol{\pi}_k \mid \boldsymbol{\alpha}\right) = \text{Dirichlet}\left(\boldsymbol{\pi}_k \mid \boldsymbol{\alpha}\right) \propto \prod_{k'=1}^{K} \pi_{k,k'}^{\alpha_{k'}-1}$, and $p\left(\boldsymbol{z}^k \mid \boldsymbol{\pi}_k\right)$ can be seen as a multinomial distribution $p\left(\boldsymbol{z}^k \mid \boldsymbol{\pi}_k\right) = \text{Multinomial}_K\left(\boldsymbol{z}^k \mid N, \boldsymbol{\pi}_k\right) \propto \prod_{k'=1}^{K} \pi_{k,j}^{\left|M_{k,k'}\right|}$, where $\left|M_{k,k'}\right|$ is the number of elements in $M_{k,k'}$. Therefore, the conditional posterior distribution is a

Dirichlet distribution:

$$p\left(\boldsymbol{\pi}_k \mid \boldsymbol{z}^k, \boldsymbol{\alpha}\right) = \text{Dirichlet}\left(\left|M_{k,1}\right| + \alpha_1, \left|M_{k,2}\right| + \alpha_2, \cdots, \left|M_{k,K}\right| + \alpha_K\right). \tag{5.10}$$

- **Sampling state sequence $\boldsymbol{z}_{1:I_d}^{(d)}$ from $p\left(\boldsymbol{z}_{1:I_d}^{(d)} \mid \boldsymbol{y}_{1:I_d}^{(d)}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)$.** For each bus sequence, we sample the entire hidden discrete state sequence $\boldsymbol{z}_{1:I_d}^{(d)}$ all at once given the sequence observations $\boldsymbol{y}_{1:I_d}^{(d)}$ and parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}$. For the Markov chain model, the forward-backward algorithm is often used to sample the state sequence (Scott, 2002). In this sampling algorithm, we are interested in finding the posterior distribution $p\left(z_i^{(d)} \mid \boldsymbol{y}_{1:I_d}^{(d)}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)$ of $z_i^{(d)}$ given the sequence observations $\boldsymbol{y}_{1:I_d}^{(d)}$ and parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}$. With Bayes' theorem, we have

$$p\left(z_i^{(d)} \mid \boldsymbol{y}_{1:I_d}^{(d)}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) = \frac{p\left(\boldsymbol{y}_{1:I_d}^{(d)} \mid z_i^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) p\left(z_i^{(d)} \mid \boldsymbol{\pi}\right)}{p\left(\boldsymbol{y}_{1:I_d}^{(d)} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)}. \tag{5.11}$$

Using the conditional independence property in Figure 5.5, we obtain

$$\begin{aligned} p\left(z_i^{(d)} \mid \boldsymbol{y}_{1:I_d}^{(d)}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) &= \frac{p\left(\boldsymbol{y}_{1:i}^{(d)}, z_i^{(d)} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) p\left(\boldsymbol{y}_{i+1:I_d}^{(d)} \mid z_i^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)}{p\left(\boldsymbol{y}_{1:I_d}^{(d)} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)} \\ &= \frac{\alpha\left(z_i^{(d)}\right) \beta\left(z_i^{(d)}\right)}{p\left(\boldsymbol{y}_{1:I_d}^{(d)} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)}, \end{aligned} \tag{5.12}$$

where we define $\alpha\left(z_i^{(d)}\right)$ and $\beta\left(z_i^{(d)}\right)$ as follows

$$\alpha\left(z_i^{(d)}\right) = p\left(\boldsymbol{y}_{1:i}^{(d)}, z_i^{(d)} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right), \tag{5.13}$$

$$\beta\left(z_i^{(d)}\right) = p\left(\boldsymbol{y}_{i+1:I_d}^{(d)} \mid z_i^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right). \tag{5.14}$$

We then derive recursive relationships that allow $\alpha\left(z_i^{(d)}\right)$ and $\beta\left(z_i^{(d)}\right)$ to be evaluated efficiently. The relationship between $\alpha\left(z_i^{(d)}\right)$ and $\alpha\left(z_{i-1}^{(d)}\right)$ can be derived as

$$\begin{aligned}\alpha\left(z_i^{(d)}\right) &= p\left(\boldsymbol{y}_i^{(d)} \mid \boldsymbol{y}_{i-1}^{(d)}, z_i^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) \sum_{z_{i-1}^{(d)}} p\left(\boldsymbol{y}_{1:i-1}^{(d)} \mid z_{i-1}^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) p\left(z_i^{(d)} \mid z_{i-1}^{(d)}\right) p\left(z_{i-1}^{(d)}\right) \\ &= p\left(\boldsymbol{y}_i^{(d)} \mid \boldsymbol{y}_{i-1}^{(d)}, z_i^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) \sum_{z_{i-1}^{(d)}} \alpha\left(z_{i-1}^{(d)}\right) p\left(z_i^{(d)} \mid z_{i-1}^{(d)}\right) \\ &= \mathcal{N}\left(\boldsymbol{y}_i^{(d)} \mid \boldsymbol{A}_{z_i^{(d)}} \boldsymbol{y}_{i-1}^{(d)} + \boldsymbol{\mu}_{z_i^{(d)}}, \boldsymbol{\Sigma}_{z_i^{(d)}}\right) \sum_{z_{i-1}^{(d)}} \alpha\left(z_{i-1}^{(d)}\right) \pi_{z_{i-1}^{(d)}}\left(z_i^{(d)}\right). \end{aligned} \tag{5.15}$$

We can similarly derive the recursive relationship for the quantities $\beta\left(z_i^{(d)}\right)$ as follows

$$\begin{aligned}\beta\left(z_i^{(d)}\right) &= \sum_{z_{i+1}^{(d)}} p\left(\boldsymbol{y}_{i+2:I_d}^{(d)} \mid z_{i+1}^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) p\left(\boldsymbol{y}_{i+1}^{(d)} \mid \boldsymbol{y}_i^{(d)}, z_{i+1}^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) p\left(z_{i+1}^{(d)} \mid z_i^{(d)}\right) \\ &= \sum_{z_{i+1}^{(d)}} \beta\left(z_{i+1}^{(d)}\right) p\left(\boldsymbol{y}_{i+1}^{(d)} \mid \boldsymbol{y}_i^{(d)}, z_{i+1}^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) p\left(z_{i+1}^{(d)} \mid z_i^{(d)}\right) \\ &= \sum_{z_{i+1}^{(d)}} \beta\left(z_{i+1}^{(d)}\right) \mathcal{N}\left(\boldsymbol{y}_{i+1}^{(d)} \mid \boldsymbol{A}_{z_{i+1}^{(d)}} \boldsymbol{y}_i^{(d)} + \boldsymbol{\mu}_{z_{i+1}^{(d)}}, \boldsymbol{\Sigma}_{z_{i+1}^{(d)}}\right) \pi_{z_i^{(d)}}\left(z_{i+1}^{(d)}\right). \end{aligned} \tag{5.16}$$

Considering that the left-hand side in Eq. (5.12) is a normalized distribution, the quantity $p\left(\boldsymbol{y}_{1:I_d}^{(d)} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)$ can be obtained as follows:

$$p\left(\boldsymbol{y}_{1:I_d}^{(d)} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) = \sum_{z_i^{(d)}} \alpha\left(z_i^{(d)}\right) \beta\left(z_i^{(d)}\right). \tag{5.17}$$

After running the recursion from $i = 1, \ldots, I_t$ to obtain $\alpha\left(z_1^{(d)}\right), \ldots, \alpha\left(z_{I_d}^{(d)}\right)$ and the recursion from $i = I_t, \ldots, 1$ to obtain $\beta\left(z_1^{(d)}\right), \ldots, \beta\left(z_{I_t}^{(d)}\right)$, then we could evaluate $p\left(z_i^{(d)} \mid \boldsymbol{y}_{1:I_d}^{(d)}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)$ and sample the state sequence $\boldsymbol{z}_{1:I_d}^{(d)}$.

- **Sampling mean and covariance $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ from $p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathcal{Y}_k, \Theta, \boldsymbol{A}_k\right)$.** Thanks to the conjugate prior distribution, the conditional distribution of the mean vector and the covariance matrix $p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathcal{Y}_k, \Theta, \boldsymbol{A}_k\right)$ is a Gaussian-inverse-Wishart distribution:

$$p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathcal{Y}_k, \Theta, \boldsymbol{A}_k\right) = \mathcal{N}\left(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0^*, \frac{1}{\lambda_0^*}\boldsymbol{\Sigma}_k\right) \mathcal{W}^{-1}\left(\boldsymbol{\Sigma}_k \mid \boldsymbol{\Psi}_0^*, \nu_0^*\right), \tag{5.18}$$

where

$$
\begin{aligned}
&\boldsymbol{\mu}_0^* = \frac{\lambda_0 \boldsymbol{\mu}_0 + |M_k|\,\boldsymbol{\delta}}{\lambda_0 + |M_k|}, \quad \lambda_0^* = \lambda_0 + |M_k|, \quad \nu_0^* = \nu_0 + |M_k|, \\
&\boldsymbol{\delta} = \frac{1}{|M_k|} \sum_{i,d \in M_k} \left(\boldsymbol{y}_i^{(d)} - \boldsymbol{A}_k \boldsymbol{y}_{i-1}^{(d)} \right), \quad \boldsymbol{\Psi}_0^* = \boldsymbol{\Psi}_0 + \mathbf{S} + \frac{\lambda_0 |M_k|}{\lambda_0 + |M_k|} \left(\boldsymbol{\delta} - \boldsymbol{\mu}_{\mathbf{0}} \right) \left(\boldsymbol{\delta} - \boldsymbol{\mu}_{\mathbf{0}} \right)^\top, \\
&\mathbf{S} = \sum_{i,d \in M_k} \left(\boldsymbol{y}_i^{(d)} - \boldsymbol{A}_k \boldsymbol{y}_{i-1}^{(d)} - \boldsymbol{\delta} \right) \left(\boldsymbol{y}_i^{(d)} - \boldsymbol{A}_k \boldsymbol{y}_{i-1}^{(d)} - \boldsymbol{\delta} \right)^\top .
\end{aligned}
\tag{5.19}
$$

- **Sampling coefficient matrix $\boldsymbol{A}_k$ from** $p\left(\boldsymbol{A}_k \mid \mathcal{Y}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$**.** As we use the conjugate prior distribution, the conditional distribution of the coefficient matrix $p\left(\boldsymbol{A}_k \mid \mathcal{Y}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ is a matrix Gaussian distribution:

$$
p\left(\boldsymbol{A}_k \mid \mathcal{Y}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \mathcal{MN}\left(\boldsymbol{A}_k \mid \boldsymbol{M}_0^*, \boldsymbol{\Sigma}_k, \boldsymbol{V}_0^*\right), \tag{5.20}
$$

where

$$
\boldsymbol{V}_0^* = \left(\boldsymbol{V}_0^{-1} + \sum_{i,d \in M_k} \boldsymbol{y}_{i-1}^{(d)} \boldsymbol{y}_{i-1}^{(d)\,\top} \right)^{-1}, \tag{5.21}
$$

$$
\boldsymbol{M}_0^* = \left(\boldsymbol{M}_0 \boldsymbol{V}_0^{-1} + \sum_{i,d \in M_k} \left(\boldsymbol{y}_i^{(d)} - \boldsymbol{\mu}_k \right) \boldsymbol{y}_{i-1}^{(d)\,\top} \right) \boldsymbol{V}_0^* . \tag{5.22}
$$

Finally, we summarize the Gibbs sampling procedure to estimate the parameters in Algorithm 6 seen in Appendix A. We drop the first $n_1 = 9000$ iterations as burn-in and then store samples of parameters $\boldsymbol{\pi}_k$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, $\boldsymbol{A}_k$ from the following $n_2 = 1000$ iterations. For hyperparameters, we set $\boldsymbol{\mu}_0 = \mathbf{0}_{2n+1}$, $\lambda_0 = 2$, $\boldsymbol{\Phi}_0 = \boldsymbol{I}_{2n+1}$, $\boldsymbol{V}_0 = \boldsymbol{I}_{2n+1}$, $\nu_0 = 2n+3$, and $\boldsymbol{\alpha} = 0.2 \times \boldsymbol{I}_K$, where n is the number of bus links. Note that model training is in fact offline based on historical data, and only Markov chains (i.e. samples) of the parameters are used in the forecasting task. We code the MCMC algorithm using Python with Numpy and Scipy packages.

5.5 Probabilistic Forecasting

We categorize the links of each bus into two groups: observed links and upcoming links, during the forecasting process. Observed links refer to the links that the bus has already traversed, and we know their respective travel times and passenger occupancies. On the other hand, upcoming links are the links that the bus is yet to traverse, and we need to forecast their travel times and passenger occupancies. The crucial step of the forecasting process is to determine the hidden states of the buses using the available observed information. Once hidden states are identified, it becomes convenient to make probabilistic forecasts for travel time and occupancy for downstream links conditional on observed information. Through the model training process, we have obtained a sample from the posterior distribution of the parameters. Using sampling techniques, we can make estimations about the hidden states of buses using the information that is currently observed. Forecasting can be achieved by approximating the distribution of unobserved/future variables conditional on the observed link travel time and passenger occupancy. Typically, there are upcoming links for both the following bus (bus j) and the leading bus (bus $(j-1)$). While it is possible to make forecasts solely based on the observed links, we adopt an autoregressive approach. This approach incorporates the forecasting of the upcoming links of the leading bus to forecast the travel time and passenger occupancy of the following bus. When considering the interdependence among buses, our forecasting method offers a more comprehensive and accurate estimation.

We use Figure 5.6 to illustrate the forecasting process. Figure 5.6 (a) presents the scenario where bus $j-1$ has completed its run, bus j has traversed the initial two links and arrived at stop #3, and bus $j+1$ has left the origin stop but has not yet arrived at stop #2. Our initial task is to estimate the hidden states of these buses (from bus 1 to bus $j+1$) based on the observed link travel times and passenger occupancies. Similarly, we used observed link information and headway to forecast the subsequent links for bus j and bus $j+1$ based on the estimated states. For bus j, we can make predictions by using its upstream links (the first two links), all the observed link travel times/passenger occupancies of the leading bus $j-1$, and the starting headway of the bus j. Concerning bus $j+1$, the observed upstream link travel times/passenger occupancies of bus j are used, along with the forecasts of the upcoming link travel times/passenger occupancies of bus j, and the observed headway. As time passes, the forecasting of relevant buses can be updated upon receiving new observed links. At the time point illustrated in Figure 5.6 (b), bus $j+1$ obtains a new observed link, enabling us to update the hidden states and perform prediction. In this example, the states of all buses remain unchanged, but with additional observed links, the forecasting could become more accurate. At the subsequent time point depicted in Figure 5.6 (c), bus j acquires a new observed link, which requires an update of the bus states. With more observed information, it is possible for the states of certain

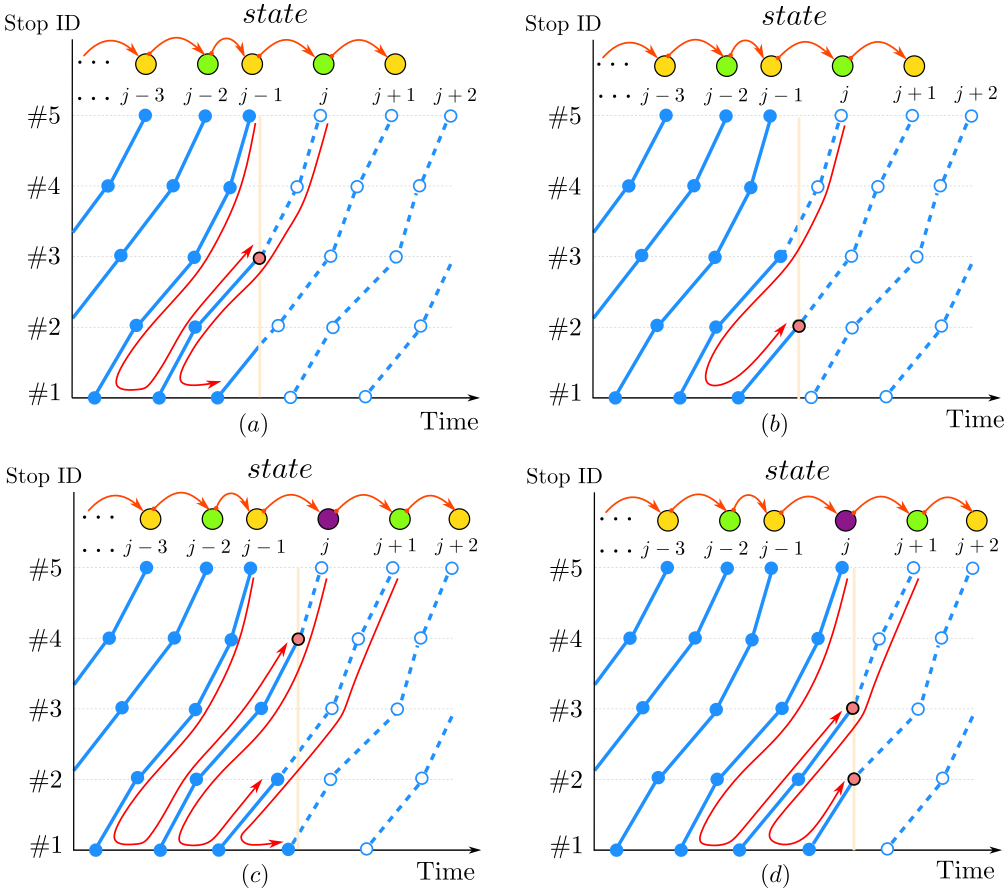


Figure 5.6: Representation of probabilistic forecasting. Solid blue circles: observed link travel time and passenger occupancy at bus stops. The hollow circles: unknown arrival time and passenger occupancy at the bus stop. Solid red circles: new observed arrival time and passenger occupancy at bus stops. Solid green, orange, and purple circles represent different hidden states. Each sub-figure is a new round of forecasting triggered by a new observation of bus arrival. For each bus, we forecast the bus's upcoming link travel times and passenger occupancies (dashed blue lines) based on observed and previously forecast link travel times and passenger occupancies (blue lines covered by a red arrow area).

buses to change, resulting in improved accuracy. In this case, the state of bus j changes from green to purple, and consequently we can update the forecast. It should be noted that even though bus $j+1$ does not have any additional observed links, updated information about its leading bus (i.e., bus j) can reinforce/enhance its forecasting. Furthermore, we can see that a new bus $j+2$ starts its trip on the route; with the observed headway, we could make forecasting for bus $j+2$. Figure 5.6 (d) demonstrates the scenario where bus j has completed its run, and bus $j+1$ and $j+2$ each obtain a new observed link. In this case, we can generate forecasts for bus $j+1$ and $j+2$ following the aforementioned procedure.

Consider a specific bus j, and the observed links (travel time, passenger occupancy, headway) can be organized into a vector denoted as $\boldsymbol{y}_j^o$, representing partial observations. Our goal is to forecast the unobserved links of bus j, and we denote the vector we want to forecast as $\boldsymbol{y}_j^f$. Therefore, the full variable of bus j is $\boldsymbol{y}_j = \left[\boldsymbol{y}_j^{o\top}, \boldsymbol{y}_j^{f\top}\right]^\top$. For most time-series problems, the Markov regime vector autoregressive model often makes forecasting based only on previous observations. For example, forecasts of $\boldsymbol{y}_j^f$ are derived solely from its immediate predecessor $\boldsymbol{y}_{j-1}$. This conventional approach relies on the dynamics inherent in the vector autoregressive structure and the probabilistic transitions between the hidden states. At any time, in additional to the observation $\boldsymbol{y}_{j-1}$, we also have partial observations $\boldsymbol{y}_j^o$ and $\boldsymbol{y}_{j+1}^o$. Our proposed forecasting approach will incorporate all observations (that is, $\boldsymbol{y}_{j-1}$, $\boldsymbol{y}_j^o$, $\boldsymbol{y}_{j+1}^o$) to enhance the forecast for the unobserved part $\boldsymbol{y}_j^f$ of bus j. With the collected n_2 samples $\left\{\boldsymbol{\theta}^{(\rho)}\right\}_{\rho=1}^{n_2}$ of parameters during the model estimation stage, we can utilize the Gibbs sampling method to obtain the predicted distribution (probabilistic forecasting) for unobserved variables. The predictive distribution over the unobserved part $\boldsymbol{y}_j^f$ of bus j given the observed data can be approximated by Monte Carlo estimation:

$$
\begin{aligned}
& p\left(\boldsymbol{y}_j^f \mid \boldsymbol{y}_{1:j-1}, \boldsymbol{y}_j^o, \boldsymbol{y}_{j+1}^o\right) \\
& = \iiiint p\left(\boldsymbol{y}_j^f \mid -\right) p\left(z_j, z_{j+1} \mid \boldsymbol{y}_j^o, \boldsymbol{y}_{1:j-1}, \boldsymbol{y}_{j+1}^o, \boldsymbol{\pi}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\pi}\right) p\left(\boldsymbol{\theta}\right) dz_j dz_{j+1} d\boldsymbol{\pi} d\boldsymbol{\theta} \\
& \approx \frac{1}{n_2} \sum_{\rho=1}^{n_2} p\left(\boldsymbol{y}_j^f \mid \boldsymbol{y}_j^o, \boldsymbol{y}_{j-1}, \boldsymbol{y}_{j+1}^o, z_j^{(\rho)}, z_{j+1}^{(\rho)}, \boldsymbol{\mu}^{(\rho)}, \boldsymbol{\Sigma}^{(\rho)}, \boldsymbol{A}^{(\rho)}\right).
\end{aligned} \tag{5.23}
$$

Figure 5.7 shows the overall graphical representation of probabilistic forecasting. The sampling scheme for probabilistic forecasting is as follows.

