Spring 2021

Essays in Applied Bayesian Analysis

Xinyuan Chen
Yale University Graduate School of Arts and Sciences, xinyuan.eric.chen@gmail.com

Abstract

Essays in Applied Bayesian Analysis

Xinyuan Chen
2021

With continuing rapid developments in computational power, Bayesian statistical methods, because of their user-friendliness and estimation capabilities, have become increasingly popular in a considerable variety of application fields. In this thesis, applied Bayesian methodological topics and empirical examples focusing on nonhomogeneous hidden Markov models (NHMMs) and measurement error models are explored in three chapters. In the first chapter, a subsequence-based variational Bayesian inference framework for NHMMs is proposed in order to address the computational problems encountered when analyzing datasets containing long sequences. The second chapter concentrates on measurement error models, where a Bayesian estimation procedure is proposed for the partial potential impact fraction (pPIF) with the presence of measurement error. The third chapter focuses on an empirical application in marketing, where a coupled nonhomogeneous hidden Markov model (CNHMM) is introduced to provide a novel framework for customer relationship management.

Essays in Applied Bayesian Analysis

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Xinyuan Chen

Dissertation Director: Dr. Joseph Chang

June 2021

Acknowledgments

I would like to thank my advisor Dr. Joseph Chang for his genuine friendship, consistent patience, and unconditional support, which made it possible for me to pursue my academic goals.

I would also like to thank Dr. Andrew Barron and Dr. John Emerson for their timely guidance and encouragement throughout my time at Yale.

I would never be able to make this far without the mentorship and expertise from Dr. Yiwei Li, Dr. Xiangnan Feng, Dr. Fan Li, Dr. Donna Spiegelman, Dr. Katarzyna Chawarska, Dr. Sahand Negahban, and Dr. Jason Lam. Special thanks to all of you for your valuable inputs and advice.

Above all, I want to thank my parents, Weixuan Kang and Yuchun Chen, for their continual love and support, and always believing that I can achieve anything I set my mind to.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With continuing rapid developments in computational power, Bayesian statistical methods, because of their user-friendliness and estimation capabilities, have become increasingly popular in a considerable variety of application fields. In this dissertation, I investigate some methodological topics and empirical applications of Bayesian analysis, emphasizing nonhomogeneous hidden Markov models and measurement error models.

This dissertation includes three chapters. The first chapter, **Variational Bayesian Analysis of Nonhomogeneous Hidden Markov Models with Long Sequences**, focuses on addressing computational issues encountered when analyzing datasets containing long sequences using nonhomogeneous hidden Markov models (NHMMs). This study is motivated by a collaboration with the Yale Child Study Center, where the main interest lies in modeling eye-tracking scan-paths of autistic children to examine the comparative social salience of puppets and people in designed social-communicative scenes, and a joint work with a large telecommunication carrier in China, where the goal is to model ultra-long sequences of customers' telecom records to uncover the relationship between their mobile Internet use behavior and conventional telecommunication (phone calls and SMS) behaviors.

Conventional Bayesian approaches, particularly Markov chain Monte Carlo (MCMC) methods, are computationally demanding especially for long observation sequences. I thus

develop a variational Bayes (VB) method for NHMMs, which utilizes a structured variational family of Gaussian distributions with factorized covariance matrices to approximate target posteriors, combining forward-backward algorithm and stochastic gradient ascent in estimation. To improve efficiency and handle ultra-long sequences, I further propose a subsequence VB (SVB) method that works on subsamples (short sequences sampled from from some of the series). The SVB method exploits a memory decay property of NHMMs and uses buffers to control for bias caused by breaking sequential dependence from subsampling. The chapter highlights that local nonhomogeneity of NHMMs substantially affects required buffer lengths and proposes the use of local Lyapunov exponents to help characterize local memory decay rates of NHMMs and determine buffer lengths adaptively.

The second chapter, **A Bayesian approach for estimating the partial potential impact fraction with exposure measurement error under a main study/internal validation design**, discusses the estimation of the partial potential impact fraction (pPIF) in the presence of measurement error. The pPIF is often used to describe the proportion of disease cases that can be prevented if the distribution of a modifiable continuous exposure is shifted in a population, while other risk factors are not modified. It is a useful quantity for evaluating the burden of disease in epidemiologic and public health studies. When exposures are measured with error, standard pPIF estimates may be biased, which necessitates methods to correct for the exposure measurement error. Motivated by the Health Professionals Follow-up Study (HPFS), I propose a Bayesian approach to adjust for exposure measurement error when estimating the pPIF under the main study/internal validation study design. I adopt a *reclassification approach* that leverages the strength of the main study/internal validation study design, and clarify *transportability* assumptions, which relate certain distributions from the main study and validation study, for valid inference. I assess the finite-sample performance of both the point and credible interval estimators via extensive simulations, and apply the proposed approach in the HPFS to estimate the pPIF for colorectal cancer (CRC) incidence under interventions exploring shifting the

distributions of red meat, alcohol, and/or folate intake.