For the sequence state inference, we use the same forward-backward sampling method as in Section 5.4.5. Next, we need to sample $\boldsymbol{y}_j^f$ from $p\left(\boldsymbol{y}_j^f \mid \boldsymbol{y}_j^o, \boldsymbol{y}_{j-1}, \boldsymbol{y}_{j+1}^o, z_j, z_{j+1}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right)$. Note that in sampling $\boldsymbol{y}_j^f$ we build the conditional distribution on $\boldsymbol{y}_{j+1}^o$ instead of $\boldsymbol{y}_{j+1}$ for efficient sampling. Here, by using the characteristics of Gaussian linear systems, we can easily obtain the joint distribution of $\boldsymbol{y}_j$ and $\boldsymbol{y}_{j+1}$ as

$$
p\left(\left[\begin{array}{c} \boldsymbol{y}_j \\ \boldsymbol{y}_{j+1} \end{array}\right] \middle| \boldsymbol{y}_{j-1}, z_j, z_{j+1}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{L}\right), \tag{5.24}
$$

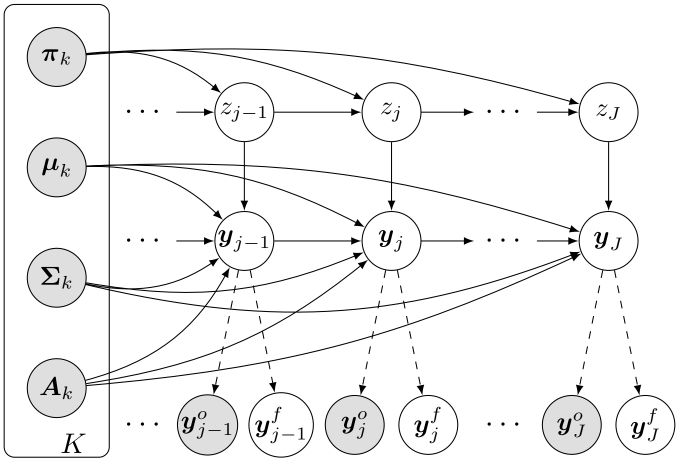


Figure 5.7: Graphical representation of probabilistic forecasting. Each bus has the partial observations represented by a vector $\boldsymbol{y}^o_j$. Our goal is to forecast the unobserved links of the buses, and the vector we want to forecast is represented by $\boldsymbol{y}^f_j$.

where

$$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{A}_{z_j}\boldsymbol{y}_{j-1} + \boldsymbol{\mu}_{z_j} \\ \boldsymbol{A}_{z_{j+1}}(\boldsymbol{A}_{z_j}\boldsymbol{y}_{j-1} + \boldsymbol{\mu}_{z_j}) + \boldsymbol{\mu}_{z_{j+1}} \end{bmatrix}, \boldsymbol{L} = \begin{bmatrix} \boldsymbol{\Sigma}_{z_j} & \boldsymbol{\Sigma}_{z_j}\boldsymbol{A}^\top_{z_{j+1}} \\ \boldsymbol{A}_{z_{j+1}}\boldsymbol{\Sigma}_{z_j} & \boldsymbol{A}_{z_{j+1}}\boldsymbol{\Sigma}_{z_j}\boldsymbol{A}^\top_{z_{j+1}} + \boldsymbol{\Sigma}_{z_{j+1}} \end{bmatrix}. \quad (5.25)$$

In this joint distribution, we have partial observations $\boldsymbol{y}^o = \left[\boldsymbol{y}^{o\top}_j, \boldsymbol{y}^{o\ \top}_{j+1}\right]^\top$. Based on the joint distribution, we can directly derive the conditional distribution of unobserved vectors $\boldsymbol{y}^f_j$:

$$p\begin{pmatrix} \boldsymbol{y}^f_j \\ \boldsymbol{y}^o \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}_f \\ \boldsymbol{m}_o \end{bmatrix}, \begin{bmatrix} \boldsymbol{L}_{f,f} & \boldsymbol{L}_{f,o} \\ \boldsymbol{L}_{o,f} & \boldsymbol{L}_{o,o} \end{bmatrix}\right). \quad (5.26)$$

The conditional distribution of $\boldsymbol{y}^f_j$ given $\boldsymbol{y}^o$ is:

$$p\left(\boldsymbol{y}^f_j \mid \boldsymbol{y}^o_j, \boldsymbol{y}_{j-1}, \boldsymbol{y}^o_{j+1}, z_j, z_{j+1}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}\right) = p\left(\boldsymbol{y}^f_j \mid \boldsymbol{y}^o\right) = \mathcal{N}\left(\boldsymbol{y}^f_j \mid \boldsymbol{m}_{f|o}, \boldsymbol{L}_{f|o}\right), \quad (5.27)$$

where $\boldsymbol{m}_{f|o}$ and $\boldsymbol{L}_{f|o}$ are the conditional mean and covariance matrix, respectively, given by:

$$\boldsymbol{m}_{f|o} = \boldsymbol{m}_f + \boldsymbol{L}_{f,o}\boldsymbol{L}^{-1}_{o,o}\left(\boldsymbol{y}^o - \boldsymbol{m}_o\right), \quad (5.28)$$

$$\boldsymbol{L}_{f|o} = \boldsymbol{L}_{f,f} - \boldsymbol{L}_{f,o}\boldsymbol{L}^{-1}_{o,o}\boldsymbol{L}_{o,f}. \quad (5.29)$$

By collecting forecasting samples from all parameter groups, we can approximate posterior predictive distributions for travel times and passenger occupancies of upcoming links. We summarize the Gibbs sampling algorithm for probabilistic forecasting in Appendix B.

5.6 Experiments

In this section, we conduct a comprehensive evaluation of our proposed model using real-world datasets. We also undertake a comparative analysis with existing models to highlight the superior performance of our approach. In addition, we explore an examination of parameter patterns to further substantiate our findings. The source code used for these experiments can be accessed from https://github.com/xiaoxuchen/Markov-Regime-switching-Model.

5.6.1 Experiment Settings

As the measurements have different units, we first perform data standardization (z-score normalization) so that all variables are centered at 0 with a standard deviation of 1. By doing so, we can better model and learn the covariance matrix. For example, $\ell_{i,m}$ (the m-th link travel time of the i-th bus) can be rescaled/standardized with

$$\tilde{\ell}_{i,m} = \frac{\ell_{i,m} - \mu_{\ell_m}}{\sigma_{\ell_m}}, \tag{5.30}$$

where μ_{ℓ_m} is the mean of travel time at the m-th link; σ_{ℓ_m} is the standard deviation of travel time at the m-th link. The dataset encompasses a period of four consecutive weeks, specifically focusing on weekdays (Monday to Friday), which amounts to a total of 20 days. We use the first 15 days for model inference, and the remaining 5 days to validate the model forecasting. In the inference, we can estimate the parameters of the model and understand the underlying structure or process that generates the observed data. The forecasting validation is crucial for evaluating how well our proposed model performs on unseen data. It helps in assessing the model's predictive accuracy and generalizability. In this experiment, we make probabilistic forecasting for bus travel time and passenger occupancy to evaluate the proposed model.

5.6.2 Performance Metrics

We use the root mean squared error (RMSE) and the mean absolute error (MAE) to evaluate the performance for point estimation based on the mean. We use the continuous rank probability score (CRPS) to evaluate the performance of probabilistic forecasting.

- RMSE and MAE are defined as:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \end{aligned} \tag{5.31}$$

  where $y_i, \hat{y}_i, i = 1, \ldots, n$ are the true values and forecasts, respectively.

- The continuous rank probability score is often used as a quantitative measure of probabilistic forecasting; it is defined as the quadratic measure of the discrepancy between the predicted cumulative distribution function (CDF, denoted by F_X) and $\mathbb{I}(x \geqslant y)$, the empirical CDF of the observation y:

$$\text{CRPS}(F_X, y) = \int_{-\infty}^{\infty} \left(F_X(x) - \mathbb{I}(x \geqslant y)\right)^2 dx, \tag{5.32}$$

  where $\mathbb{I}(\cdot)$ is the indicator function, which is defined as follows: If the condition inside the parentheses is true, then $\mathbb{I}(\cdot)$ equals 1. We use the average CRPS of all observations as one metric. Essentially, the CRPS calculates the mean squared difference between the predicted probabilities and the observed outcome, integrated over all possible threshold values. Lower CRPS values indicate better forecast accuracy.

5.6.3 Model Comparison

In this paper, we compare the performance of our proposed model with the Bayesian time-dependent Gaussian mixture model developed in Chen et al. (2023). Here, we detail the method outlined in Section 5.2. Figure 5.8 shows the overall graphical representation of the time-dependent Bayesian Gaussian mixture model. The random variable at the t-th period follows a multivariate Gaussian mixture model:

$$p^t\left(\boldsymbol{y}^t\right) = \sum_{k=1}^{K} \pi_k^t \mathcal{N}\left(\boldsymbol{y}^t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{5.33}$$

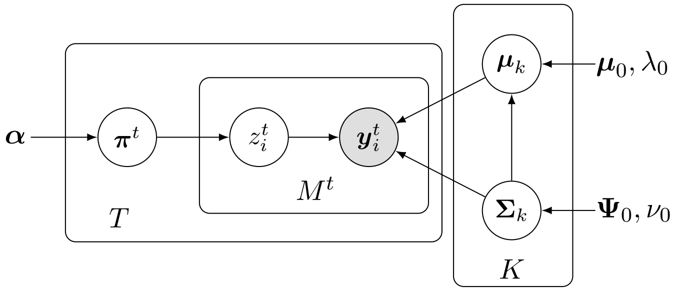


Figure 5.8: Graphical representation of Bayesian time-dependent Gaussian mixture model.

where the superscript $(\cdot)^t$ denotes the time period, K is the number of components, $0 \leqslant \pi_k^t \leqslant 1$ is a mixing coefficient with $\sum_{k=1}^{K} \pi_k^t = 1$, and each of the K components follows a multivariate Gaussian distribution.

The random variable of each period is characterized by a mixture of several shared Gaussian distributions. In the graphical model, z_i^t is a component label, indicating which component $\boldsymbol{y}_i^t$ belongs to. In a Bayesian setting, they use a conjugate Gaussian-inverse-Wishart prior on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ and a Dirichlet prior on $\boldsymbol{\pi}^t$ for efficient inference. The overall data generation process is summarized as:

$$\boldsymbol{\pi}^t \sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right), \tag{5.34}$$

$$\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi}_0, \nu_0\right), \tag{5.35}$$

$$\boldsymbol{\mu}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda_0}\boldsymbol{\Sigma}_k\right), \tag{5.36}$$

$$z_i^t \sim \text{Categorical}\left(\boldsymbol{\pi}^t\right), \tag{5.37}$$

$$\boldsymbol{y}_i^t \mid z_i^t = k \sim \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{5.38}$$

where $\boldsymbol{\alpha}$ is the concentration parameter of the Dirichlet distribution; $\mathcal{W}^{-1}\left(\boldsymbol{\Psi}_0, \nu_0\right)$ is the inverse-Wishart distribution with a scale matrix $\boldsymbol{\Psi}_0$ and ν_0 degrees of freedom; $\boldsymbol{\mu}_0$ and λ_0 are parameters for the Gaussian prior.

Next, we design the following models to demonstrate the effect of interactions between bus travel time and passenger occupancy and dependencies between hidden states on probabilistic forecasting for bus travel time and passenger occupancy.

- BGMM-S: Apply Bayesian time-dependent Gaussian mixture model to make bus travel time and passenger occupancy forecasting separately. We develop two independent models: one model for bus travel time and the other model for passenger occupancy. These independent models are the particular cases where the correlation between travel time and passenger occupancy is not considered.

- BGMM-J: Apply a Bayesian time-dependent Gaussian mixture model to make bus travel time and passenger occupancy forecasting jointly. We model the bus travel time and passenger occupancy as a random variable and therefore could consider the interactions between them.
- MSAR-S: Apply Bayesian Markov regime-switching vector autoregressive model to make bus travel time and passenger occupancy forecasting separately. Similarly, there are two independent models: one model for bus travel time and the other model for passenger occupancy.
- MSAR-J: This is our proposed model, and we utilize the Bayesian Markov regime-switching vector autoregressive model to jointly forecast bus travel time and passenger occupancy.

5.6.4 Forecast Performance

We apply Algorithm 6 to estimate model parameters and Algorithm 7 to make probabilistic forecasting for bus travel time and passenger occupancy. We test models with different numbers of clusters/states ($K = 1, 5, 10, 20, 30, 40, 50$) to select an optimal state number. For each model, we start with an initial value $K = 1$ and evaluate performance. When performance improves, we will select a larger K to evaluate the model again and continue this process until there is no substantial improvement in performance. Table 5.1 shows the performance of probabilistic forecasting for bus travel time and passenger occupancy with different models. First, we can see that all models show improved performance as the number of clusters/states (K) increases, indicating that they are significant in forecasting bus travel time and passenger occupancy. Second, we can observe that RSMM-J and BGMM-J outperform MSAR-S and MSAR-S, which demonstrates the importance of joint modeling of bus travel time and passenger occupancy. This finding shows that considering the interactions between bus travel time and passenger occupancy can significantly improve probabilistic forecast performance for bus travel time and passenger occupancy. Third, we can see that MSAR-S/J outperforms BGMM-S/J, which indicates that modeling states transition/connection could help make better probabilistic forecasting for bus travel time and passenger occupancy.

Table 5.1: Performance of probabilistic forecasting of link travel time, passenger occupancy, and trip travel time.

		Link travel time (sec)			Passenger occupancy (pax)			Trip travel time (sec)		
		RMSE	MAE	CRPS	RMSE	MAE	CRPS	RMSE	MAE	CRPS
BGMM-S	$K = 1$	64.31	51.32	32.44	10.05	8.13	6.20	258.83	215.61	175.18
	$K = 5$	54.51	42.67	29.60	9.16	7.34	5.87	246.89	201.36	157.85
	$K = 10$	47.41	36.82	27.44	8.00	6.89	5.21	236.14	189.39	144.83
	$K = 20$	45.73	35.54	25.91	7.81	6.70	5.43	218.65	176.94	132.44
	$K = 30$	**40.53**	**31.49**	**20.59**	7.67	6.60	5.37	**199.49**	**161.29**	**115.78**
	$K = 40$	43.96	33.97	22.28	**6.94**	**5.95**	**4.95**	212.09	172.35	125.93
BGMM-J	$K = 1$	47.23	35.71	27.25	8.97	7.01	5.45	221.09	184.14	135.29
	$K = 5$	44.63	34.35	25.69	8.64	6.92	5.28	213.53	174.15	131.21
	$K = 10$	38.94	30.28	19.96	6.46	5.27	4.26	183.32	149.51	106.41
	$K = 20$	36.87	28.32	19.65	6.16	5.04	3.96	180.88	143.22	103.60
	$K = 30$	24.16	17.28	17.63	5.76	4.69	3.84	170.43	112.47	79.89
	$K = 40$	**18.25**	**13.02**	**14.35**	**4.53**	**3.60**	**3.57**	**164.22**	**102.23**	**72.36**
	$K = 50$	18.41	13.06	14.64	4.60	3.74	3.75	166.80	105.52	73.99
MSAR-S	$K = 1$	62.87	49.48	31.15	10.23	8.53	6.32	257.11	214.04	174.87
	$K = 5$	51.86	40.46	25.32	7.65	6.46	5.25	235.50	194.67	143.99
	$K = 10$	42.61	32.73	21.74	6.89	5.87	4.73	205.99	169.55	120.36
	$K = 20$	**34.45**	**25.96**	**19.59**	**6.35**	**5.39**	**4.27**	**195.90**	**156.32**	**109.32**
	$K = 30$	39.09	30.28	20.01	6.58	5.61	4.53	205.60	157.92	112.60
MSAR-J	$K = 1$	48.32	35.99	27.45	9.20	7.03	5.48	222.28	184.37	135.60
	$K = 5$	39.38	30.36	20.39	6.43	5.45	4.82	197.60	157.43	114.00
	$K = 10$	30.16	22.57	18.25	5.27	4.41	3.90	190.47	141.01	105.90
	$K = 20$	18.35	13.38	14.36	4.86	4.07	3.79	164.47	103.55	72.60
	$K = 30$	**16.11**	**11.66**	**12.14**	**3.48**	**2.92**	**3.07**	**137.13**	**83.48**	**57.98**
	$K = 40$	17.02	12.57	13.37	4.50	3.98	3.71	153.35	95.72	63.06

Best results are highlighted in bold fonts.

5.6.5 Interpreting Analysis

Bayesian models are powerful tools for interpreting parameters and uncovering patterns in probabilistic forecasting of bus travel time and passenger occupancy. In our study, as depicted in Figure 5.9, we showcase the estimated transition matrix, a cornerstone of the Markov regime-switching model. Each element of this matrix provides insight into the probability of transitioning from one state to another. The structure and values within this matrix are instrumental in understanding how frequently and likely certain state transitions occur, which in turn, can be linked to specific conditions or patterns in bus travel time and occupancy. We can see that most buses would like to keep or transit to state 2. Furthermore, Figure 5.10 presents the estimated coefficient matrices of different states and we can find that they have different patterns, indicating that different states show different relationships between adjacent buses. Figure 5.11 and Figure 5.12 illustrate the estimated mean vectors and covariance matrices of the random error term. The visual representation of mean vectors gives us an understanding of the central tendencies of errors across different states. The covariance matrices, on the other hand, unravel the relationships between errors of different variables. These visualizations reveal clear distinctions among different states in both mean values and covariance matrices.

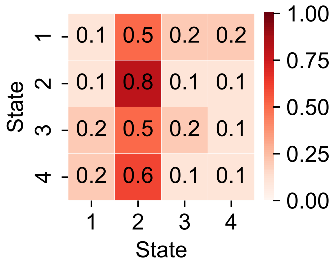


Figure 5.9: Estimated transition matrix.

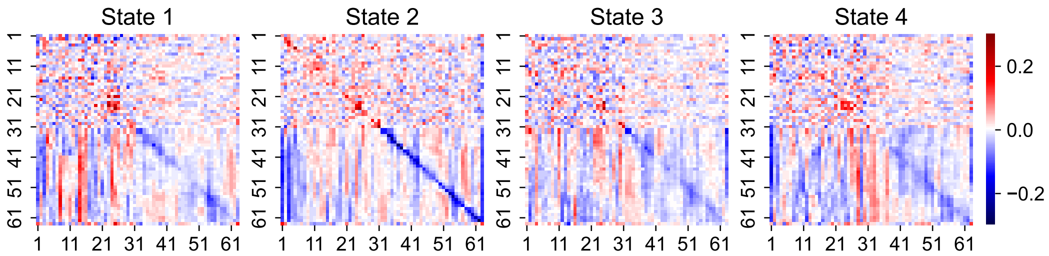


Figure 5.10: Estimated coefficient matrices.

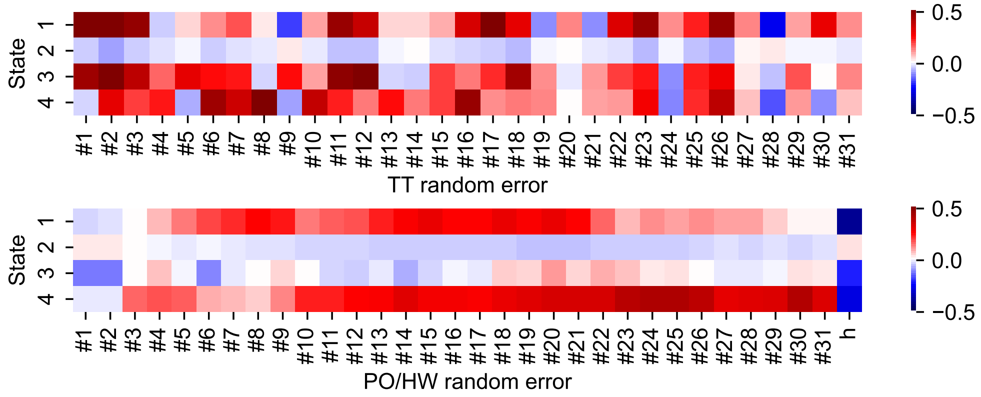


Figure 5.11: Estimated mean vectors of the random error term.

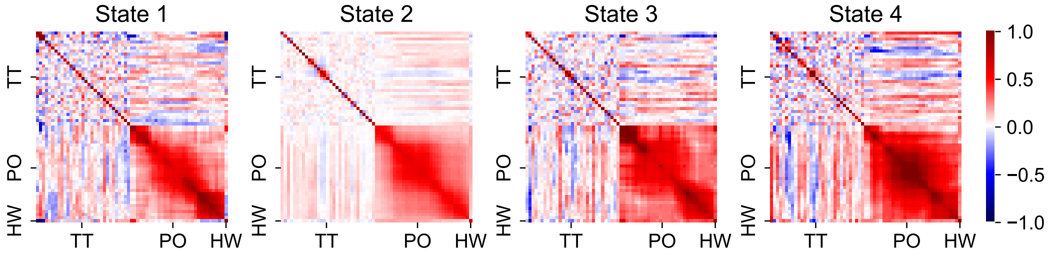


Figure 5.12: Estimated covariance matrices of the random error term.

5.6.6 Forecast Distribution

Figure 5.13 shows that the last bus has arrived at stop #14 and the goal is to provide predicted distributions for travel time and passenger occupancy of downstream links (that is, from stop 14 to 32), and provide trip travel time/arrival time distributions. In this figure, we plot the predicted trajectories with passenger occupancy. In particular, the model generates multiple outcomes for each bus run, depicted by the spread of five sampled trajectories, which collectively offer a distribution that encapsulates the possible variance in travel times and occupancy. This spread of predictions signifies the model's robust approach to capturing the uncertainties inherent in the bus systems.

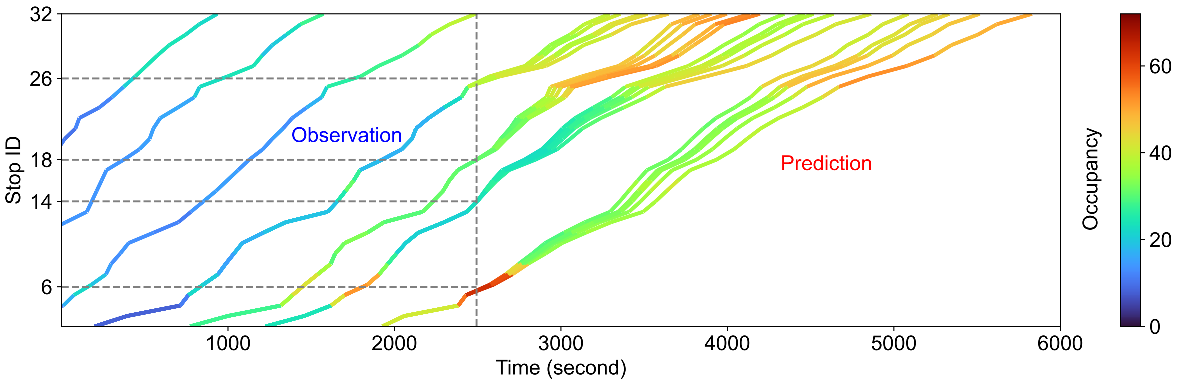


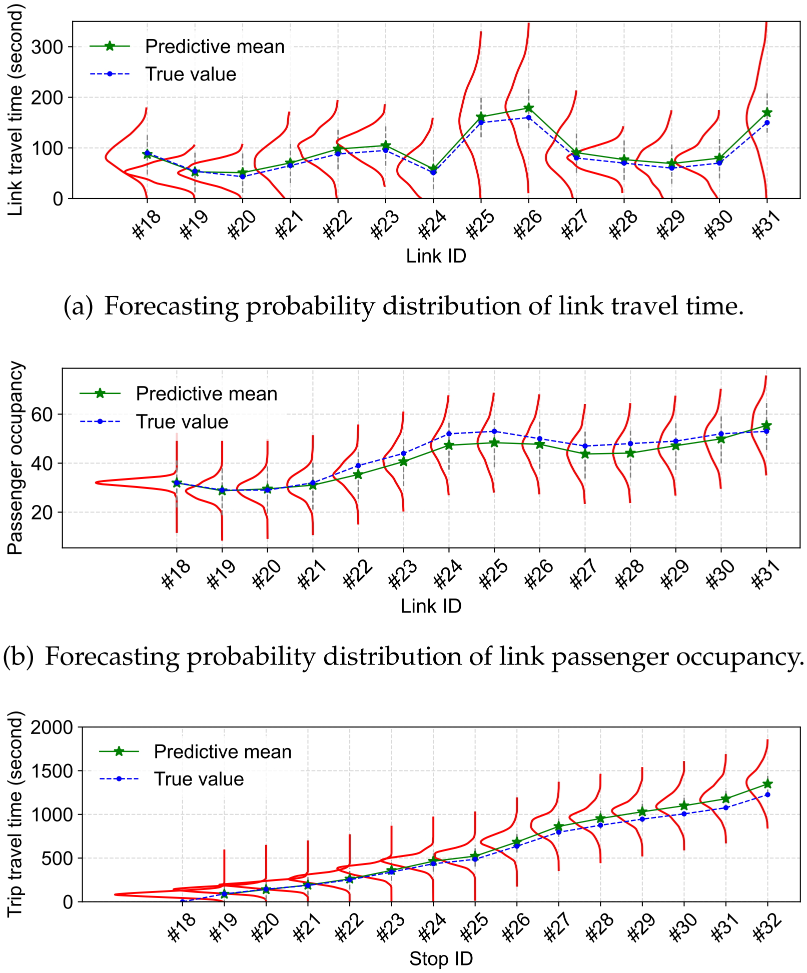
Figure 5.13: Trajectory plot to show the bus travel time and occupancy forecasting. The vertical line represents the current time that separates past and future. Each colored curve shows the trajectory of one bus run, with color indicating the number of passengers onboard (i.e., occupancy). For each bus, we plot five samples of the predicted downstream travel time and passenger occupancy based on the proposed model.

We present visualizations of predicted probability distributions for a specific bus at a particular time point. Using a sampling method, obtaining the predicted trip travel time for the bus becomes straightforward. The visualization encompasses forecasting probability distributions for link travel time, passenger occupancy, and trip travel time, as demonstrated in Figure 5.14. In this figure, the bus has already traversed the first 17 links and our goal is to forecast the next 14 links (from #18 to #31). The blue points represent the true values, while the green points represent the predictive mean values. The first two panels show the predicted probability distributions for link travel times and passenger occupancy. Evidently, the predictive means closely align with the true values, which confirms the good accuracy of our forecasting for both bus link travel time and passenger occupancy. Furthermore, we observe that link travel times with larger values tend to exhibit larger variances, indicated by the larger variance in those red density functions. Additionally, upcoming links situated near the current links have smaller variances, suggesting that more precise predictions. The bottom panel shows the forecasting probability distributions for trip travel times. As the number of links in a trip increases, we notice that the red bell curves become more spread out, reflecting an increased variance in trip travel time. This is because longer trips inherently introduce more uncertainty in travel time predictions.

5.7 Discussion

In this paper, we propose a Bayesian Markov regime-switching vector autoregressive model for probabilistic forecasting of bus travel time and passenger occupancy. Our approach can effectively capture and address several critical factors: the correlations between travel time and passenger occupancy, the relationship between adjacent buses, and the multimodality/skewness of bus travel time and passenger occupancy distributions. To validate our proposed model, we conduct extensive numerical experiments on a real-world dataset. Our results demonstrate the superiority of the proposed approach compared to benchmark models and its effectiveness in providing accurate forecasts for bus travel time and passenger occupancy.

Our approach has implications for both practice and research. First, the proposed Bayesian Markov regime-switching vector autoregressive model could forecast trip travel time (i.e., estimated time of arrival) and passenger occupancy distributions, which could be incorporated into real-time bus information systems to help passengers and bus agencies make better decisions. Second, the proposed Bayesian model is also an interpretable tool for bus agencies to better understand bus operation patterns with uncertainty. For example, if one has access to enormous historical bus operation data (including some special events such as extreme weather, sporting events, large-scale concerts, etc.), we could also learn the pattern of special events, which would be helpful for bus agencies to provide better



(a) Forecasting probability distribution of link travel time.

(b) Forecasting probability distribution of link passenger occupancy.

(c) Forecasting probability distribution of trip travel time.

Figure 5.14: Probability distributions of forecasting travel time and passenger occupancy.

and robust management and operations. Third, our model can also be used to model other transport systems with interactions between adjacent vehicles. For instance, we can use the same model to model train/metro operation in a network and study how delay propagates. The proposed Bayesian Markov regime-switching vector autoregressive model can offer valuable insights into understanding and mitigating the delays that frequently affect train systems. Last, the proposed Bayesian Markov regime-switching vector autoregressive model could be utilized to perform imputation for series data with missing values.

5.8 Appendix

Algorithm 6 Gibbs sampling for parameter estimation.

Input: Sequential observations $\left\{\boldsymbol{y}_i^{(d)}\right\}_{i=1,d=1}^{I_d,D}$, hyperparameters Θ and $\boldsymbol{\alpha}$, random initialization of sequential states $\left\{z_i^{(d)}\right\}_{i=1,d=1}^{I_d,D}$, iterations n_1, n_2.

Output: Samples of transition matrix $\left\{{\boldsymbol{\pi}_k}^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$, samples of mean vectors $\left\{\boldsymbol{\mu}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$, and samples of covariance matrices $\left\{\boldsymbol{\Sigma}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$.

```
for iter = 1 to $n_1 + n_2$ do
    for $k = 1$ to $K$ do
        Draw $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ according to Eq. (5.5) and Eq. (5.6).
        Draw $\boldsymbol{A}_k$ according to Eq. (5.7).
        if iter $> n_1$ then
            Collect $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{A}_k$ to the output sets.
        end if
    end for
    for $k = 1$ to $K$ do
        Draw $\boldsymbol{\pi}_k$ according to Eq. (5.8).
        if iter $> n_1$ then
            Collect $\boldsymbol{\pi}_k$ to the output set.
        end if
    end for
    Conduct the forward-backward algorithm to obtain $\alpha(\cdot)$ and $\beta(\cdot)$.
    for $d = 1$ to $T$ do
        Calculate $\boldsymbol{\pi}^*$ and draw $z_1^{(d)}$ according to Eq. (5.4).
        for $i = 2$ to $I_d$ do
            Calculate $p\left(z_i^{(d)}\right)$ according to Eq. (5.12).
            Draw $z_i^{(d)}$ according to Eq. (5.3).
        end for
    end for
    for $k = 1$ to $K$ do
        Update the parameters $\Theta = \{\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0\}$ by Eq. (5.19).
    end for
    Update the parameters $\boldsymbol{\alpha}$ by Eq. (5.10).
end for
return $\left\{{\boldsymbol{\pi}_k}^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$,$\left\{\boldsymbol{\mu}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$, $\left\{\boldsymbol{\Sigma}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$, $\left\{\boldsymbol{A}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$.
```

Algorithm 7 Gibbs sampling for probabilistic forecasting.

Input: Sequential observations $\left\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{j-1}, \boldsymbol{y}_j^o, \boldsymbol{y}_{j+1}^o, \ldots, \boldsymbol{y}_J^o\right\}$, samples of transition matrix $\left\{\boldsymbol{\pi}_k{}^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$, samples of mean vectors $\left\{\boldsymbol{\mu}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$, samples of covariance matrices $\left\{\boldsymbol{\Sigma}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$, and samples of coefficient matrices $\left\{\boldsymbol{A}_k^{(\rho)}\right\}_{k=1,\rho=1}^{K,n_2}$.

Output: Forecasting sample set $\left\{\boldsymbol{y}_j^{f\,(\rho)}, \ldots, \boldsymbol{y}_J^{f\,(\rho)}\right\}_{\rho=1}^{n_2}$.

```
for $\rho = 1$ to $n_2$ do
    for $j' = j$ to $J$ do
        Draw $\boldsymbol{y}_{j'}^f$ as the forecasting sample.
        Collect $\boldsymbol{y}_{j'}^f$ to the forecasting sample set.
    end for
    Calculate $\boldsymbol{\pi}^*$ and draw $z_1$ according to Eq. (5.4).
    Conduct the forward-backward algorithm to obtain $\alpha(\cdot)$ and $\beta(\cdot)$.
    for $j = 2$ to $J$ do
        Calculate $p(z_j)$ according to Eq. (5.12).
        Draw $z_j$ according to Eq. (5.3).
    end for
end for
return $\left\{\boldsymbol{y}_j^{f\,(\rho)}, \ldots, \boldsymbol{y}_J^{f\,(\rho)}\right\}_{\rho=1}^{n_2}$.
```

Chapter 6

Bayesian Inference for Time-varying Transit Origin-Destination Matrices

This chapter is a research article submitted to *Transportation Science*:

- **Chen, X.**, Cheng, Z., Sun, L. 2024. Bayesian Inference of Time-varying Origin-Destination Matrices from Boarding/Alighting Counts for Transit Services.

This chapter corresponds to the Bayesian inference model for transit origin-destination demand matrices.

6.1 Abstract

Origin-destination (OD) demand matrices are crucial for transit agencies to optimize the management and operation of transit systems. This paper introduces a temporal Bayesian model to estimate transit OD matrices at the individual bus level using counts of boarding and alighting passengers at each stop. Specifically, we model the number of alighting passengers at subsequent bus stops, given a boarding stop, by a multinomial distribution. Next, we assume that the parameters (i.e., assignment probabilities) of the multinomial distribution smoothly vary over time. Directly modeling the temporal dependencies among the parameters is difficult due to the constraint of parameters (e.g., the sum of a vector of assignment probabilities is one). To address this challenge, we introduce a latent variable matrix to parameterize the time-varying multinomial distributions through the softmax transformation. As the latent matrix contains a large number of elements, we employ matrix factorization to parameterize it into a mapping factor matrix and a temporal factor matrix, which substantially reduces the number of parameters. To encode a temporally smooth structure in the matrix, we impose Gaussian process priors on the columns of the temporal factor matrix, which ensures the alighting probabilities are smoothly time-varying. For model inference, we develop a two-stage algorithm with a Markov Chain Monte Carlo approach. In the first stage, we sample latent OD matrices conditional on parameters using a Metropolis-Hastings sampling algorithm with a Markov model-based proposal distribution. In the second stage, we sample parameters conditional on latent OD matrices using slice and elliptical slice sampling algorithms. We validate our model using real-world data of three bus routes (short, medium, long) and results demonstrate that our model can achieve accurate estimation and outperforms the iterative proportional fitting method. Moreover, our model can provide uncertainty quantification associated with estimation and parameter interpretation.

6.2 Introduction

The origin-destination (OD) matrix for a bus route captures passenger flows from one stop to another, serving as a comprehensive representation of passenger demand. These matrices can be defined either in an aggregated manner (e.g., overall demand during morning rush hours) or in detail for each bus journey. An accurate estimation of the OD matrices and a good understanding of how these matrices evolve over time are critical for transit agencies in making planning and operational decisions, such as route design (Ahern et al., 2022), service scheduling (Martínez et al., 2014), timetabling (Sun et al., 2014a), and fleet allocation (Gkiotsalitis et al., 2019).

Estimating OD matrices for transit systems from available operational data has been an important application and a long-standing research question for both practitioners and researchers (Hussain et al., 2021; Mohammed and Oke, 2023). A traditional approach to collecting transit OD matrices is through onboard surveys. This approach, as expected, is time-consuming and labor-intensive (Agrawal et al., 2017). More importantly, there could be systematic bias when using the OD matrix obtained from a specific bus journey as the base to estimate the overall demand profile due to the inherent randomness in bus operations. For example, the OD matrices of two consecutive bus journeys could be substantially different when bus bunching occurs. The emergence of advanced data collection techniques has offered alternative solutions to collect/estimate OD data. For example, automated fare collection (AFC) systems in cities such as Beijing and Singapore can register both the boarding and alighting transactions of passenger trips (Sun et al., 2014b). For such cases, the AFC data contains the complete information of the OD matrices. However, most cities around the world adopt a "tap-in-only" AFC system with no alighting information recorded, which no longer supports direct inference of OD matrices. Although there exists a large body of literature on inferring passenger alighting stops from "tap-in-only" AFC data based on the trip chain continuity assumption (see e.g., Trépanier et al., 2007; Assemi et al., 2020; Hussain et al., 2021; Mohammed and Oke, 2023), these methods are often inaccurate for infrequent transit users (Cheng et al., 2021). A more prevalent data collection technique is the Automated Passenger Counting (APC) system, which registers the boarding and alighting counts when the bus arrives at a bus stop. APC systems have been widely used by transit agencies and allow one to estimate the number of passengers onboard during the trip in real time. However, because boarding/alighting counts are in fact column/row sums of an OD matrix, we cannot directly infer the OD matrices from APC data.

The focus of our study is to infer time-varying OD matrices from boarding/alighting counts with high temporal resolution. The iterative proportional fitting (IPF) method has long been the primary solution adopted by transit agencies to estimate OD matrices from APC data. IPF estimates OD matrices by adjusting a reference/seed matrix derived from additional data sources such as onboard surveys (Ben-Akiva et al., 1985; Ji et al., 2014). Using the observed counts (i.e., row/column sums), the OD matrix is adjusted through an iterative process: For each row in the matrix, each entry is multiplied by a constant, ensuring that the sum of the row is in accordance with the actual count. Then, this adjustment is applied to each column. This process is repeated iteratively for both rows and columns until the matrix entries converge (Lamond and Stewart, 1981). There are several major issues associated with the IPF method. First, the accuracy of IPF is highly dependent on the quality of the seed matrix, while in practice creating a good seed matrix is not an easy task—it has the same number of variables as in the OD matrix. In general, these seed matrices come from onboard surveys. However, as mentioned above, the seed matrix obtained could be biased to represent and reproduce the true demand patterns of OD, and thus the quality of the seed matrix becomes very important for IPF. Second, IPF often struggles with the issue of zeros in the seed matrix (Ben-Akiva et al., 1985), which will remain zero in all iterations. In cases where the reference matrix contains an entire row or column of zeros but the corresponding actual boarding or alighting counts are not zeros, IPF will fail to find a feasible solution. Third, from a statistical perspective, the estimated matrix after convengency is no longer a "count" matrix, thus violating the underlying "count" nature of passenger demand. This solution is acceptable when we have a large number of passengers; however, at the journey level, the true OD matrix is likely to be a sparse count matrix, and modeling it using continuous values will result in biased estimation.

Bayesian statistical solutions are developed to combine prior information with observed boarding and alighting counts to achieve accurate inference (Li, 2009; Hazelton, 2010; Blume et al., 2022). In contrast to deterministic IPF, Bayesian statistical models have the capability to generate posterior distributions for the elements in the OD matrix, thus providing estimates along with their associated uncertainties. In general, these models treat OD matrix estimation as a linear inverse problem and focus on estimating the alighting probabilities instead of the counts (see e.g., Li and Cassidy, 2007; Ji et al., 2015). A landmark work in probabilistic inference is the Bayesian approach proposed by Li (2009), which incorporates a Markov model to describe the relationships between the entries of a transit OD matrix. This approach assumes that the probability of an onboard passenger alighting at the next approaching stop is independent of his/her boarding stop. In other words, the onboard passengers are assumed to be memoryless. While this innovative approach reduces the model parameters significantly and enables elegant likelihood construction, the Markov assumption is too restrictive and unrealistic for practical applications. To address this limitation, Hazelton (2010) introduced a novel Bayesian method for static OD matrix inference. A key challenge in building the statistical model is that the calculation of the likelihood of observed counts requires an enumeration of all possible OD matrices (with elements being counts) that match the marginal counts. To tackle this challenge, Hazelton

(2010) developed a two-stage sampling algorithm for model inference using the Markov chain Monte Carlo (MCMC) method. The first stage samples latent OD matrices using the Markov model by Li (2009) as the proposal distribution. The second stage samples model parameters conditional on the OD matrices in the first stage. Blume et al. (2022) developed a Bayesian inference approach to estimate the static OD matrix in large-scale networked transit systems but considering elements as continuous random variables. This problem is approached as an inverse linear regression, and the posterior distributions of OD matrix entries are estimated using Hamiltonian Monte Carlo. Overall, the statistical methods summarized above are essentially designed for static OD inference, and they generally require many observed bus journeys to estimate one single OD matrix. As a result, it is infeasible to apply these methods directly to infer time-varying OD matrices. It should be noted that although Hazelton (2010) presented results for time-varying OD demand, the estimation is achieved by fitting several static models over a day.

The challenges in inferring time-varying OD matrices from APC data are summarized as follows: (1) Transit OD flows need to rely on discrete count models and the likelihood structures are more complicated than continuous models. (2) The distributions of OD matrices are not static, and time-varying models are needed to characterize how demand evolves over time. For example, we expect the OD demand matrices in two consecutive hours to be similar, while the matrix during the morning rush hour should be substantially different from the one during the evening rush hour. When considering individual buses, the variability in OD matrices between adjacent bus journeys can depend on the reliability and uniformity of the operation. In scenarios where services are regular and consistent, the OD matrices observed from two consecutive journeys are likely to be similar. However, in cases of service interruptions or irregularities, such as bus bunching or delays, the OD matrices can differ significantly from one journey to the next. (3) A model that takes entries in the OD matrix as parameters will involve a larger number of parameters and become difficult to estimate, not to mention when estimating time-varying OD matrices. (4) Due to the underdetermined nature of the linear inverse problem, the set of feasible solutions for OD matrices is extraordinarily large. The application of likelihood-based methods for discrete count data in this context would involve an exhaustive enumeration of all possible OD matrices and therefore will face practical computational limitations, especially when considering bus routes with a large number of stops. For a given set of boarding/alighting counts along a bus journey, Figure 6.1 shows the real OD matrix together with potential solutions estimated with different assumptions. We can observe that these solutions could be substantially different from each other. (5) When making operational and planning

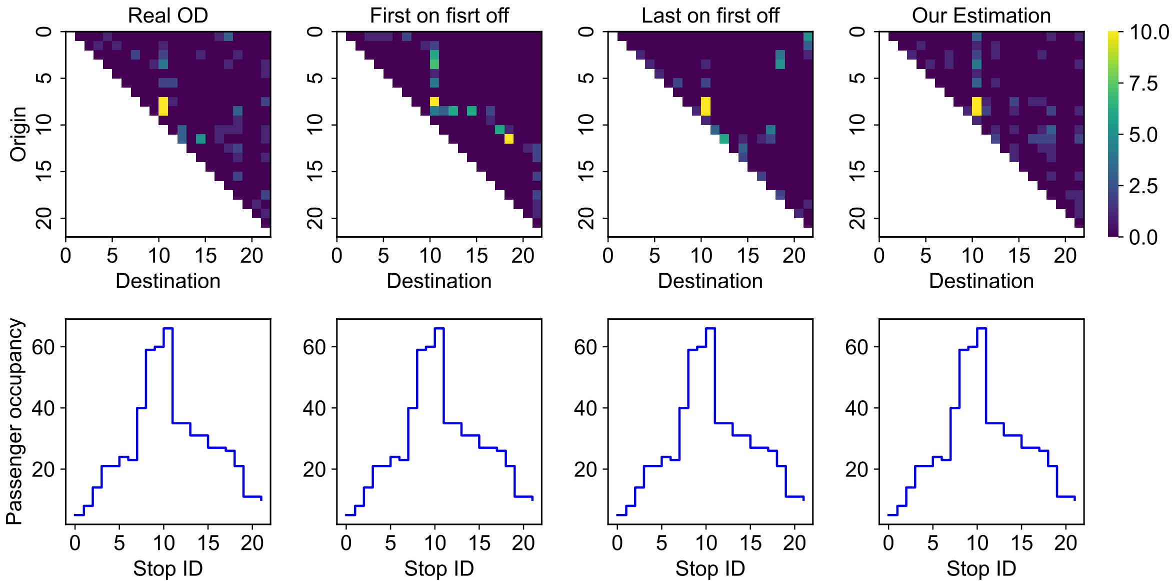


Figure 6.1: Illustration of the uncertainty of OD solution. There are two rows: the top row presents the OD matrices and the bottom row shows the passenger occupancy. The first OD matrix is the real OD matrix of a specific bus trip. The last three OD matrices are estimated with different methods. The second matrix is estimated using the "First on first off" principle which assumes that passengers who board first will alight first. The third matrix is estimated with the "Last on first off" principle which assumes that the last passengers to board are the first ones to alight at subsequent stops. The last matrix is estimated from our proposed method and it shows a similar pattern to the real OD matrix.

decisions, we are more interested in the distribution of the underlying OD demand, rather than a point estimate.