The third chapter, **When customer dynamics is more than relationship: A coupled hidden Markov model framework**, concentrates on an empirical application of the coupled nonhomogeneous hidden Markov model (CNHMM) in marketing to study customer dynamics and implement customer relationship management (CRM). In this research, I propose a CNHMM-based framework that simultaneously considers two correlated Markov processes, respectively representing the latent relational and monetary value of customers. Leveraging data from a major telecommunication carrier in China, the findings indicate that the proposed method is able to uncover the two-dimensional latent states of customers (dynamic customer values) and possible effects of covariates of interest (including marketing mixes) on the evolutions of the latent states. Consumers' choice of products (and services) are jointly influenced by their relational and monetary value over time, and the evolution of customers' relational states is significantly dependent on their monetary states (but not vice versa), suggesting customer heterogeneity in monetary value is a potential antecedent of customer-firm relationship. Furthermore, scenario analyses are conducted to showcase how the proposed model can help firms formulate effective multidimensional dynamic segmentation strategies for customer relationship management.

Chapter 2

Variational Bayesian Analysis of Nonhomogeneous Hidden Markov Models with Long Sequences[1]

Abstract

Nonhomogeneous hidden Markov models (NHMMs) are useful in modeling sequential and autocorrelated data. Bayesian approaches, particularly Markov chain Monte Carlo (MCMC) methods, are principal statistical inference tools for NHMMs. However, MCMC sampling is computationally demanding especially for long observation sequences. We develop a variational Bayes (VB) method for NHMMs, which utilizes a structured variational family of Gaussian distributions with factorized covariance matrices to approximate target posteriors, combining forward-backward algorithm and stochastic gradient ascent in estimation. To improve efficiency and handle ultra-long sequences, we further propose a subsequence VB (SVB) method that works on subsamples. The SVB method exploits the memory decay property of NHMMs and uses buffers to control for bias caused by breaking sequential dependence from subsampling. We highlight that local nonhomogeneity of NHMMs substantially affects required buffer lengths and propose the use of local Lya-

[1] Co-authored with Yiwei Li, Xiangnan Feng, Fred Volkmar, Katarzyna Chawarska, and Joseph Chang

punov exponents which characterizes local memory decay rates of NHMMs and determines buffer lengths adaptively. Our methods are applied in modeling eye-tracking scan-paths of autistic children to examine the comparative social salience of humanoid representation and person in designed social-communicative scenes and in modeling ultra-long sequences of customers' telecom records to uncover the relationship between their mobile Internet use behavior and conventional telecommunication behaviors.

2.1 Introduction

Hidden Markov models (HMMs) are a class of discrete-time finite state-space models, which are particularly suitable for modeling sequentially correlated observations via the evolution of a set of hidden states. Because of their structural flexibility in uncovering the dynamic transitions between hidden states and interpretation power of summarizing complex behavioral patterns implied from the observation sequences, HMMs and their variants have been applied widely (Cappé et al., 2005). HMMs model observed sequences through two components: a latent Markov chain governing the sequential evolution of hidden states, and an emission model generating state-specific observations. Most applications have focused on time-homogeneous HMMs, where transition probability matrices are time-invariant. This homogeneous setting ignores the impacts of time-varying influential factors on the transition probabilities and is over-simplified in certain applications. As an extension, nonhomogeneous HMMs (NHMMs), which relax the homogeneous assumption and model the temporal changes in the hidden process of HMMs, have been developed and applied (Heaps et al., 2015; Holsclaw et al., 2017; Netzer et al., 2008; Hughes et al., 1999).

Bayesian approaches, particularly Markov chain Monte Carlo (MCMC) methods, are widely adopted for conducting inference on NHMMs (e.g., Heaps et al., 2015; Ascarza et al., 2018; Montoya et al., 2010). The Bayesian paradigm is appealing for estimating complex models such as NHMMs because it provides reliable results through incorporating

valuable prior information and facilitates statistical inference by generating entire posterior distributions for unknown quantities. However, Bayesian approaches for NHMMs are computationally demanding, especially when the observation sequences are long. Specifically, parameter inference for the transition model is complicated, not only because using the logistic function jeopardizes the conjugacy of posterior distributions, but also because the estimation of unobservable hidden states depends on the forward-backward algorithm which requires iterative updates across entire sequences. For instance, the length of each observed sequence is 2.6×10^3 and 8.7×10^4 in our first and second applications respectively, which is overly long for conventional methods to handle efficiently.