To address these challenges, in this paper, we extend the work of Hazelton (2010) and develop a temporal Bayesian model for inferring transit OD matrices at the individual bus level. To model the discrete count data, we assume that the number of alighting passengers at subsequent bus stops, given a boarding stop, follows a multinomial distribution. To better characterize the temporal patterns in passenger demand, we assume that the parameters (i.e., assignment probabilities) of the multinomial distribution vary smoothly over time, thus allowing for building a time-varying model using counts observed from a limited number of bus journeys. We introduce a latent variable matrix and use it to parameterize the time-varying multinomial distributions through the softmax transformation. In addition, we propose using matrix factorization to parameterize the latent matrix as the product of a mapping factor matrix and a temporal factor matrix, which substantially reduces the number of parameters. To encode a temporally smooth structure in the matrix, we impose Gaussian process priors on the columns of the temporal factor matrix, which consequently ensure that the assignment probabilities vary smoothly over time. For model inference, we follow Hazelton (2010) and also develop a two-stage algorithm based on MCMC. In the first stage, we sample latent OD matrices conditional on parameters using the Metropolis-Hastings sampling algorithm with the proposal distribution proposed by Hazelton (2010), which efficiently bypasses the need to enumerate the large number of feasible OD matrices that align with observed boarding and alighting counts for each bus trip. In the second stage, we sample model parameters conditional on latent OD matrices obtained from the first stage. The key challenge in this step is to efficiently sample latent Gaussian processes with non-Gaussian likelihood, where the posterior no longer has an analytical formulation. To address this issue, we use the efficient elliptical slice sampling (ESS) algorithms developed by Murray et al. (2010) to sample the temporal factor matrix. We evaluate our proposed model using real-world APC data and true OD matrices from three bus routes in an anonymous city. We compare the performance of the proposed temporal model to a non-temporal variant, and the results show that the temporal Bayesian model outperforms the non-temporal variant, confirming the importance and value of developing a time-varying model. In addition, we also compare our model with the widely used IPF method, and the results show that our model can achieve superior performance in deterministic estimation.

The remainder of this paper is organized as follows. In Section 6.3, we define the problem and introduce the notation used throughout the paper. In Section 6.4, we introduce the proposed temporal Bayesian model. We elaborate on the theoretical underpinnings of our approach and explain how it addresses those identified challenges. Section 6.5 develops an efficient inference algorithm based on MCMC, in which elliptical slice sampling is used to sample the temporal factor matrix. We then evaluate the effectiveness and performance of our proposed model based on real-world data in Section 6.6. Finally, Section 6.7 summarizes our key findings and discusses future research directions.

6.3 Problem Definition

We follow Hazelton (2010) to define the OD matrix inference problem using boarding/alighting counts. Consider a bus route comprising S stops at which passengers can board and alight. Let u_i and v_i denote the numbers of boarding passengers and alighting passenger at stop i, respectively, for $i = 1, 2, \dots, S$. Such boarding/alighting counts are available from the APC systems. In general, we will see neither alighting passengers at stop 1 nor boarding passengers at stop S, so we can fix $v_1 = u_S = 0$.

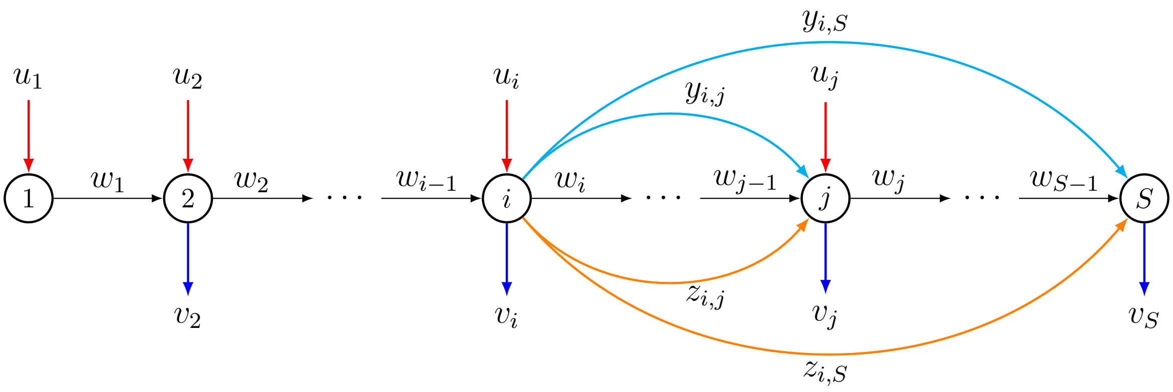


Figure 6.2: A graphical representation of notations in a bus route. There is a bus route with S stops. u_i and v_i are boarding and alighting counts of passengers at stop i, respectively; $y_{i,j}$ is the number of passengers boarding at stop i and alighting at stop j; w_i is the number of passengers on the bus immediately after leaving stop i; $z_{i,j}$ is the number of passengers boarded at stop i and are on the bus as it approaches stop j.

For a bus journey (i.e., a trip from stop 1 to stop S), we denote by $y_{i,j}$ the number of passengers who board at stop i and alight at stop j, which cannot be observed directly. We define w_i as the number of passengers on the bus immediately after leaving stop i. This can be expressed recursively as

$$w_i = w_{i-1} + u_i - v_i \qquad (i = 1, 2, \dots, S)\,, \tag{6.1}$$

with the initial condition $w_0 = 0$. Let $z_{i,j}$ represent the unobserved number of passengers who board at stop i and remain on the bus as it approaches stop j. The relationships among these variables are given by

$$z_{j,j+1} = u_j, \tag{6.2}$$

$$z_{i,j+1} = z_{i,j} - y_{i,j} \qquad (i = 1, 2, \dots, j-1)\,, \tag{6.3}$$

$$w_j = \sum_{i=1}^{j} z_{i,j}. \tag{6.4}$$

These notations of a bus route are graphically represented in Figure 6.2. Let $\boldsymbol{u} = (u_1, u_2, \dots, u_S)^\top$ and $\boldsymbol{v} = (v_1, v_2, \dots, v_S)^\top$ be the vectors of boarding and alighting counts at the stops, respectively; we then denote by $\boldsymbol{x} = \left(\boldsymbol{u}^\top, \boldsymbol{v}^\top\right)^\top = (u_1, u_2, \dots, u_S, v_1, v_2, \dots, v_S)^\top$ as the aggregation of observed counts for a bus trip.

Our study aims to infer the OD matrix $\boldsymbol{Y} = \left(y_{i,j}\right)_{S\times S}$. For a bus route/service, it is clear that passengers can only travel to downstream stops. Thus, we fix $y_{i,j} = 0$ for all cases where $i \geqslant j$, and focus exclusively on the upper-triangular part of $\boldsymbol{Y}$. Following Hazelton (2010), we denote by $\boldsymbol{y}_i$ the number of passengers traveling from the i-th stop to the subsequent stops along the bus route. For instance, $\boldsymbol{y}_1 = (y_{1,2}, y_{1,3}, \ldots, y_{1,S})^\top$ denotes the passenger counts from the initial stop to all subsequent stops along the route. This definition continues to $\boldsymbol{y}_{S-1} = (y_{S-1,S})$, which represents the number of passengers traveling from the second-to-last stop to the last stop. Next, we stack these passenger counts into a single OD vector $\boldsymbol{y} = \left(\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_{S-1}^\top\right)^\top \in \mathbb{R}^{S(S-1)/2}$. Although $\boldsymbol{y}$ is not directly observable, its relationship with the observed boarding and alighting counts can be expressed as follows:

$$\sum_{j=i+1}^{S} y_{i,j} = u_i \qquad (i = 1, 2, \ldots, S-1), \tag{6.5}$$

and

$$\sum_{i=1}^{j-1} y_{i,j} = v_j \qquad (j = 2, \ldots, S-1). \tag{6.6}$$

This relationship can be encapsulated as

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{y}, \tag{6.7}$$

where both $\boldsymbol{x}$ and $\boldsymbol{y}$ are count-valued vectors, and $\mathbf{A}$ is a $2S \times M$ binary routing matrix defined by

$$a_{i,j} = \begin{cases} 1, & \text{if } i = 1, \ldots, S \text{ and} \\ & j = S(i-1) - i(i+1)/2 + k \\ & \text{for } k = i+1, \ldots, S, \\ 1, & \text{if } i = S+1, \ldots, 2S \text{ and} \\ & j = S(k-2) - k(k+1)/2 + i \\ & \text{for } k = 1, \ldots, i-S-1, \\ 0, & \text{otherwise.} \end{cases} \tag{6.8}$$

Notably, the S-th and $(S+1)$-th rows of $\mathbf{A}$ contain only zero elements, corresponding to the non-existent boarding and alighting counts at the terminal and initial stops, respectively. Although these two rows of $\mathbf{A}$ are redundant, they can maintain a direct correspondence between the row indices of the matrix and the bus stop numbers. Using a bus route with six stops as an example, the linear relationship between the observation $\boldsymbol{x}$ and the OD vector $\boldsymbol{y}$ can be expressed as

$$\underbrace{\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} y_{1,2} \\ y_{1,3} \\ y_{1,4} \\ y_{1,5} \\ y_{1,6} \\ y_{2,3} \\ y_{2,4} \\ y_{2,5} \\ y_{2,6} \\ y_{3,4} \\ y_{3,5} \\ y_{3,6} \\ y_{4,5} \\ y_{4,6} \\ y_{5,6} \end{bmatrix}}_{\boldsymbol{y}}. \tag{6.9}$$

To account for multiple bus journeys, we extend the notation to include a bus index n, which represents the n-th bus journey, with $\boldsymbol{x}^n = \left(\boldsymbol{u}^{n\top}, \boldsymbol{v}^{n\top}\right)^\top$ and $\boldsymbol{y}^n = \left(\boldsymbol{y}_1^{n\top}, \ldots, \boldsymbol{y}_{S-1}^{n}{}^\top\right)^\top$. To effectively model the dynamic/time-varying nature of OD matrices/vectors, our model incorporates temporal information. Specifically, we denote by t^n the departure time at the initial stop for the n-th bus trip/journey. For a total of N bus journeys over a studied period, we define $\mathcal{X} = \{\boldsymbol{x}^n \mid n = 1, 2, \ldots, N\}$ as the set of observed boarding and alighting counts and $\boldsymbol{t} = \left(t^1, t^2, \ldots, t^N\right)^\top$ as the vector of observed departure time. The primary objective is to estimate the set of OD vectors, denoted as $\mathcal{Y} = \{\boldsymbol{y}^n \mid n = 1, 2, \ldots, N\}$, using observed data set $\mathcal{X}$ and $\boldsymbol{t}$. This problem is challenging because the number of unknown quantities (OD vector) is much larger than the number of observations (boarding and alighting counts) in the linear system expressed by Eq. (6.7), resulting in a challenging statistical linear inverse problem (Vardi, 1996; Hazelton, 2010). We denote by $\mathcal{H}\left(\boldsymbol{x}^n\right) = \{\boldsymbol{y}^n \mid \boldsymbol{x}^n = \mathbf{A}\boldsymbol{y}^n\}$ the solution space that encompasses all feasible OD vectors consistent with the observation $\boldsymbol{x}^n$. In general, the solution space could be very large even for a route with a modest number of stops.

6.4 Bayesian Model Formulation

6.4.1 Likelihood for Static and Time-varying Models

Hazelton (2010) provides two general parameterizations for the static OD inference problem. The first approach models the entries in the OD matrix using Poisson distributions and introduces a set of intensity parameters. The second approach considers each row in the OD matrix as a realization from a multinomial distribution, and treats the alighting probabilities for each bus stop as model parameters. In this paper, we follow the second approach due to the reasons discussed in Section 6.2. For example, the exact OD entries in two consecutive matrices (e.g., $y_{i,j}^n$ and $y_{i,j}^{n+1}$) could vary substantially due to operational randomness and factors such as bus bunching, and it becomes problematic to use a fixed Poisson intensity to model both observations. On the other hand, we can safely assume that the alighting probabilities are universal for both journeys, and the variation in $y_{i,j}$ is due to the variation in the boarding counts u_i.

In terms of notation, let $\lambda_{i,j}^n$ be the probability that a passenger boarding at stop i will alight at stop j during the n-th bus trip. Furthermore, let $\boldsymbol{\lambda}_i^n = \left(\lambda_{i,i+1}^n, \ldots, \lambda_{i,S}^n\right)^\top$ be the alighting probabilities of downstream stops for a passenger boarding at stop i, and the sum of these probabilities is one, i.e., $\sum_{j=i+1}^S \lambda_{i,j}^n = 1$. Next, let $\boldsymbol{\lambda}^n = \left({\boldsymbol{\lambda}_1^n}^\top, \ldots, {\boldsymbol{\lambda}_{S-1}^n}^\top\right)^\top$ denote probabilities for all the corresponding OD entries of the n-th bus trip. Assuming that passengers make decisions independently, $\boldsymbol{y}_i^n$ follows a multinomial distribution $\boldsymbol{y}_i^n \sim \text{Multinomial}\left(u_i^n, \boldsymbol{\lambda}_i^n\right)$. Specifically, it can be represented as

$$p\left(\boldsymbol{y}_i^n \mid u_i^n, \boldsymbol{\lambda}_i^n\right) = u_i^n! \prod_{j=i+1}^{S} \frac{{\lambda_{i,j}^n}^{y_{i,j}^n}}{y_{i,j}^n!}, \tag{6.10}$$

and the likelihood of observing $\boldsymbol{x}^n$ becomes

$$\begin{aligned} L\left(\boldsymbol{\lambda}^n\right) = p\left(\boldsymbol{x}^n \mid \boldsymbol{\lambda}^n\right) &= \sum_{\boldsymbol{y}^n} p\left(\boldsymbol{x}^n \mid \boldsymbol{y}^n, \boldsymbol{\lambda}^n\right) p\left(\boldsymbol{y}^n \mid \boldsymbol{\lambda}^n\right) \\ &= \sum_{\boldsymbol{y}^n \in \mathcal{H}(\boldsymbol{x}^n)} p\left(\boldsymbol{y}^n \mid \boldsymbol{\lambda}^n\right) \\ &= \sum_{\boldsymbol{y}^n \in \mathcal{H}(\boldsymbol{x}^n)} \prod_{i=1}^{S-1} u_i^n! \prod_{j=i+1}^{S} \frac{{\lambda_{i,j}^n}^{y_{i,j}^n}}{y_{i,j}^n!}. \end{aligned} \tag{6.11}$$

Clearly, such a model is not identifiable if we have only one bus journey. Hazelton (2010) assumes that over a certain period of time we have access to repeated and independent observations of $\boldsymbol{x}$ following the same distribution with parameters $\boldsymbol{\lambda}$. Under such an assumption, $\boldsymbol{\lambda}$ becomes identifiable. For modeling multiple bus journeys in a day, we expect $\boldsymbol{\lambda}_i^n$ to vary smoothly from one bus to the next (or over time). In this case, we need an effective parameterization that produces time-varying multinomial probabilities. Since $\boldsymbol{\lambda}_i$ is the parameter of a multinomial distribution, Hazelton (2010) suggested using conjugate Dirichlet priors for $\boldsymbol{\lambda}_i$, which can model the uncertainty about $\boldsymbol{\lambda}_i$. However, in practice, it becomes challenging to adapt the Dirichlet distribution and encode temporal dynamics to generate time-varying samples of $\{\boldsymbol{\lambda}_i^n\}$.

6.4.2 Parametrization of Time-varying Assignment Probabilities

As mentioned, although the Dirichlet distribution is a natural prior for modeling $\boldsymbol{\lambda}_i$, it is difficult to adapt it to dynamic/time-varying processes. To address this issue and effectively characterize the time-varying nature of $\boldsymbol{\lambda}_i^n$, we employ a natural softmax parameterization:

$$\boldsymbol{\lambda}_i^n = \text{Softmax}\left(\rho \boldsymbol{G}_i^n\right) = \begin{pmatrix} \frac{\exp\left(\rho G_{i,i+1}^n\right)}{1+\sum_{j=i+1}^{S-1}\exp\left(\rho G_{i,j}^n\right)} \\ \frac{\exp\left(\rho G_{i,i+2}^n\right)}{1+\sum_{j=i+1}^{S-1}\exp\left(\rho G_{i,j}^n\right)} \\ \vdots \\ \frac{\exp\left(\rho G_{i,S-1}^n\right)}{1+\sum_{j=i+1}^{S-1}\exp\left(\rho G_{i,j}^n\right)} \\ \frac{1}{1+\sum_{j=i+1}^{S-1}\exp\left(\rho G_{i,j}^n\right)} \end{pmatrix}, \tag{6.12}$$

where $\boldsymbol{G}_i^n = \left(G_{i,i+1}^n, G_{i,i+2}^n, \dots, G_{i,S-1}^n\right)^\top \in \mathbb{R}^{S-i-1}$, and $\rho > 0$ is the temperature parameter, which can help to learn good sharpness/smoothness of the probability distribution. Next, we denote the collection of $\boldsymbol{G}_i^n$ over N bus journeys by the matrix

$$\boldsymbol{G}_i = \left[\boldsymbol{G}_i^1, \boldsymbol{G}_i^2, \dots, \boldsymbol{G}_i^N\right]. \tag{6.13}$$

Now we can see that $\boldsymbol{G}_i$ contains $(S-i-1)\times N$ parameters to be estimated. It should also be noted that there is no need to introduce $\boldsymbol{G}_{S-1}$ as there is only one possible alighting stop, i.e., stop S, and we always have $\lambda_{S-1,S}^n = 1$. For dynamic models, a general approach in the literature is to impose a state-space model with Gaussian noise, for instance:

$$\boldsymbol{G}_i^n \mid \boldsymbol{G}_i^{n-1} \sim \mathcal{N}\left(\boldsymbol{G}_i^{n-1}, \sigma^2 \boldsymbol{I}\right). \tag{6.14}$$

This parameterization of time-varying multinomial probabilities has been used in dynamic topic models (Blei and Lafferty, 2006). However, for the entire bus route, we have to create a dynamic model for each bus stop $i = 1, \ldots, S-2$, and the number of variables in the latent state for each bus becomes $(S-1)(S-2)/2$. Despite having a simple formulation, this parameterization shows several critical issues. First, we have a non-Gaussian state-space model in which the likelihood of observing $\boldsymbol{x}^n$ is computationally intractable (see Eq. (6.11)). Although the likelihood of $\boldsymbol{y}^n$ can be computed, estimating $\boldsymbol{G}_i$ requires filtering based on the non-Gaussian likelihood, which becomes computationally prohibitive given the large dimensionality of the latent state. Second, the state transition model in Eq. (6.14) assumes that the $(S-1)(S-2)/2$ latent states vary independently over time, which ignores potential structures over space and time. For example, in transportation systems, it is likely that the probabilities $\lambda_{i,k}$ and $\lambda_{j,k}$ $(i \neq j)$ share similar temporal patterns, which is determined by the land-use profile of stop k.

To address these issues, we next introduce an alternative parameterization for $\boldsymbol{G}_i$. Specifically, we assume $\boldsymbol{G}_i$ has a low-rank structure:

$$\boldsymbol{G}_i = \boldsymbol{\Phi}_i \boldsymbol{\Psi}^\top = \sum_{d=1}^{D} \phi_{i,d} \psi_d^\top, \tag{6.15}$$

where $\boldsymbol{\Phi}_i \in \mathbb{R}^{(S-i-1)\times D}$, $\boldsymbol{\Psi_d} \in \mathbb{R}^{N\times D}$, and $\phi_{i,d}$ and ψ_d are the d-th column of $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Psi}$, respectively. Stacking Eq. (6.15) for bus stops $i = 1, \ldots, S-2$ together, we have

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_1 \\ \vdots \\ \boldsymbol{G}_{S-2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_1 \\ \vdots \\ \boldsymbol{\Phi}_{S-2} \end{bmatrix} \times \boldsymbol{\Psi}^\top = \boldsymbol{\Phi}\boldsymbol{\Psi}^\top, \tag{6.16}$$

where we refer to $\boldsymbol{\Phi}$ as the *mapping factor matrix* and $\boldsymbol{\Psi}$ as the *temporal factor matrix*. The low-rank assumption posits that $D \ll N$ and $D \ll (S-2)(S-1)/2$, so that the factorization of $\boldsymbol{G}$ substantially reduces the number of parameters.

In order to encode temporal smoothness in $\boldsymbol{G}$, we assume that each column ψ_d in $\boldsymbol{\Psi}$ is generated from a latent Gaussian process by taking values at bus departure times $\boldsymbol{t}$ with kernel/covariance function $k_d\left(t, t'; \boldsymbol{\eta}_d\right)$ where $\boldsymbol{\eta}_d$ is the vector of kernel hyperparameters.

For example, a widely used kernel function that can produce smooth functions is the squared-exponential kernel:

$$k\left(t, t'; l, \sigma^2\right) = \sigma^2 \exp\left(-\frac{(t-t')^2}{2l^2}\right), \tag{6.17}$$

with two hyperparamters—lengthscale l and variance σ^2. We further assume the mean of the Gaussian process to be zero, and this gives for $d = 1, \ldots, D$

$$\boldsymbol{\psi}_d \sim \mathcal{N}\left(\mathbf{0}_N, \boldsymbol{K}_d\right),\ [\boldsymbol{K}_d]_{ij} = k_d(t_i, t_j; \boldsymbol{\eta}_d). \tag{6.18}$$

The Gaussian process factor model specified in Eqs. (6.16) and (6.18) provides an effective framework to model high-dimensional processes with a temporal structure, and it has been extensively used to model high-dimensional spatial and temporal data (see e.g., in Lopes et al., 2008; Luttinen and Ilin, 2009; Lei et al., 2022). The specification of $\boldsymbol{\psi}_d$ in Eq. (6.18) facilitates temporally-smooth variations of G_i. Consequently, the multinomial probability $\boldsymbol{\lambda}_i$ also exhibits smooth temporal variations. This specification assumes that $\boldsymbol{\lambda}_i^n \approx \boldsymbol{\lambda}_i^{n+1}$, which corresponds to a homogeneous assumption between the two groups of passengers who board bus n and bus $n+1$, respectively, at stop i. This assumption is reasonable, considering that both groups of passengers essentially arrive at stop i at the same time.

6.4.3 Prior Specification

In our numerical experiment, for simplicity, we assume that all columns in $\boldsymbol{\Psi}$ are generated independently from the same Gaussian process with a squared-exponential kernel, $\boldsymbol{\eta}_d = \{\sigma, l\}$, with $\sigma = 1$ and $l = 3600$ sec. We set $\sigma = 1$ because $\boldsymbol{\Phi}\boldsymbol{\Psi}^\top = \sigma\boldsymbol{\Phi} \times \frac{1}{\sigma}\boldsymbol{\Psi}^\top$ so it is not necessary to introduce the variance hyperparameter. For lengthscale l, we can further put a prior distribution on it to make the model fully Bayesian; however, learning covariance hyperparameters in latent Gaussian process models is known to be a challenging task with convergence issues (Murray and Adams, 2010). For model simplicity, we fix $l = 1$ hour based on prior knowledge.

For factor matrix $\boldsymbol{\Phi}_i$ ($i = 1, \ldots, S-2$), we simply put an independent univariate Gaussian prior for each entry. Alternatively, we have the column $\boldsymbol{\phi}_{i,d}$ to be independent and identically distributed following a zero-mean isotropic Gaussian distribution:

$$\boldsymbol{\phi}_{i,d} \sim \mathcal{N}\left(\mathbf{0}_{S-i-1}, \sigma_0^2\boldsymbol{I}\right), \quad d = 1, \ldots, D, \tag{6.19}$$

where we set hyperparameter $\sigma_0^2 = 1$.

As the temperature parameter has to be positive, we use Gaussian prior on the log-transformed ρ:

$$\log\left(\rho\right) \sim \mathcal{N}\left(\mu_\rho, \sigma_\rho^2\right), \tag{6.20}$$

where we set hyperparameters $\mu_\rho = \ln(0.1)$ and $\sigma_\rho^2 = 1$.

6.5 Bayesian Inference

In this section, we focus on performing Bayesian inference for the proposed model using MCMC. Let $\Theta = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}, \rho\}$ be the set of model parameters. Given $\mathcal{X}$ and $\boldsymbol{t}$, our aim is to infer the posterior distributions of Θ and $\mathcal{Y}$. The joint posterior of Θ and $\mathcal{Y}$ can be specified by

$$\begin{aligned} p\left(\Theta, \mathcal{Y} \mid \mathcal{X}, \boldsymbol{t}\right) = & \, p\left(\Theta \mid \mathcal{Y}, \boldsymbol{t}\right) p\left(\mathcal{Y} \mid \mathcal{X}\right) \\ \propto & \, p\left(\Theta \mid \boldsymbol{t}\right) p\left(\mathcal{Y} \mid \Theta\right) p\left(\mathcal{Y} \mid \mathcal{X}\right) \\ \propto & \, p\left(\Theta \mid \boldsymbol{t}\right) \prod_{n=1}^{N} p\left(\boldsymbol{y}^n \mid \Theta\right) I\left(\boldsymbol{y}^n \in \mathcal{H}\left(\boldsymbol{x}^n\right)\right), \end{aligned} \tag{6.21}$$

where $I\left(E\right)$ is the indicator variable for the event E. The factorization of the posterior distribution in Eq. (6.21) naturally provides a two-stage iterative sampling algorithm: (1) draw $\mathcal{Y}$ conditional on Θ and $\mathcal{X}$, and (2) draw Θ conditional on $\mathcal{Y}$ and $\boldsymbol{t}$.

There are two critical challenges in the sampling process. First, for the sampling of OD vector $\boldsymbol{y}^n$, computing the likelihood of observing $\boldsymbol{x}^n$ in Eq. (6.11) involves the enumeration of all solutions in $\mathcal{H}\left(\boldsymbol{x}^n\right)$, which becomes computationally intractable. Second, columns in the temporal $\boldsymbol{\Psi}$ are in fact latent Gaussian processes with strong dependencies, which require careful consideration when designing the MCMC method.

For the first challenge, Hazelton (2010) has provided an effective Metropolis-Hastings sampling solution to generate a candidate OD vector $\boldsymbol{y}^n$ from $\mathcal{H}\left(\boldsymbol{x}^n\right)$ rather than trying to enumerate it. This Metropolis-Hastings sampling strategy can be directly integrated into our time-varying model without much adaptation. We briefly illustrate this approach in Section 6.5.1.

6.5.1 Conditional Sampling of OD Vectors

The subsection gives a brief summary of the method to sample the OD vector $\boldsymbol{y}^n$ conditional on parameters Θ and boarding/alighting counts $\boldsymbol{x}^n$ developed by Hazelton (2010). The conditional distribution of $\mathcal{Y}$ is given by

$$p\left(\mathcal{Y} \mid \Theta, \mathcal{X}\right) \propto \prod_{n=1}^{N} p\left(\boldsymbol{y}^n \mid \Theta\right) I\left(\boldsymbol{y}^n \in \mathcal{H}\left(\boldsymbol{x}^n\right)\right). \tag{6.22}$$

The Metropolis-Hastings approach for sampling $\boldsymbol{y}^n$ is summarized as follows.

- Sample candidate $\boldsymbol{y}^*$ from proposal distribution q^n: Let $\boldsymbol{z}_j^* = \left(z_{1,j}^*, \ldots, z_{j-1,j}^*\right)^\top$, where $z_{i,j}^*$ be the candidate number of passengers who boarded at stop i and are currently on the bus as it approaches stop j. This approach assumes that the passengers on board have the same alighting probability at any stop. Consequently, one can randomly sample $\left(y_{1,j}^*, \ldots, y_{j-1,j}^*\right)$ at stop j from $\boldsymbol{z}_j^*$ with the constraint $\sum_{i=1}^{j-1} y_{i,j}^* = v_j$. Then $\boldsymbol{z}_{j+1}^*$ can be updated with Eq. (6.2) and Eq. (6.3), which can be used for the sampling at stop $j+1$. The Markov chain $\left\{\boldsymbol{z}_j^*\right\}$ has transition probability given by

$$\pi_{j+1} = P\left(\boldsymbol{z}_{j+1}^* \mid \boldsymbol{z}_j^*\right) = \left\{\binom{w_{j-1}}{v_j}\right\}^{-1} \prod_{i=1}^{j-1} \binom{z_{i,j}^*}{y_{i,j}^*}. \tag{6.23}$$

  Through the process, we can sample $\boldsymbol{y}^*$ and calculate the probability density by $q^n\left(\boldsymbol{y}^*\right) = \prod_{j=1}^{N-1} \pi_j$.

- Update the OD vector: Accept $\boldsymbol{y}^*$ with probability $\min\left(1, \frac{p(\boldsymbol{y}^* \mid \Theta) q^n(\boldsymbol{y}^n)}{p(\boldsymbol{y}^n \mid \Theta) q^n(\boldsymbol{y}^*)}\right)$. If candidate $\boldsymbol{y}^*$ is accepted, then $\boldsymbol{y}^n$ is updated to equal $\boldsymbol{y}^*$. If the candidate is not accepted, $\boldsymbol{y}^n$ remains unchanged.

6.5.2 Conditional Sampling of Model Parameters

The second stage is to sample the model parameters, including the temperature parameter ρ, the mapping factor matrix $\boldsymbol{\Phi}$ and the temporal factor matrix $\boldsymbol{\Psi}$, conditional on $\mathcal{Y}$. For these three parameters, a straightforward solution is to use Gibbs sampling to sequentially sample from:

- $p\left(\boldsymbol{\Phi} \mid \boldsymbol{\Psi}, \rho, \mathcal{Y}\right)$,
- $p\left(\boldsymbol{\Psi} \mid \boldsymbol{\Phi}, \rho, \mathcal{Y}\right)$,

Algorithm 8 Elliptical slice sampling for each column $\boldsymbol{\psi}_d$ of factor matrix $\boldsymbol{\Psi}$.

```
Input: Current state $\boldsymbol{\psi}_d$, covariance matrix $\boldsymbol{K}_d$, likelihood function $L(\boldsymbol{\psi}_d)$
Output: a new state $\boldsymbol{\psi}'_d$
 1: Choose ellipse: $\boldsymbol{\nu} \sim \mathcal{N}\left(\mathbf{0}_N, \boldsymbol{K}_d\right)$
 2: Log-likelihood threshold: $\gamma \sim \text{Uniform}\left[0,1\right], \log c = \log L\left(\boldsymbol{\psi}_d\right) + \log \gamma$
 3: Draw an initial sampling range: $\theta \sim \text{Uniform}\left[0, 2\pi\right], \theta_{\min} = \theta - 2\pi, \theta_{\max} = \theta$
 4: $\boldsymbol{\psi}'_d = \boldsymbol{\psi}_d \cos\theta + \boldsymbol{\nu}\sin\theta$
 5: if $\log L\left(\boldsymbol{\psi}'_d\right) > \log c$ then
 6:     return $\boldsymbol{\psi}'_d$
 7: else
 8:     Shrink the sampling range and try a new point:
 9:     if $\theta \leqslant 0$ then: $\theta_{\min} = \theta$ else: $\theta_{\max} = \theta$
10:     $\theta \sim \text{Uniform}\left[\theta_{\min}, \theta_{\max}\right]$
11:     GoTo Step 4.
12: end if
```

- $p\left(\rho \mid \boldsymbol{\Psi}, \boldsymbol{\Phi}, \mathcal{Y}\right)$.

However, as mentioned, since the likelihood is multinomial, we can no longer derive the analytical posterior distributions for $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. We next introduce in detail the solution for sampling $\boldsymbol{\Psi}$.

Given the independent assumption for the columns in $\boldsymbol{\Psi}$, we can sample the whole matrix in a column-based manner. Taking the column $\boldsymbol{\psi}_d$ as an example, we can update $\boldsymbol{\psi}_d$ conditional on $\boldsymbol{\Psi}_{:,h,h\neq d}$, which represents the matrix obtained by removing the d-th column vector from $\boldsymbol{\Psi}$. The posterior distribution $p\left(\boldsymbol{\psi}_d \mid \mathcal{Y}, \boldsymbol{\Phi}, \boldsymbol{\Psi}_{:,h,h\neq d}, \rho\right)$ is proportional to the product of the multinomial likelihood $L(\boldsymbol{\psi}_d) = p\left(\mathcal{Y} \mid \boldsymbol{\psi}_d, \boldsymbol{\Psi}_{:,h,h\neq d}, \boldsymbol{\Phi}, \rho\right)$ and the Gaussian process prior specified by $\mathcal{N}\left(\boldsymbol{\psi}_d; \mathbf{0}_N, \boldsymbol{K}_d\right)$. The problem becomes sampling a latent Gaussian process in a non-conjugate setting. For such problems, Murray et al. (2010) has developed an ESS method that can efficiently explore the parameter space without the need for manually tuning of step sizes or the proposal distributions. This efficiency is achieved by proposing new samples in a manner that directly leverages the underlying correlation structure of the Gaussian process, and ESS has shown superior performance over other methods. Therefore, we use ESS to update $p\left(\boldsymbol{\psi}_d\right)$ and summarize the procedure in Algorithm 8. The calculation of likelihood $L(\boldsymbol{\psi}_d)$ is straightforward following Eq. (6.10):

$$L\left(\boldsymbol{\psi}_d\right) = p\left(\mathcal{Y} \mid \boldsymbol{\Phi}, \boldsymbol{\Psi}, \rho\right) = \prod_{n=1}^{N} p\left(\boldsymbol{y}^n \mid \boldsymbol{\lambda}^n\right), \tag{6.24}$$

where $\boldsymbol{\lambda}^n$ are computed using the current values of all parameters.

Algorithm 9 Slice sampling for temperature parameter ρ.

```
Input: Current state $\rho$, likelihood function $L(\rho)$, slice sampling scale $\epsilon$
Output: a new state $\rho'$
 1: Log-likelihood threshold: $\gamma \sim \text{Uniform}[0,1], \log c = \log L(\rho) + \log p(\rho) + \log \gamma$
 2: Draw an initial sampling range: $\kappa \sim \text{Uniform}[0,\epsilon]$, $\rho_{\min} = \rho - \kappa$, $\rho_{\max} = \rho_{\min} + \epsilon$
 3: $\rho' \sim \text{Uniform}[\rho_{\min}, \rho_{\max}]$
 4: if $\log L(\rho') + \log p(\rho') > \log c$ then
 5:     return $\rho'$
 6: else
 7:     Shrink the sampling range:
 8:     if $\rho' < \rho$ then: $\rho_{\min} = \rho'$ else: $\rho_{\max} = \rho'$
 9:     GoTo Step 3.
10: end if
```

For sampling $\mathbf{\Phi}$, a straightforward approach is to use an element-wise Metropolis-Hasting algorithm given the independent prior. However, as $\mathbf{\Phi}$ contains a large number of entries, entry-by-entry sampling is computationally too expensive due to the considerable cost of likelihood evaluation. For efficiency, we sample $\mathbf{\Phi}$ in a block-based manner, i.e., updating $\mathbf{\Phi}_i$ one by one for $i = 1, \ldots, S-2$. For a given block ϕ_i, we can once again use elliptical slice sampling to update each column $\phi_{i,d}$ in $\mathbf{\Phi}_i$. The likelihood term can be computed in the same way, and the key difference from the procedure for ψ_d is that the prior distribution becomes $\mathcal{N}\left(\phi_{i,d}; \mathbf{0}_{S-i-1}, \sigma_0^2 \boldsymbol{I}\right)$.

For the temperature parameter ρ, the posterior distribution is $p\left(\rho \mid \mathbf{\Phi}, \mathbf{\Psi}, \mathcal{Y}\right) \propto p\left(\mathcal{Y} \mid \mathbf{\Phi}, \mathbf{\Psi}, \rho\right) p\left(\rho\right)$. We propose using slice sampling for ρ and the algorithm is summarized in Algorithm 9. The likelihood $L(\rho)$ has the same formulation as in Eq. (6.24).

6.5.3 Approximating Posterior Distribution of OD Vectors

In the Bayesian framework, the posterior distribution of an OD vector $\boldsymbol{y}^n$ conditional on observed counts and departure times is obtained by integrating out the model parameters:

$$\begin{aligned} p\left(\boldsymbol{y}^n \mid \mathcal{X}, \boldsymbol{t}\right) &= \int p\left(\boldsymbol{y}^n \mid \boldsymbol{x}^n, \Theta\right) p\left(\Theta \mid \mathcal{X}, \boldsymbol{t}\right) d\Theta \\ &\approx \frac{1}{M} \sum_{m=1}^{M} p\left(\boldsymbol{y}^n \mid \boldsymbol{x}^n, \Theta^{(m)}\right), \end{aligned} \tag{6.25}$$

where M is the number of samples used for posterior approximation, and $\Theta^{(m)}$ denotes the m-th sample in the stationary Markov chain. Therefore, the posterior distributions of OD vectors are approximated by the set of samples $\left\{\mathcal{Y}^{(m)}\right\}_{m=1}^{M}$ during the sampling process, where $\mathcal{Y}^{(m)}$ denotes the m-th sample in the Markov chain.

6.6 Experiments

Here we conduct numerical experiments using real-world data to evaluate the performance of our proposed model.

6.6.1 Data and Experiment Settings

To evaluate our approach, we use high-quality AFC data from three distinct bus routes in a city—a short route with 22 stops, a medium route with 40 stops, and a long route with 72 stops. These bus routes are in operation daily between 6:00 AM and 11:00 PM. The short route operates 103 bus runs daily with a peak frequency of 8.5 buses/hour and an off-peak frequency of 5.5 buses/hour; the medium route operates 85 bus runs daily with a peak frequency of 7.0 buses/hour and an off-peak frequency of 4.5 buses/hour; the long route operates 68 bus runs daily with a peak frequency of 6.0 buses/hour and an off-peak frequency of 4.0 buses/hour. For each bus journey, the AFC data allows us to reconstruct the true OD matrices, serving as the ground truth. We obtain boarding/alighting counts based on the true OD matrices and then apply the proposed model to infer/estimate OD matrices based on the counts. This enables us to directly evaluate the performance of our proposed model alongside other baseline methods by comparing the estimated OD matrices with the ground truth. Figure 6.3 visualizes the boarding/alighting counts at all stops over one week (from Monday to Friday). Notably, the data reveals significant fluctuations in passenger counts within a day, delineating distinct peak and off-peak hours. Furthermore, a clear daily periodicity is evident over the course of the week. Figure 6.4 presents the one-week OD vectors/flows. These passenger counts exhibit similar temporal patterns as observed in the boarding/alighting counts. Both figures demonstrate the time-varying structure of passenger demand and OD matrices, highlighting the importance of considering temporal dynamics in the estimation of OD matrices.

6.6.2 Iterative Proportional Fitting (IPF)

We first compare the performance of our model with the widely used IPF method (Ben-Akiva et al., 1985). Here, we briefly describe the IPF algorithm as follows:

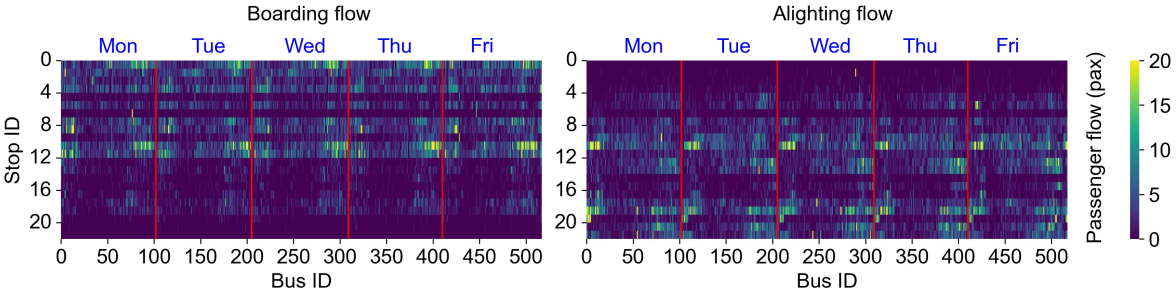


Figure 6.3: One-week boarding and alighting passenger flows of buses at stops. There are two panels: the left shows the boarding counts and the right presents the alighting counts. The x-axis represents different bus IDs and the y-axis represents the stop IDs. The color indicates the volume of passenger flows.

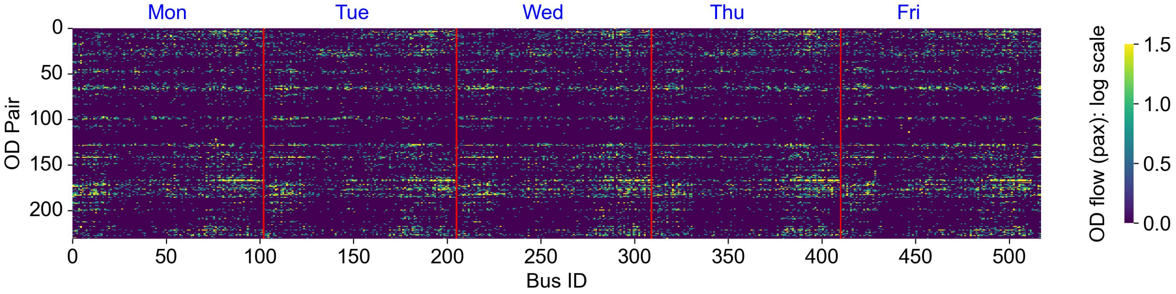


Figure 6.4: One-week OD vectors/flows of all buses. The x-axis represents different bus IDs and the y-axis represents the OD pair IDs. The color indicates the volume of flows.

- Reference/seed matrix: We divide one day into four periods, i.e., morning peak, midday off-peak, afternoon peak, and evening off-peak hours. For each period, we randomly select three true OD matrices and calculate the average OD matrix as the reference matrix. The reference matrices are initial estimates of OD matrices.