The development of an efficient Bayesian method for NHMMs that handles long observation sequences is challenging. To our knowledge, Holsclaw et al. (2017)'s work is the only study that attempts to propose an efficient Bayesian method for an NHMM with long observation sequences. They focus on the lack of conjugacy problem that arises from the presence of the logistic function in the transition model and adopt the Pólya-Gamma data augmentation method (Polson et al., 2013) as a remedy to reduce the mixing time and hyper-parameter tuning complexity of the MCMC approach. In their example, the proposed method is able to handle an NHMM with long sequences of a length up to 10^4, where the NHMM contains a transition model of a reduced form tailored for the application.

In this paper, we consider an alternative Bayesian approach for general NHMMs with long sequences, Variational Bayes (VB), which demonstrates appealing computational efficiency and accuracy and features satisfactory scalability to large-scale data (Blei et al., 2017). The aim of MCMC and VB methods is to approximate complex posterior distributions; MCMC methods achieve this goal by generating samples from a Markov chain that converges to the target posteriors, whereas VB methods use variational posteriors as the approximation, obtained by minimizing a distance between the variational and true posteriors through optimization. Though not enjoying theoretical guarantees of producing samples from the true target posteriors, VB methods typically provide sufficiently close

approximations with a significant speedup (Braun and McAuliffe, 2010; Blei et al., 2017). To our understanding, no VB method has been proposed for general NHMMs. Hence, we contribute to the literature by developing an efficient VB method for NHMMs capable of handling datasets with long sequences. A structured mean-field variational family that allows for dependence among variational posteriors is used to preserve the structural dependence implied by the model and thereby better approximate the true posteriors (Ong et al., 2018). We utilize stochastic gradient ascent (SGA, Robbins and Monro, 1951) as the optimization approach which directly works on the gradient of the objective function and does not rely on the conjugacy of posteriors, which alleviates the lack of conjugacy problem in the transition model. The proposed VB method for NHMMs is capable of handling moderately long sequences ($T \leq 10^4$). However, neither the MCMC method in Holsclaw et al. (2017) nor the proposed VB method can efficiently process datasets with ultra-long sequences ($T > 10^4$).

To reduce the computational complexity of the proposed VB method in handling ultra-long sequences, we propose a subsequence VB (SVB) method that works on subsamples. Subsampling methods, especially stochastic gradient methods (Hoffman et al., 2013; Nemirovski et al., 2009) which employ noisy estimates of the gradient using minibatches of the data, have been proposed to avoid costly gradient computation using the full dataset. Marked progress has been made by adapting subsampling methods in various models to analyze massive datasets (e.g., Ansari et al., 2018; Blei et al., 2017; Gentzkow et al., 2019). However, these subsampling methods are mainly developed for independent or exchangeable data and not directly applicable for sequential and correlated data as modeled by HMMs, because simply sampling several subsequences may break crucial dependence among data points and lead to significant bias (Aicher et al., 2019). Recently, subsequence methods exploiting the memory decay property of homogeneous HMMs have been proposed to control for the bias with the aid of buffers adjacent to the sampled subsequences (Foti et al., 2014; Aicher et al., 2019; Ma et al., 2017; Ye, 2018). Our proposed SVB method

extends this methodology and uses buffers to reduce bias caused by subsampling. Our work highlights a new aspect of NHMM analysis, in that, in comparison to homogeneous HMMs, NHMMs require more careful consideration of local nonhomogeneity when developing the subsampling method. Specifically, local nonhomogeneity of an NHMM significantly affects its local memory decay rates, and thus, the required buffer lengths. We thus propose the use of *Local Lyapunov Exponents* (LLEs, Abarbanel et al., 1992), a measure characterizing local memory decay rates of NHMMs, to estimate the buffer lengths. With LLEs, our subsampling method adaptively determines the buffer lengths for each subsequence and is demonstrated to be efficient and effective in processing ultra-long sequences.