- Scaling rows and columns: The IPF algorithm iteratively scales the rows and columns of the reference matrix to match the observed boarding and alighting counts for each bus journey. Let $\boldsymbol{Y}$ be the initial reference matrix. In each iteration, the rows of $\boldsymbol{Y}$ are scaled so that their sums match the elements of $\boldsymbol{u}$, and then the columns of $\boldsymbol{Y}$ are scaled to match the elements of $\boldsymbol{v}$. This row and column scaling can be represented as follows:

$$y_{i,j}^{(\text{new})} = y_{i,j}\frac{u_i}{\sum_j y_{i,j}} \qquad (\text{row scaling})\,, \tag{6.26}$$

$$y_{i,j}^{(\text{new})} = y_{i,j}\frac{v_j}{\sum_i y_{i,j}} \qquad (\text{column scaling})\,. \tag{6.27}$$

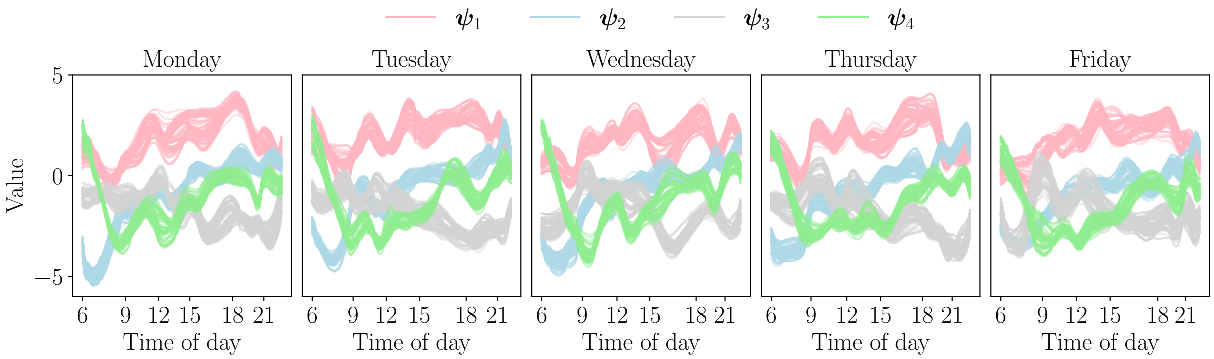


Figure 6.5: Posterior samples of $\mathbf{\Psi}$ with rank $D = 4$. We show 100 samples of $\mathbf{\Psi}$. Samples of different columns are plotted in different colors.

- Convergence: The process iterates until convergence, i.e., when the change between two iterations is below a predetermined threshold.

We apply this IPF method to estimate all OD matrices for different periods, and then compare the performance to our proposed model.

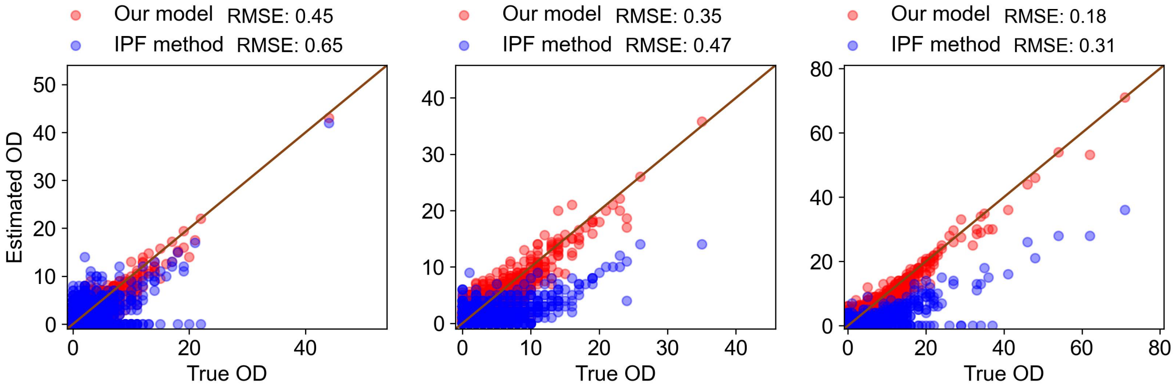
6.6.3 Estimation Results

We implement the developed MCMC algorithm and run a total of 100,000 iterations to sample the model parameters. We use the first 95,000 iterations as "burn-in" and the last $M = 5,000$ iterations to approximate the posterior distributions. As an example, Figure 6.5 shows the 100 randomly selected samples of $\mathbf{\Psi}$ with rank $D = 4$. Although we use a simple squared-exponential kernel function in the prior, the posterior samples of $\mathbf{\Psi}$ still show a clear daily periodic pattern, which confirms the consistent time-dependent characteristics of travel demand.

To demonstrate the importance of integrating temporal dynamics in OD matrix estimation, we compare the performance of the temporal Bayesian model with a non-temporal variant. The non-temporal approach assumes static parameters and is derived from our model with rank $D = 1$ and $\mathbf{\Psi} = \mathbf{1}_{N\times 1}$. This ensures that the N journeys share the same alighting probabilities. We evaluate the log-likelihood of true OD matrices given the estimated multinomial parameters. A larger log-likelihood value signifies a better model. Table 6.1 presents the log-likelihood of different models for the estimation of OD matrices. For the temporal Bayesian model, we implement four variants with different ranks (1, 2, 4 and 6). First, we compare the static model with the temporal model with $D = 1$. The key difference between these two models is the assumption of $\mathbf{\Psi}$—the static model defines $\mathbf{\Psi}$ as

Table 6.1: Log-likelihood of different models for OD matrices estimation.

		Static model	Temporal Bayesian model			
			$D = 1$	$D = 2$	$D = 4$	$D = 6$
Short route	Mean	-35328.77	-34829.07	-33728.49	-33064.69	-32874.45
	Standard deviation	98.63	86.65	79.14	86.75	85.54
Medium route	Mean	-63599.00	-62915.15	-62141.88	-61652.34	-61539.26
	Standard deviation	165.46	150.24	134.89	156.80	123.00
Long route	Mean	-47449.85	-47078.95	-46179.38	-45722.42	-45722.37
	Standard deviation	158.43	139.10	134.59	134.88	124.12



(1) Short route (2) Medium route (3) Long route

Figure 6.6: True and estimated OD flow of IPF and our proposed model for different routes. These three scatter plots show the actual and estimated OD flow of the short, medium, and long routes. Red scatters are estimated from our model and blue scatters are from the IPF model. The top texts show the RMSE of different methods.

a column vector of ones, while the temporal model treats $\boldsymbol{\Psi}$ as a random vector generated from a Gaussian process. From the results, we can see that having a temporal component (even with $D = 1$) can greatly enhance the quality of the model, confirming the importance of designing a model with time-varying parameters for OD estimation. In addtion, we can see that the log-likelihood evalutions increase with the rank for all the three bus routes. However, while more factors can enhance model performance, the improvement becomes rather marginal when $D \geqslant 4$. We use the results from models with rank $D = 4$ in the following analysis.

Figure 6.6 presents true and estimated OD flows at the journey level derived from the IPF method and our model for the three routes. Because IPF is a deterministic method, for the Bayesian method, we use the posterior mean as the estimated demand for model evaluation. The diagonal line of each plot presents the reference line of perfect estimation where estimated flows would align exactly with the true flows. The method with the dots closer to the reference line has the more accurate estimation. We can observe that the dots

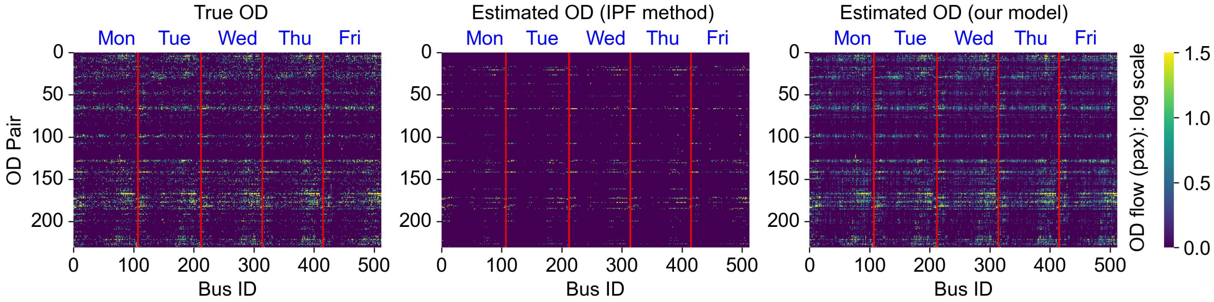


Figure 6.7: True and estimated OD vectors of all buses on the short bus route. There are three panels displaying the OD vectors/flows of all buses: the first is the true OD flows, the second is estimated with the IPF method, and the last is derived from our proposed model. The x-axis represents different bus IDs and the y-axis represents the OD pair IDs. The color represents the volume of passenger flows.

obtained from our model are closer to the reference line for all bus routes, indicating that our model outperforms the IPF method. Moreover, we use the root mean square error (RMSE) to compare the performance of our Bayesian model and the IPF method. The proposed model gives much smaller RMSE values than those obtained from IPF. Figure 6.7 shows the estimated OD vectors of all buses with the IPF method and our model on the short route. The results of the medium and long routes are shown in Appendix 6.8.1. Upon inspection, the OD vectors estimated by our model exhibit a closer resemblance to the true OD vectors compared to those generated by the IPF method. This observation underscores the superior performance of our model over the IPF method in accurately capturing and representing the OD flows.

In Figure 6.8, we visualize how the posterior mean of $\{\boldsymbol{\lambda}_1^n\}$, i.e., the vector of alighting probabilities for the first stop, varies with the sequence of journeys. As can be seen, the parameters show clear time-varying characteristics with substantial differences from morning to evening. Moreover, we can observe that while there are slight day-to-day variations in the parameters, the temporal patterns of the parameters exhibit clear similarity/periodicity across days, which is consistent with the estimate of $\boldsymbol{\Psi}$ (see Figure 6.5).

In addition to journey-level analysis, transit agencies often use aggregated OD matrices during a certain time period as a proxy for travel demand, serving as input for downstream operational tasks such as timetabling and fleet allocation. To get temporally aggregated OD matrices, we can simply aggregate those OD matrices derived from each bus journey over a defined time window. In practice, the period/window of interest by operators typically includes morning peak (commuting for work/school trips), midday off-peak, afternoon peak (commuting for home trips), and evening off-peak. Specifically, we define

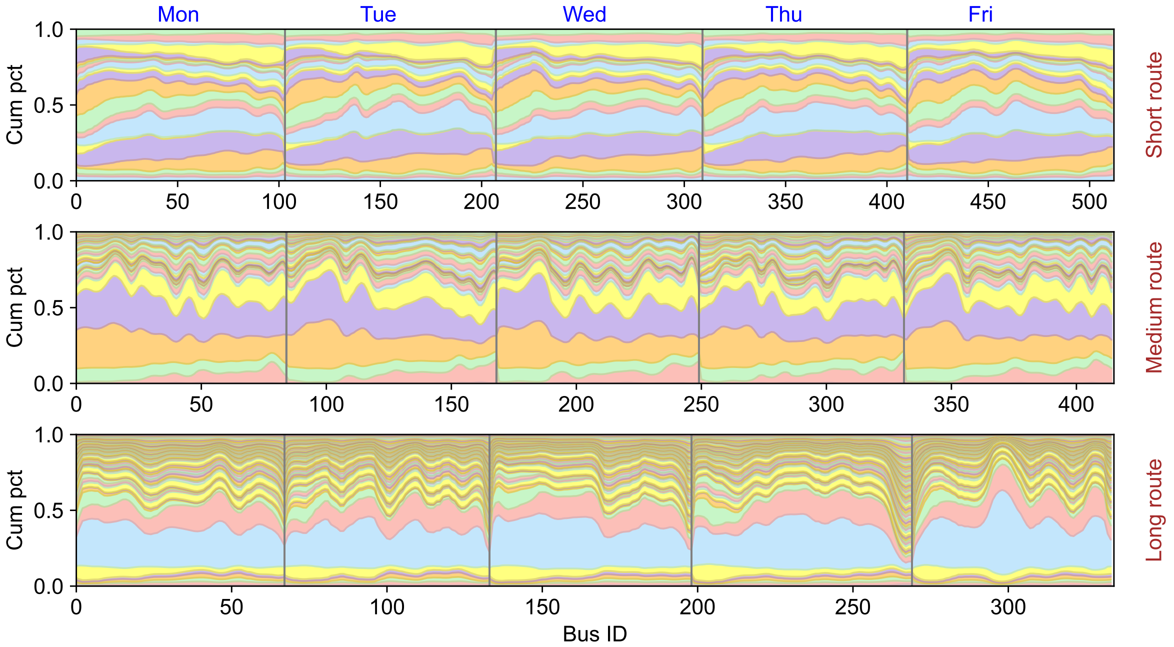


Figure 6.8: Temporal patterns of $\boldsymbol{\lambda}_1$ for different bus routes. There are three panels for different bus routes. Each panel presents the alighting probabilities of downstream stops (from 2 to S) for a passenger boarding at the initial stop. The cumulative alighting probabilities of the stops are shown in order (i.e., stop 2 to stop S) from the bottom to the top.

the periods/windows as morning peak (7:00–9:00 AM), midday off-peak (9:00 AM–5:00 PM), afternoon peak (5:00–7:00 PM), and evening off-peak (7:00–11:00 PM). We aggregate the estimated OD matrices of all buses into OD matrices of the four periods and further calculate the average hourly OD matrices of the periods to validate the performance of our model. For aggregated OD matrices, the journey-based IPF model is expected to perform poorly due to overfitting. A more appropriate benchmark is to fit a single IPF model for each period using aggregated marginal counts. We refer to this approach as "aggregated IPF". Figure 6.9 provides aggregated hourly OD matrices for different periods for all bus routes. These scatter plots present the hourly variability in performance and reveal temporal patterns in the estimation accuracy of our model. We can see that our model outperforms aggregated IPF and achieves accurate estimations of different periods for all bus routes.

A unique advantage of Bayesian inference is that we can get the posterior distributions associated with each entry in the journey-level OD matrix. Figure 6.10 presents estimations with uncertainties (95% credible interval) of all buses on the short route. The results of the medium and long routes are shown in Appendix 6.8.2. As can be seen, the presented entries indeed vary substantially over the sequence of journeys. This further supports our

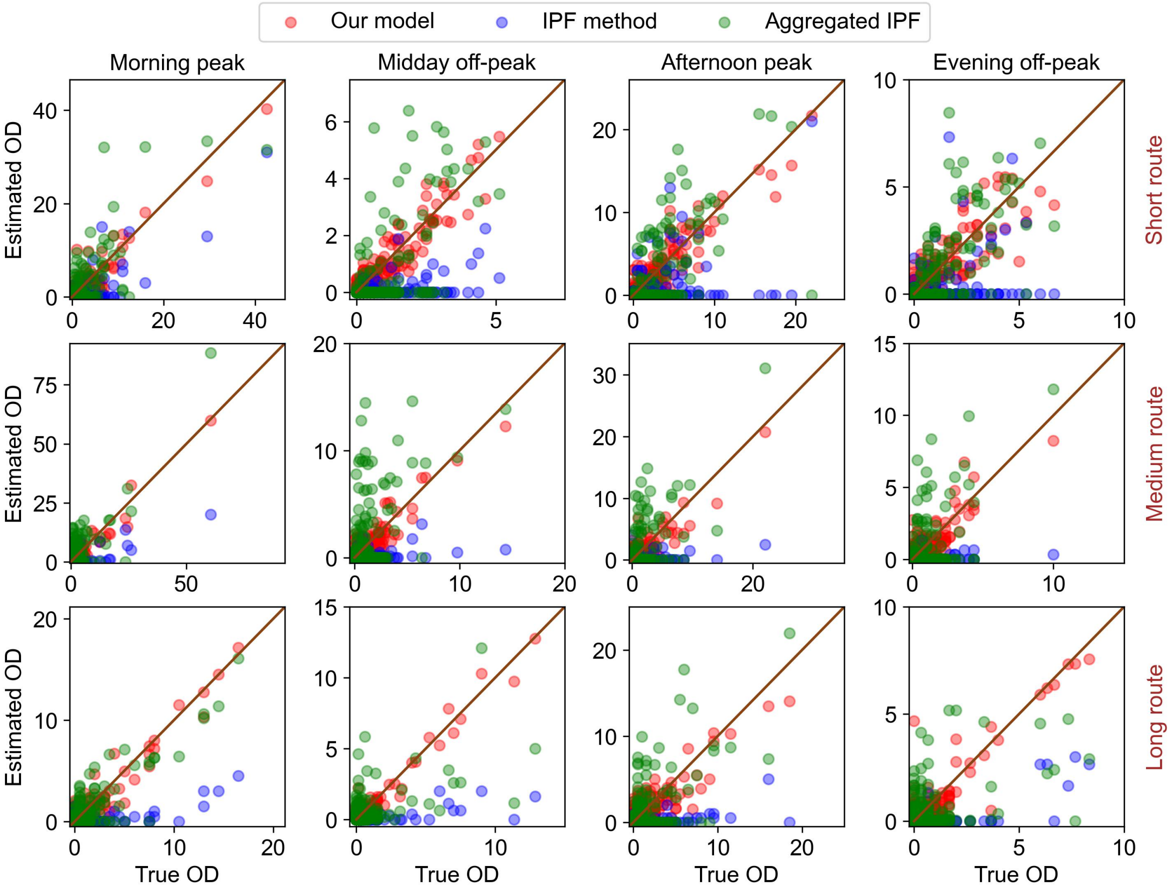


Figure 6.9: True and estimated average hourly OD flow of IPF and our model for different routes. The panels are shown in three rows (represent the short, medium, and long routes, respectively) and four columns (represent the period of morning peak, midday off-peak, afternoon peak, and evening off-peak, respectively). Each panel shows the actual and estimated average hourly OD flow of a route during a specific period. Red scatters are estimated from our proposed model, blue scatters are from the IPF method, and green scatters are from the aggregated IPF method.

choice of using the alighting probability instead of demand intensity to build the time-varying model. We can see that the performance of the estimations varies across different buses and OD pairs. In most cases, we observe good uncertainty quantification where the true values align closely with the mean estimates and fall within the 95% credible intervals. Overall, the model demonstrates robust estimation results with high-quality uncertainty quantification.

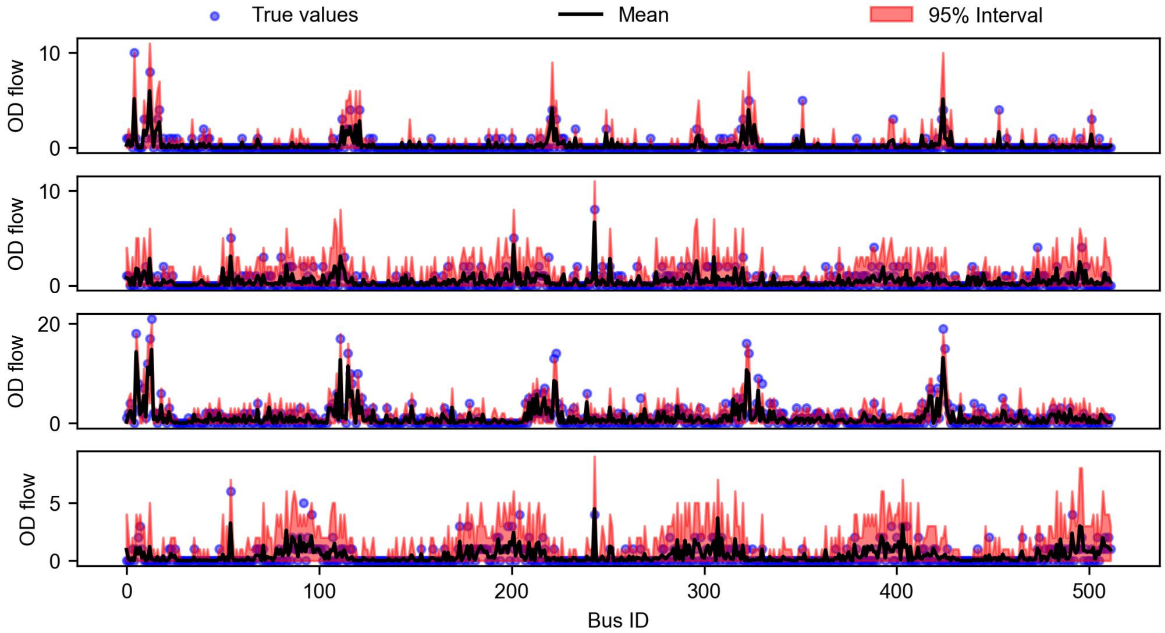


Figure 6.10: Estimation with the uncertainty of some OD pairs on the short bus route. There are four panels and each panel is for a specific OD pair. The x-axis represents the bus IDs and the y-axis represents the OD flow. The blue dots represent the true OD flows of buses, the black line represents the estimated mean of the OD flows, and the red shade areas represent the estimated 95% confidence intervals.

6.7 Conclusion

In this paper, we propose a novel temporal Bayesian model for inferring transit OD matrices at the individual bus journey level based on boarding/alighting counts at each stop. Given a boarding stop, we model the number of alighting passengers at subsequent bus stops with a multinomial distribution parameterized by a vector of alighting probabilities, and we assume that these probabilities vary smoothly over time. Given the scale of the problem, we design an efficient and effective parameterization using a matrix factorization model with a mapping factor matrix and a temporal factor matrix. In particular, we use a Gaussian process prior to model the temporal factor matrix, thus ensuring temporal smoothness in the estimated alighting probabilities. For model inference, we develop an efficient two-stage algorithm based on the MCMC method. Our approach can effectively capture the dynamic nature of OD matrices and bypass the exhaustive enumeration of feasible OD matrices which aligns with observed boarding and alighting counts. We evaluate the proposed model using real-world data, and the results confirm its effectiveness in terms of accurate OD matrix estimation and robust uncertainty quantification.

Our approach has potential implications for both practice and research. First, the proposed temporal Bayesian model can produce posterior distributions for transit OD matrices. It is important to highlight that distributions are more valuable than point estimates. This is because the associated uncertainty in travel demand distributions could benefit many downstream operational tasks, such as network design and service scheduling, where it is important to make decisions for a range of possible scenarios. Second, we find that the inferred model parameters are highly interpretable. The learned patterns could help agencies better understand how travel demand varies spatially and temporally and further improve the design of transit networks.

Our proposed model has a limitation on the assumption that multinomial probability parameters vary smoothly over time. While this assumption is generally valid for recurrent travel demand, it cannot characterize sudden changes in travel patterns resulted from special or unexpected events, where one could observed abrupt chanages in passenger demand. Therefore, in future work, we could focus on developing estimation models tailored for abnormal scenarios. Moreover, we could extend the proposed model to incorporate more prior information from additional data sources such as AFC data, which might improve the accuracy of OD matrices inference.