The methods for analyzing NHMMs with long and ultra-long sequences are motivated by two real data problems. The first motivating example is the problem of modeling eye-tracking scan-path data obtained from children with autism spectrum disorders (ASD). ASD refers to complex developmental brain disorders of early onset marked by a profound social dysfunction affecting an individual's social interaction and behavioral pattern (American Psychiatric Association, 2013). One critical deficit among children with ASD is diminished attention to social cues from others, such as face directions, eye gaze, gestures, and language, which impacts the development of critical social-cognitive skills such as joint attention, social play, language development, and theory of mind (Chawarska et al., 2012; Klin et al., 2002). Therefore, increasing children's social attention is a major target of ASD interventions. Recent research on ASD interventions suggests that humanoid representations (e.g., robots and puppets) may be more effective in teaching ASD children certain social skills than human-delivered training (Kim et al., 2013; Scassellati et al., 2018). Although the findings generate considerable excitement, the underlying mechanisms of the advantageous performance of humanoid representations remain uncertain. For instance, it is unclear whether humanoid representations are particularly suitable to teach certain social skills, such as joint attention, are generally more salient to children with ASD than a human, especially with the presence of higher saliency cues (HSCs) such as speech. In

this motivating example, we focus on the latter potential mechanism and use an NHMM to model eye-tracking scan-paths recorded from children with ASD watching a video that features a puppet and a person engaging in a conversation, so as to examine the comparative social salience of the puppet and the person in the setting of designed social-communicative scenes.

The data of this example came from a research program on the social and emotional development of children with ASD conducted by the Child Study Center of the Yale School of Medicine. Children were eye-tracked during the experiment to obtain precise and dense measurements of their eye movements in the form of sequences of gaze point coordinates scan-paths. Previous methods for analyzing eye-tracking scan-path data in autism research mainly lie in two directions. The first direction focuses on spatial information of the scan-paths by analyzing fractions of time participants looked at predefined regions of interest (ROIs, Wang et al., 2019; Shic et al., 2019) or constructing a cohesion metric to quantify children's gaze behaviors (Wang et al., 2018). This direction loses potentially valuable temporal information of the scan-paths. The second direction regards the scan-paths as time-series data and applies sequential models such as HMMs to summarize the temporal gaze patterns or to classify gaze points into data-driven categories (Alie et al., 2011; Mavadati et al., 2014). This latter direction may provide data-driven ROIs and investigate both spatial and temporal aspects of gaze behaviors. However, the latter approach generally use homogeneous HMMs, which cannot capture the effects of time-varying influential factors (e.g., HSCs in our study) on dynamic transitions between hidden states (e.g., attention shifts across ROIs). We therefore propose an NHMM for analyzing eye-tracking scan-paths, providing data-driven ROIs on which children's attention focuses and describing children's attention shifts across ROIs. In addition, the NHMM captures the effects of time-varying HSCs on children's attention shifts. The proposed framework presents an appealing alternative for analyzing eye-tracking data in autism research, which, however, requires an efficient computational approach to deal with the methodological challenge

caused by the long sequences ($T = 2.6 \times 10^3$ in this application).

The second motivating example arises from the need for better capturing and understanding the mobile Internet use patterns of customers in the telecom service industry. Recent technological innovations of mobile Internet are reshaping functions of modern smartphones. The evolution of mobile functions has started to shift the telecom industry's core business from conventional telecom services to mobile Internet-based services (Andrews et al., 2016; Fong et al., 2015; Luo et al., 2014). Consequently, there has been much debate among managers and even lawmakers on whether conventional telecom services are competitive substitutes or useful complements for mobile Internet services. While some managers think that mobile Internet messaging apps are taking the place of phone calls and SMS and therefore hurt company profits from metered consumption and billing, some other decision makers believe that the relationship between conventional telecom services and mobile Internet services is complementary and helps the company generate more profits through better catering to customers' needs for communication and social connectivity. These competing viewpoints motivate us to examine the relationship between conventional telecom services and mobile Internet services through modeling the dynamics of customers' mobile Internet use behaviors. Our results may guide companies in adjusting their business strategies. The analysis was conducted based on densely recorded individual-level data from a major telecom service provider in China, which contains information (calls, texts/short message service [SMS], and mobile Internet usage) at a frequency of every five minutes for 10 months from September 2013 to June 2014 for a group of customers. We set out to utilize an NHMM to model customers' dynamic mobile Internet use behaviors, uncover their latent needs for mobile Internet through analysis of hidden states, assess the influence of their conventional telecom behaviors (i.e., calls and SMS) on the latent needs, and forecast their future mobile Internet usage.