6.8 Appendix

6.8.1 True and estimated OD vectors of all buses.

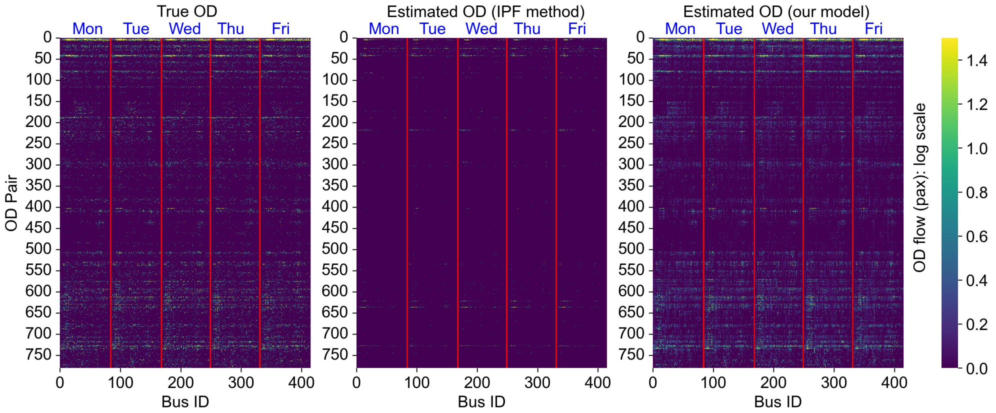


Figure 6.11: True and estimated OD vectors of all buses on the medium bus route.

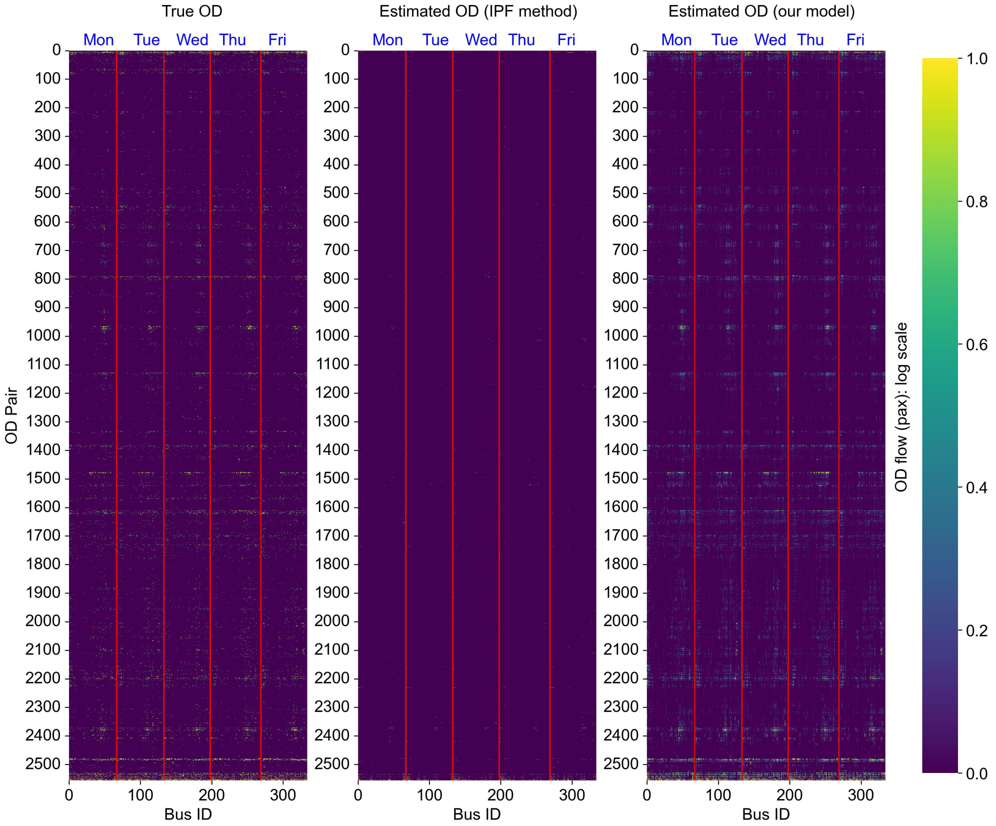


Figure 6.12: True and estimated OD vectors of all buses on the long bus route.

6.8.2 Estimation with uncertainty of some OD pairs.

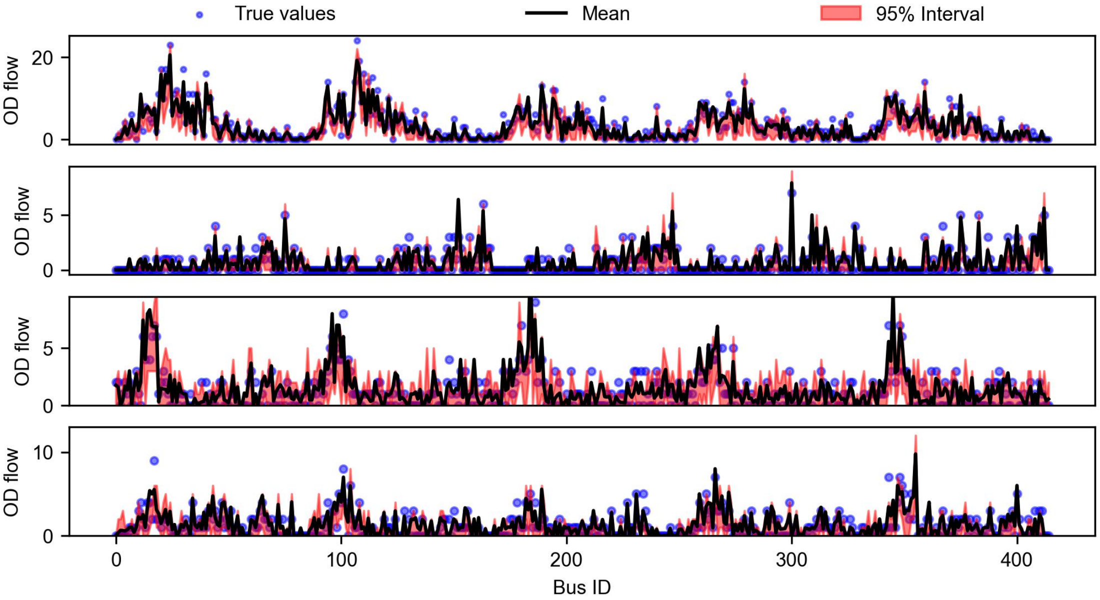


Figure 6.13: Estimation with the uncertainty of some OD pairs on the medium bus route.

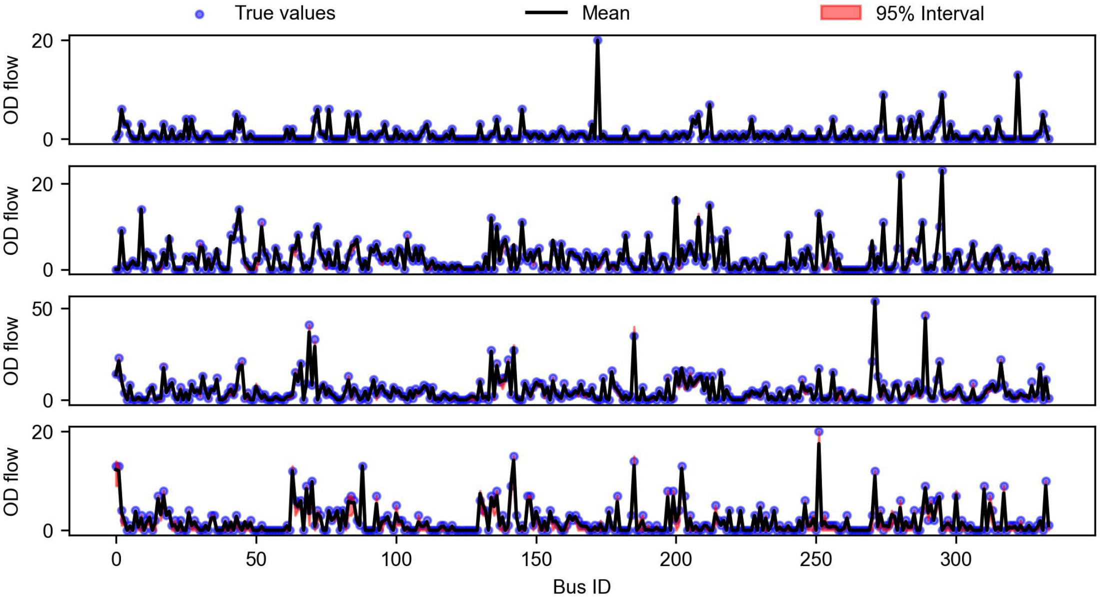


Figure 6.14: Estimation with the uncertainty of some OD pairs on the long bus route.

Chapter 7

Final Conclusion & Future Work

7.1 Summary of Results

With the increased availability of data within public transit systems, extensive research has focused on developing data-driven inference and forecasting approaches specific to these systems. However, most previous studies on transit problems have predominantly relied on deterministic models, which overlooked the uncertainty of the complex transit systems caused by stochastic factors such as traffic conditions and passenger behaviors. This thesis proposes Bayesian inference and forecasting methods tailored to address the research problems inherent to transit systems, which not only provide the point estimation but also offer the estimated probability distribution. Results demonstrate that the developed Bayesian model not only enhances forecasting accuracy but also offers robust uncertainty quantification. In the following, we summarize the results of this research in four parts:

In Chapter 3, a Bayesian probabilistic model is proposed to estimate the link travel time correlations in a bus route. The travel time of links in a bus route is assumed to follow a multivariate Gaussian distribution. This method makes use of incomplete observations with missing, ragged values and route segments from multiple bus routes. The conditional distribution of missing and ragged values can be viewed as a multivariate Gaussian distribution truncated on the intersection with a hyperplane. Next, an efficient MCMC sampling algorithm is developed to marginalize the missing and ragged values and obtain the posterior distribution of the covariance matrix. In a test with synthetic data, results show that our method produces an accurate estimation of link travel time covariance and the incorporation of incomplete data substantially improves the estimation. Moreover, the model is used to empirically quantify the link travel time correlations of a twenty-link bus route in Guangzhou, China; results reveal strong local and long-range correlation patterns in link travel time of the bus route. Finally, this chapter demonstrates an example of probabilistic forecasting of link/trip travel time in a bus route using the estimated covariance matrix; the forecasting method is more accurate than the historical average.

Chapter 4 focuses on developing a Bayesian probabilistic model for bus travel time. This model uses a new representation that combines bus link travel time and headway from a pair of adjacent buses and assumes it follows Multivariate Gaussian mixture distributions for probabilistic bus travel time forecasting. The approach naturally captures/handles the link travel time correlations of a bus route, the interactions between adjacent buses, the multimodality of bus travel time distribution, and missing values in data. Moreover, it also integrates the Gaussian mixture model with a Bayesian hierarchical framework to capture bus travel time patterns in different periods of a day. We test the proposed probabilistic forecasting model using a dataset from two bus lines in Guangzhou, China. Results show our approach that considers the dependencies between adjacent buses and the headway relationships significantly outperforms baseline models that overlook these factors, in terms of both predictive means and distributions. Besides forecasting, the parameters of the proposed model contain rich information for understanding/improving the bus service, e.g., analyzing link travel time correlation using correlation matrices and understanding temporal patterns of the bus route from mixing coefficients.

Chapter 5 continues the probabilistic forecasting of bus travel time and passenger occupancy. We develop a joint Bayesian model for bus travel time and passenger occupancy, building upon the foundation of Chapter 4. To model the correlations between travel time and passenger occupancy, we construct a variable that combines the link travel time vector, the passenger occupancy vector, and the departure headway. We employ a Bayesian Markov regime-switching vector autoregressive model to characterize the dynamic relationship among multiple buses. This new approach effectively captures essential interactions between adjacent buses, along with the multimodality and skewness of bus travel time and passenger occupancy distributions. Furthermore, it adeptly models intricate state transitions, particularly crucial when forecasting bus travel time and passenger occupancy with limited observations for the following bus. As we follow the Bayesian paradigm to estimate the parameters of the model, predictions are obtained by approximating the posterior predictive distribution. We fit the proposed model to the smart card data of one bus route in an anonymous city. The experimental results confirm that the proposed Markov regime-switching vector autoregressive model outperforms existing methods in terms of both point estimates and uncertainty quantification. This holistic approach contributes to a more robust and nuanced understanding of bus travel time and passenger occupancy dynamics, offering improved forecasting capabilities in real-world scenarios.

Chapter 6 develops a temporal Bayesian model for inferring transit OD matrices at the individual bus level. To model the discrete count data, we assume that the number of alighting passengers at subsequent bus stops, given a boarding stop, follows a multinomial distribution. To better characterize the temporal patterns in passenger demand, we assume that the parameters (i.e., assignment probabilities) of the multinomial distribution vary smoothly over time. We introduce a latent variable matrix and use it to parameterize the time-varying multinomial distributions through the softmax transformation. In addition, we propose using matrix factorization to parameterize the latent matrix as the product of a mapping factor matrix and a temporal factor matrix, which substantially reduces the number of parameters. To encode a temporally smooth structure in the matrix, we impose Gaussian process priors on the columns of the temporal factor matrix, which consequently ensure that the assignment probabilities vary smoothly over time. We evaluate our proposed model using real-world APC data and true OD matrices from three bus routes in an anonymous city. We compare the performance of the proposed temporal model to a non-temporal variant, and the results show that the temporal Bayesian model outperforms the non-temporal variant, confirming the importance and value of developing a time-varying model. In addition, we also compare our model with the widely used IPF method, and the results show that our model can achieve superior performance in deterministic estimation.

7.2 Limitations and Future Research

This thesis has explored and demonstrated the effectiveness of Bayesian statistical models for research problems in public transit systems. In this section, we aim to summarize the identified limitations and outline potential directions for future research.

- The proposed Bayesian model for inferring link travel time correlation has two limitations. First, our defined link travel time includes dwelling time. However, different bus routes have distinct characteristics of dwell time due to factors including passenger demand, bus schedule, and bus types. For example, bus routes with lower passengers flow will have shorter link travel times, while a larger passenger flow will cause longer travel times. In this case, our assumption that related bus routes share the same link travel time distribution may no longer hold. The influence of dwell time from multiple bus routes could be studied in further research. Second, the way we model the covariance structure of different time periods is by dividing samples into several periods and estimating the proposed model independently. Although simple, this approach ignores the temporal dynamic of the covariance structure—the covariance structure may vary smoothly and continuously over time. We could consider using generalized Wishart process proposed by Wilson and Ghahramani (2010) to model the time-varying covariance matrices.

- This thesis has made significant contributions in advancing the accuracy of forecasting bus travel time and passenger occupancy. By leveraging data on travel time, passenger occupancy, and headway, the models demonstrate excellent predictive performance. However, one limitation lies in its narrowed focus on internal operational factors, while external influences, such as weather conditions are not incorporated into the model. The weather has an impact on both travel times and passenger behaviors (Tao et al., 2018; Ricard et al., 2022); for example, rain, snow, and extreme temperatures can significantly alter traffic conditions and passenger demand. The exclusion of these external variables means that the model may not fully capture the complexities of real-world operations, potentially limiting its applicability in diverse environmental conditions. Future research could integrate external factors into the Bayesian forecasting models to enhance their predictive accuracy. Specifically, incorporating weather-related variables such as precipitation, temperature, and visibility could offer more comprehensive insights into their effects on bus travel times and passenger occupancy rates. Additionally, future studies could explore the inclusion of other external factors, such as special events, road constructions, and changes in urban infrastructure, which could also impact public transportation dynamics.

- The proposed Bayesian temporal model for inferring time-varying OD matrices demonstrates the ability to provide accurate estimation with uncertainty quantification. The assumption that multinomial probability parameters evolve smoothly over time is typically applicable to recurrent travel demand scenarios but falls short in capturing abrupt shifts in travel patterns that arise from extraordinary or unforeseen events, leading to sudden changes in passenger demand. Recognizing this limitation, future studies could develop estimation models specifically designed to address these abnormal scenarios. Additionally, there is potential to enhance the proposed model by integrating prior information from supplementary data sources, such as AFC data. This integration could significantly improve the accuracy of OD matrix inference.

Bibliography

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion 76, 243–297.

Achar, A., Bharathi, D., Kumar, B.A., Vanajakshi, L., 2019. Bus arrival time prediction: A spatial kalman filter approach. IEEE Transactions on Intelligent Transportation Systems 21, 1298–1307.

Achar, A., Natarajan, A., Regikumar, R., Kumar, B.A., 2022. Predicting public transit arrival: A nonlinear approach. Transportation Research Part C: Emerging Technologies 144, 103875.

Agrawal, A.W., Granger-Bevan, S., Newmark, G.L., Nixon, H., 2017. Comparing data quality and cost from three modes of on-board transit surveys. Transport Policy 54, 70–79.

Ahern, Z., Paz, A., Corry, P., 2022. Approximate multi-objective optimization for integrated bus route design and service frequency setting. Transportation Research Part B: Methodological 155, 1–25.

Alam, O., Kush, A., Emami, A., Pouladzadeh, P., 2021. Predicting irregularities in arrival times for transit buses with recurrent neural networks using gps coordinates and weather data. Journal of Ambient Intelligence and Humanized Computing 12, 7813–7826.

Assemi, B., Alsger, A., Moghaddam, M., Hickman, M., Mesbah, M., 2020. Improving alighting stop inference accuracy in the trip chaining method using neural networks. Public Transport 12, 89–121.

Bachu, A.K., Reddy, K.K., Vanajakshi, L., 2021. Bus travel time prediction using support vector machines for high variance conditions. Transport 36, 221.

Bai, C., Peng, Z.R., Lu, Q.C., Sun, J., 2015. Dynamic bus travel time prediction models on road with multiple bus routes. Computational intelligence and neuroscience 2015.

Bańbura, M., Giannone, D., Reichlin, L., 2010. Large bayesian vector auto regressions. Journal of applied Econometrics 25, 71–92.

Bao, W., Yu, Q., Kong, Y., 2020. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2682–2690.

Bapaume, T., Côme, E., Ameli, M., Roos, J., Oukhellou, L., 2023. Forecasting passenger flows and headway at train level for a public transport line: Focus on atypical situations. Transportation Research Part C: Emerging Technologies 153, 104195.

Bartholdi III, J.J., Eisenstein, D.D., 2012. A self-coördinating bus route to resist bus bunching. Transportation Research Part B: Methodological 46, 481–491.

Ben-Akiva, M., Macke, P.P., Hsu, P.S., 1985. Alternative methods to estimate route-level trip tables and expand on-board surveys. Transportation Research Record 1037, 1–11.

Bernard, M., Hackney, J.K., Axhausen, K.W., 2006. Correlation of link travel speeds: conference paper strc 2006. Working Paper/IVT 399, 1–19.

Bin, Y., Zhongzhen, Y., Baozhen, Y., 2006. Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems 10, 151–158.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer New York, NY.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, pp. 113–120.

Blume, S.O., Corman, F., Sansavini, G., 2022. Bayesian origin-destination estimation in networked transit systems using nodal in-and outflow counts. Transportation Research Part B: Methodological 161, 60–94.

Botev, Z.I., 2017. The normal law under linear restrictions: simulation and estimation via minimax tilting. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79, 125–148.

Büchel, B., Corman, F., 2018. Modelling probability distributions of public transport travel time components, in: 18th Swiss Transport Research Conference (STRC 2018), STRC.

Büchel, B., Corman, F., 2020. Review on statistical modeling of travel time variability for road-based public transport. Frontiers in Built Environment 6, 70.

Büchel, B., Corman, F., 2022a. Modeling conditional dependencies for bus travel time estimation. Physica A: Statistical Mechanics and Its Applications 592, 126764.

Büchel, B., Corman, F., 2022b. What do we know when? modeling predictability of transit operations. IEEE Transactions on Intelligent Transportation Systems 23, 15684–15695.

Carrel, A., Halvorsen, A., Walker, J.L., 2013. Passengers' perception of and behavioral adaptation to unreliability in public transportation. Transportation Research Record 2351, 153–162.

Cathey, F., Dailey, D.J., 2003. A prescription for transit arrival/departure prediction using automatic vehicle location data. Transportation Research Part C: Emerging Technologies 11, 241–264.

Cats, O., Gkioulou, Z., 2017. Modeling the impacts of public transport reliability and travel information on passengers' waiting-time uncertainty. EURO Journal on Transportation and Logistics 6, 247–270.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society Series A: Statistics in Society 158, 419–444.

Chen, B.Y., Lam, W.H., Sumalee, A., Li, Z.l., 2012. Reliable shortest path finding in stochastic networks with spatial correlated link travel times. International Journal of Geographical Information Science 26, 365–386.

Chen, M., Liu, X., Xia, J., Chien, S.I., 2004. A dynamic bus-arrival time prediction model based on apc data. Computer-Aided Civil and Infrastructure Engineering 19, 364–376.

Chen, M., Yu, G., Chen, P., Wang, Y., 2017. A copula-based approach for estimating the travel time reliability of urban arterial. Transportation Research Part C: Emerging Technologies 82, 1–23.

Chen, P., Zeng, W., Chen, M., Yu, G., Wang, Y., 2019. Modeling arterial travel time distribution by accounting for link correlations: a copula-based approach. Journal of Intelligent Transportation Systems 23, 28–40.

Chen, X., Cheng, Z., Jin, J.G., Trépanier, M., Sun, L., 2023. Probabilistic forecasting of bus travel time with a bayesian gaussian mixture model. Transportation Science .

Chen, X., Cheng, Z., Sun, L., 2022. Bayesian inference for link travel time correlation of a bus route. arXiv preprint arXiv:2202.09485 .

Cheng, Z., Trépanier, M., Sun, L., 2021. Probabilistic model for destination inference and travel pattern mining from smart card data. Transportation 48, 2035–2053.

Chepuri, A., Joshi, S., Arkatkar, S., Joshi, G., Bhaskar, A., 2020. Development of new reliability measure for bus routes using trajectory data. Transportation Letters 12, 363–374.

Cheung, R.K., 1998. Iterative methods for dynamic stochastic shortest path problems. Naval Research Logistics (NRL) 45, 769–789.

Chien, S.I.J., Ding, Y., Wei, C., 2002. Dynamic bus arrival time prediction with artificial neural networks. Journal of transportation engineering 128, 429–438.

Cong, Y., Chen, B., Zhou, M., 2017. Fast simulation of hyperplane-truncated multivariate normal distributions. Bayesian Analysis 12, 1017–1037.

Daganzo, C.F., 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. Transportation Research Part B: Methodological 43, 913–921.

Dai, Z., Ma, X., Chen, X., 2019. Bus travel time modelling using gps probe and smart card data: A probabilistic approach considering link travel time and station dwell time. Journal of Intelligent Transportation Systems 23, 175–190.

Dhivya Bharathi, B., Anil Kumar, B., Achar, A., Vanajakshi, L., 2020. Bus travel time prediction: a log-normal auto-regressive (ar) modelling approach. Transportmetrica A: Transport Science 16, 807–839.

Erhardt, G.D., Hoque, J.M., Goyal, V., Berrebi, S., Brakewood, C., Watkins, K.E., 2022. Why has public transit ridership declined in the united states? Transportation Research Part A: Policy and Practice 161, 68–87.

Fan, Y., Nie, Y., 2006. Optimal routing for maximizing the travel time reliability. Networks and Spatial Economics 6, 333–344.

Farhan, A., Shalaby, A., Sayed, T., 2002. Bus travel time prediction using avl and apc, in: Applications of Advanced Technologies in Transportation (2002), pp. 616–623.

Fox, E.B., Hughes, M.C., Sudderth, E.B., Jordan, M.I., 2014. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. The Annals of Applied Statistics 8, 1281 – 1313.

Gajewski, B.J., Rilett, L.R., 2005. Estimating link travel time correlation: an application of Bayesian smoothing splines. Journal of Transportation and Statistics 7, 53–70.

Gelfand, A.E., Smith, A.F., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American statistical association 85, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. 3rd ed., Chapman and Hall/CRC.

Gelman, A., Gilks, W.R., Roberts, G.O., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. The annals of applied probability 7, 110–120.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence , 721–741.

Geroliminis, N., Skabardonis, A., 2006. Real time vehicle reidentification and performance measures on signalized arterials, in: 2006 IEEE Intelligent Transportation Systems Conference, IEEE. pp. 188–193.

Gkiotsalitis, K., Wu, Z., Cats, O., 2019. A cost-minimization model for bus fleet allocation featuring the tactical generation of short-turning and interlining options. Transportation Research Part C: Emerging Technologies 98, 14–36.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association 102, 359–378.

Green, P.J., 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82, 711–732.

Green, P.J., Mira, A., 2001. Delayed rejection in reversible jump metropolis–hastings. Biometrika 88, 1035–1053.

Gurmu, Z.K., Fan, W.D., 2014. Artificial neural network travel time prediction model for buses using only gps data. Journal of Public Transportation 17, 3.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384.

Harsha, M., Mulangi, R.H., 2021. Probability distributions analysis of travel time variability for the public transit system. International Journal of Transportation Science and Technology .

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Hazelton, M.L., 2010. Statistical inference for transit system origin-destination matrices. Technometrics 52, 221–230.

He, P., Jiang, G., Lam, S.K., Sun, Y., 2020. Learning heterogeneous traffic patterns for travel time prediction of bus journeys. Information Sciences 512, 1394–1406.

He, P., Jiang, G., Lam, S.K., Tang, D., 2018. Travel-time prediction of bus journey with multiple bus trips. IEEE Transactions on Intelligent Transportation Systems 20, 4192–4205.

Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., et al., 2019. A case study competition among methods for analyzing large spatial data. Journal of Agricultural, Biological and Environmental Statistics 24, 398–425.

Huang, Y., Chen, C., Su, Z., Chen, T., Sumalee, A., Pan, T., Zhong, R., 2021. Bus arrival time prediction and reliability analysis: An experimental comparison of functional data analysis and bayesian support vector regression. Applied Soft Computing 111, 107663.

Hunter, T., Herring, R., Abbeel, P., Bayen, A., 2009. Path and travel time inference from gps probe vehicle data. NIPS Workshop on Analyzing Networks and Learning with Graphs 12, 2.

Hussain, E., Bhaskar, A., Chung, E., 2021. Transit od matrix estimation using smartcard data: Recent developments and future research challenges. Transportation Research Part C: Emerging Technologies 125, 103044.

Ibarra-Rojas, O.J., Delgado, F., Giesen, R., Muñoz, J.C., 2015. Planning, operation, and control of bus transport systems: A literature review. Transportation Research Part B: Methodological 77, 38–75.

Jenelius, E., 2019. Data-driven metro train crowding prediction based on real-time load data. IEEE Transactions on Intelligent Transportation Systems 21, 2254–2265.

Jenelius, E., 2020. Personalized predictive public transport crowding information with automated data sources. Transportation Research Part C: Emerging Technologies 117, 102647.

Jenelius, E., Koutsopoulos, H.N., 2013. Travel time estimation for urban road networks using low frequency probe vehicle data. Transportation Research Part B: Methodological 53, 64–81.

Jenelius, E., Koutsopoulos, H.N., 2017. Urban network travel time prediction based on a probabilistic principal component analysis model of probe data. IEEE Transactions on Intelligent Transportation Systems 19, 436–445.

Jeong, R., Rilett, R., 2004. Bus arrival time prediction using artificial neural network model, in: Proceedings. The 7th international IEEE conference on intelligent transportation systems (IEEE Cat. No. 04TH8749), IEEE. pp. 988–993.

Ji, Y., Mishalani, R.G., McCord, M.R., 2014. Estimating transit route od flow matrices from apc data on multiple bus trips using the ipf method with an iteratively improved base: Method and empirical evaluation. Journal of Transportation Engineering 140, 04014008.

Ji, Y., Mishalani, R.G., McCord, M.R., 2015. Transit passenger origin–destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. Transportation Research Part C: Emerging Technologies 58, 178–192.

Jiang, M., Zhang, Y., Zhang, Y., 2021. Optimal electric bus scheduling under travel time uncertainty: A robust model and solution method. Journal of Advanced Transportation 2021, 1–19.

Jordan, A., Krüger, F., Lerch, S., 2017. Evaluating probabilistic forecasts with scoringrules. arXiv preprint arXiv:1709.04743 .

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, 425–464.

Kieu, L.M., Bhaskar, A., Chung, E., 2015. Public transport travel-time variability definitions and monitoring. Journal of Transportation Engineering 141, 04014068.

Kim, C.J., Nelson, C.R., 2017. State-space models with REGIME Switching: Classical and Gibbs-sampling approaches with applications. MIT press.

Kononenko, I., 1989. Bayesian neural networks. Biological Cybernetics 61, 361–370.

Krolzig, H.M., 2013. Markov-switching vector autoregressions: Modelling, statistical inference, and application to business cycle analysis. volume 454. Springer Science & Business Media.

Kruschke, J.K., Liddell, T.M., 2018. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. Psychonomic bulletin & review 25, 178–206.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Kumar, B.A., Jairam, R., Arkatkar, S.S., Vanajakshi, L., 2019. Real time bus travel time prediction using k-nn classifier. Transportation Letters 11, 362–372.

Kumar, B.A., Vanajakshi, L., Subramanian, S., 2013. Pattern-based bus travel time prediction under heterogeneous traffic conditions. Transportation Research Record, Transportation Research Board, National Research Council, Washington, DC .

Kumar, B.A., Vanajakshi, L., Subramanian, S.C., 2017. Bus travel time prediction using a time-space discretization approach. Transportation Research Part C: Emerging Technologies 79, 308–332.

Kumar, B.A., Vanajakshi, L., Subramanian, S.C., 2018. A hybrid model based method for bus travel time estimation. Journal of Intelligent Transportation Systems 22, 390–406.

Lam, T.C., Small, K.A., 2001. The value of time and reliability: measurement from a value pricing experiment. Transportation Research Part E: Logistics and Transportation Review 37, 231–251.

Lam, W.H., Shao, H., Sumalee, A., 2008. Modeling impacts of adverse weather conditions on a road network with uncertainties in demand and supply. Transportation research part B: methodological 42, 890–910.

Lamond, B., Stewart, N.F., 1981. Bregman's balancing method. Transportation Research Part B: Methodological 15, 239–248.

Lei, M., Labbe, A., Wu, Y., Sun, L., 2022. Bayesian kernelized matrix factorization for spatiotemporal traffic data imputation and kriging. IEEE Transactions on Intelligent Transportation Systems 23, 18962–18974.

Li, B., 2009. Markov models for bayesian analysis about transit route origin–destination matrices. Transportation Research Part B: Methodological 43, 301–310.

Li, C., Bai, L., Liu, W., Yao, L., Waller, S.T., 2020. Graph neural network for robust public transit demand prediction. IEEE Transactions on Intelligent Transportation Systems 23, 4086–4098.

Li, C., Ling, S., Zhang, H., Zhao, H., Liu, L., Jia, N., 2023a. A sequence and network embedding method for bus arrival time prediction using gps trajectory data only. IEEE Transactions on Intelligent Transportation Systems .

Li, R., 2004. Examining travel time variability using AVI data, in: Conference of Australian Institutes of Transport Research (CAITR).

Li, X., Cottam, A., Wu, Y.J., 2023b. Transit arrival time prediction using interaction networks. IEEE Transactions on Intelligent Transportation Systems .

Li, Y., Cassidy, M.J., 2007. A generalized and efficient algorithm for estimating transit route ods from passenger counts. Transportation Research Part B: Methodological 41, 114–125.

Liang, F., 2005. Bayesian neural networks for nonlinear time series forecasting. Statistics and computing 15, 13–29.

Liao, Y., Gil, J., Pereira, R.H., Yeh, S., Verendel, V., 2020. Disparities in travel times between car and transit: Spatiotemporal patterns in cities. Scientific Reports 10, 1–12.

Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., Paninski, L., 2017. Bayesian learning and inference in recurrent switching linear dynamical systems, in: Artificial intelligence and statistics, PMLR. pp. 914–922.

Liu, T., Ma, J., Guan, W., Song, Y., Niu, H., 2012. Bus arrival time prediction based on the k-nearest neighbor method, in: 2012 Fifth International Joint Conference on Computational Sciences and Optimization, IEEE. pp. 480–483.

Liu, Y., Zhang, H., Jia, J., Shi, B., Wang, W., 2023. Understanding urban bus travel time: Statistical analysis and a deep learning prediction. International Journal of Modern Physics B 37, 2350034.

Lopes, H.F., Salazar, E., Gamerman, D., 2008. Spatial dynamic factor analysis. Bayesian Analysis 3, 759–792.

Luttinen, J., Ilin, A., 2009. Variational gaussian-process factor analysis for modeling spatio-temporal data. Advances in Neural Information Processing Systems 22.

Ma, J., Chan, J., Ristanoski, G., Rajasegarar, S., Leckie, C., 2019. Bus travel time prediction with real-time traffic information. Transportation Research Part C: Emerging Technologies 105, 536–549.

Ma, Z., Ferreira, L., Mesbah, M., Zhu, S., 2016. Modeling distributions of travel time variability for bus operations. Journal of Advanced Transportation 50, 6–24.

Ma, Z., Koutsopoulos, H.N., Ferreira, L., Mesbah, M., 2017. Estimation of trip travel time distribution using a generalized markov chain approach. Transportation Research Part C: Emerging Technologies 74, 1–21.

Madzlan, N., Ibrahim, K., et al., 2010. Arima models for bus travel time prediction .

Martínez, H., Mauttone, A., Urquhart, M.E., 2014. Frequency optimization in public transportation systems: Formulation and metaheuristic approach. European Journal of Operational Research 236, 27–36.

May, A., Bonsall, P.W., Marler, N., 1989. Travel time variability of a group of car commuters in north london .

Mazloumi, E., Currie, G., Rose, G., 2010. Using gps data to gain insight into public transport travel time variability. Journal of transportation engineering 136, 623–631.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21, 1087–1092.

Metropolis, N., Ulam, S., 1949. The monte carlo method. Journal of the American statistical association 44, 335–341.

Miller-Hooks, E., 2001. Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. Networks: An International Journal 37, 35–52.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. Journal of the american statistical association 83, 1023–1032.

Mohammed, M., Oke, J., 2023. Origin-destination inference in public transportation systems: A comprehensive review. International Journal of Transportation Science and Technology 12, 315–328.

Murray, I., Adams, R., MacKay, D., 2010. Elliptical slice sampling, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings. pp. 541–548.

Murray, I., Adams, R.P., 2010. Slice sampling covariance hyperparameters of latent gaussian models. Advances in neural information processing systems 23.

Neal, R.M., 2003. Slice sampling. The annals of statistics 31, 705–767.

Osman, O., Rakha, H., Mittal, A., 2021. Application of long short term memory networks for long-and short-term bus travel time prediction .

Pasini, K., Khouadjia, M., Same, A., Ganansia, F., Oukhellou, L., 2019. Lstm encoder-predictor for short-term train load forecasting, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 535–551.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19, 557–568.

Petersen, N.C., Rodrigues, F., Pereira, F.C., 2019. Multi-output bus travel time prediction with convolutional lstm neural network. Expert Systems with Applications 120, 426–435.

Qin, W., Ji, X., Liang, F., 2020. Estimation of urban arterial travel time distribution considering link correlations. Transportmetrica A: Transport Science 16, 1429–1458.

Rachtan, P., Huang, H., Gao, S., 2013. Spatiotemporal link speed correlations: Empirical study. Transportation Research Record 2390, 34–43.

Rahbar, M., Hickman, M., Mesbah, M., Tavassoli, A., 2018. Calibrating a bayesian transit assignment model using smart card data. IEEE Transactions on Intelligent Transportation Systems 20, 1574–1583.

Rahman, M.M., Wirasinghe, S., Kattan, L., 2018. Analysis of bus travel time distributions for varying horizons and real-time applications. Transportation Research Part C: Emerging Technologies 86, 453–466.

Rakha, H.A., El-Shawarby, I., Arafeh, M., Dion, F., 2006. Estimating path travel-time reliability, in: 2006 IEEE Intelligent Transportation Systems Conference, IEEE. pp. 236–241.

Ramezani, M., Geroliminis, N., 2012. On the estimation of arterial route travel time distribution with markov chains. Transportation Research Part B: Methodological 46, 1576–1590.

Ricard, L., Desaulniers, G., Lodi, A., Rousseau, L.M., 2022. Predicting the probability distribution of bus travel time to measure the reliability of public transport services. Transportation Research Part C: Emerging Technologies 138, 103619.

Richardson, S., Green, P.J., 1997. On bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society Series B: Statistical Methodology 59, 731–792.

Rodriguez, Deniz, H., Jenelius, E., Villani, M., 2017. Urban network travel time prediction via online multi-output gaussian process regression, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 1–6.

Rodriguez-Deniz, H., Villani, M., 2022. Robust real-time delay predictions in a network of high-frequency urban buses. IEEE Transactions on Intelligent Transportation Systems .

Roos, J., Bonnevay, S., Gavin, G., 2017. Dynamic bayesian networks with gaussian mixture models for short-term passenger flow forecasting, in: 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), IEEE. pp. 1–8.

Schofer, J.L., Khattak, A., Koppelman, F.S., 1993. Behavioral issues in the design and evaluation of advanced traveler information systems. Transportation Research Part C: Emerging Technologies 1, 107–117.

Scott, S.L., 2002. Bayesian methods for hidden markov models: Recursive computing in the 21st century. Journal of the American statistical Association 97, 337–351.

Sen, A., Thakuriah, P., Zhu, X.Q., Karr, A., 1999. Variances of link travel time estimates: implications for optimal routes. International Transactions in Operational Research 6, 75–87.

Seshadri, R., Srinivasan, K.K., 2010. Algorithm for determining most reliable travel time path on network with normally distributed and correlated link travel times. Transportation Research Record 2196, 83–92.

Smeed, R., Jeffcoate, G., 1971. The variability of car journey times on a particular route. Traffic Engineering and Control 13, 238–243.

Soize, C., 2017. Uncertainty quantification. Springer.

Srinivasan, K.K., Prakash, A., Seshadri, R., 2014. Finding most reliable paths on networks with correlated and shifted log–normal travel times. Transportation Research Part B: Methodological 66, 110–128.

Sullivan, T.J., 2015. Introduction to uncertainty quantification. Springer.

Sun, L., Jin, J.G., Lee, D.H., Axhausen, K.W., Erath, A., 2014a. Demand-driven timetable design for metro services. Transportation Research Part C: Emerging Technologies 46, 284–299.

Sun, L., Lu, Y., Jin, J.G., Lee, D.H., Axhausen, K.W., 2015. An integrated bayesian approach for passenger flow assignment in metro networks. Transportation Research Part C: Emerging Technologies 52, 116–131.

Sun, L., Tirachini, A., Axhausen, K.W., Erath, A., Lee, D.H., 2014b. Models of bus boarding and alighting dynamics. Transportation Research Part A: Policy and Practice 69, 447–460.

Sun, W., Schmöcker, J.D., Fukuda, K., 2021. Estimating the route-level passenger demand profile from bus dwell times. Transportation Research Part C: Emerging Technologies 130, 103273.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American statistical Association 82, 528–540.

Tao, S., Corcoran, J., Rowe, F., Hickman, M., 2018. To travel or not to travel:'weather'is the question. modelling the effect of local weather conditions on bus ridership. Transportation research part C: emerging technologies 86, 147–167.

Taylor, M., 1982. Travel time variability—the case of two public modes. Transportation Science 16, 507–521.

Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems 11, 1–14.

Triantafyllopoulos, K., et al., 2021. Bayesian inference of state space models. Springer.

United Nations, 2015. Transforming our world: the 2030 Agenda for Sustainable Development. UN Doc. A/RES/70/1 .

Uno, N., Kurauchi, F., Tamura, H., Iida, Y., 2009. Using bus probe data for analysis of travel time variability. Journal of Intelligent Transportation Systems 13, 2–15.

Vardi, Y., 1996. Network tomography: Estimating source-destination traffic intensities from link data. Journal of the American Statistical Association 91, 365–377.

Waller, S.T., Ziliaskopoulos, A.K., 2002. On the online shortest path problem with limited arc cost dependencies. Networks: An International Journal 40, 216–227.

Wang, P., Chen, X., Chen, J., Hua, M., Pu, Z., 2021a. A two-stage method for bus passenger load prediction using automatic passenger counting data. IET Intelligent Transport Systems 15, 248–260.

Wang, P., Chen, X., Zheng, Y., Cheng, L., Wang, Y., Lei, D., 2021b. Providing real-time bus crowding information for passengers: a novel policy to promote high-frequency transit performance. Transportation Research Part A: Policy and Practice 148, 316–329.

Westgate, B.S., Woodard, D.B., Matteson, D.S., Henderson, S.G., 2016. Large-network travel time distribution estimation for ambulances. European Journal of Operational Research 252, 322–333.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian processes for machine learning. MIT press Cambridge, MA.

Wilson, A.G., Ghahramani, Z., 2010. Generalised wishart processes. arXiv preprint arXiv:1101.0240 .

Wood, J., Yu, Z., Gayah, V.V., 2023. Development and evaluation of frameworks for real-time bus passenger occupancy prediction. International Journal of Transportation Science and Technology 12, 399–413.

Woodard, D., Nogin, G., Koch, P., Racz, D., Goldszmidt, M., Horvitz, E., 2017. Predicting travel time reliability using mobile phone gps data. Transportation Research Part C: Emerging Technologies 75, 30–44.

Xu, H., Ying, J., 2017. Bus arrival time prediction with real-time and historic data. Cluster Computing 20, 3099–3106.

Xuan, Y., Argote, J., Daganzo, C.F., 2011. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. Transportation Research Part B: Methodological 45, 1831–1845.

Yang, M., Chen, C., Wang, L., Yan, X., Zhou, L., 2016. Bus arrival time prediction using support vector machine with genetic algorithm. Neural Network World 26, 205.

Yetiskul, E., Senbil, M., 2012. Public bus transit travel-time variability in ankara (turkey). Transport Policy 23, 50–59.

Yu, B., Lam, W.H., Tam, M.L., 2011. Bus arrival time prediction at bus stop with multiple routes. Transportation Research Part C: Emerging Technologies 19, 1157–1170.

Yu, B., Wang, H., Shan, W., Yao, B., 2018. Prediction of bus travel time using random forests based on near neighbors. Computer-Aided Civil and Infrastructure Engineering 33, 333–350.

Yu, Z., Wood, J.S., Gayah, V.V., 2017. Using survival models to estimate bus travel times and associated uncertainties. Transportation Research Part C: Emerging Technologies 74, 366–382.

Zeng, W., Miwa, T., Wakita, Y., Morikawa, T., 2015. Application of lagrangian relaxation approach to α-reliable path finding in stochastic networks with correlated link travel times. Transportation Research Part C: Emerging Technologies 56, 309–334.

Zhan, X., Hasan, S., Ukkusuri, S.V., Kamga, C., 2013. Urban link travel time estimation using large-scale taxi data with partial information. Transportation Research Part C: Emerging Technologies 33, 37–49.

Zhan, X., Ukkusuri, S.V., Yang, C., 2016. A bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data. Automation in construction 72, 237–246.

Zhang, X., Yan, M., Xie, B., Yang, H., Ma, H., 2021. An automatic real-time bus schedule redesign method based on bus arrival time prediction. Advanced Engineering Informatics 48, 101295.

Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2018. Individual mobility prediction using transit smart card data. Transportation research part C: emerging technologies 89, 19–34.

Zheng, Y., Zhang, Y., Li, L., 2016. Reliable path planning for bus networks considering travel time uncertainty. IEEE Intelligent Transportation Systems Magazine 8, 35–50.

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2018. Inferring left behind passengers in congested metro systems from automated data. Transportation Research Part C: Emerging Technologies 94, 323–337.