This analysis not only benefits decision makers by enhancing their understanding of the relationship between conventional telecom services (phone call and SMS) and mobile

Internet services, but also provides companies with a useful tool to analyze valuable real-time customer information on mobile Internet usage. In particular, our modeling effort is performed on the individual level, delineating patterns of customers' mobile Internet use and assisting companies in timely monitoring and forecasting customers' latent needs (and thus, usage) of mobile Internet. This application showcases the potential of NHMMs in customer relationship management (CRM). CRM, defined as the process of managing customers' information to create value for customers and maximize their loyalty, is the outcome of the evolution and integration of new data, technologies, and strategies (Boulding et al., 2005; Kotler and Keller, 2016). NHMMs have been recognized in the literature as an effective tool for analyzing customer behavior data, uncovering meaningful hidden states such as customer preference, customer satisfaction, and customer relationship, as well as assessing the influential factors for transitions among different hidden states (e.g., Netzer et al., 2008; Ascarza et al., 2018; Ma et al., 2015b; Montoya et al., 2010). Our analysis further suggests that the NHMM framework can be applied by companies to obtain real-time customer information through identifying the latent needs that drive customers' mobile Internet use behaviors and dynamically monitoring and forecasting the transitions of customers' hidden states. This analysis enables companies to achieve better planning for their mobile network capacity so as to provide better services to customers, increase customer loyalty, and thereby generate more profits. All of these potential benefits, however, rely on the availability of an efficient technique to analyze ultra-long sequences with NHMMs.

The reminder of this paper is organized as follows: Section 2.2 describes the general setting of NHMMs. Section 2.3 introduces variational Bayesian inference and proposes a VB and an SVB method for NHMMs with long and ultra-long sequences, respectively. Section 2.4 conducts simulation studies to examine the performance of the proposed methods in comparison with conventional methods. Section 2.5 reports the analysis of applying the VB method in the NHMM modeling of the eye-tracking scan-paths. Section 2.6 presents the study using the SVB method to handle ultra-long mobile Internet usage

sequences in the NHMM modeling. Section 2.7 concludes the paper with some discussion.

2.2 Nonhomogeneous Hidden Markov Models

In this section, we introduce the specification of NHMMs. Consider an observed multivariate temporal process $\boldsymbol{y}_{\mathcal{T}} = \{\boldsymbol{y}_t\}_{t=1}^{T}$, where $\mathcal{T}$ denotes the entire time indices within the sequence, $\boldsymbol{y}_t = (y_{t1}, ..., y_{tR_y})'$, and R_y denotes the dimensionality of $\boldsymbol{y}_t$. Note that we assume one process here to suppress the subject index n for notational simplicity. To model dynamics of the observed process, an NHMM assumes $\boldsymbol{y}_{\mathcal{T}}$ to be a stochastic function of a hidden sequential process $z_{\mathcal{T}} = \{z_t\}_{t=1}^{T}$ which follows a nonhomogeneous discrete-time Markov chain with a finite state space $\{1, ..., K\}$. Given state z_t, each vector $\boldsymbol{y}_t$ is assumed to be conditionally independent of other $\boldsymbol{y}_{t^*}$ vectors and states z_{t^*}, for $t^* \neq t$.

The hidden process $z_{\mathcal{T}}$ is formulated by two components, initial state distribution and transition probability matrices. The initial state distribution $p(z_1|\boldsymbol{\pi})$ is defined as follows:

$$\mathbb{P}(z_1 = \pi_k|\boldsymbol{\pi}) = \pi_k, \quad k = 1, \ldots, K, \tag{2.1}$$

where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)'$. We follow a common practice to assume a uniform distribution for the initial state in this study (Meligkotsidou and Dellaportas, 2011).

The transition probability matrices of hidden process $z_{\mathcal{T}}$ are time varying and given as follows: for $t = 1, \ldots, T$,

$$\boldsymbol{Q}_t = [q_{t,k_1k_2}] = \begin{pmatrix} q_{t,11} & \cdots & q_{t,1K} \\ \vdots & \ddots & \vdots \\ q_{t,K1} & \cdots & q_{t,KK} \end{pmatrix}, \tag{2.2}$$

where

$$q_{t,k_1k_2} = \mathbb{P}(z_t = k_2|z_{t-1} = k_1)$$

denotes the transition probability from state $z_{t-1} = k_1$ at time $t-1$ to state $z_t = k_2$ at time t and

$$\sum_{k_2=1}^{K} q_{t,k_1k_2} = 1.$$

We consider two ways of modeling transition probabilities q_{t,k_1k_2}, depending on whether the identified hidden states follow a rank order. The transition probabilities q_{t,k_1k_2} for unordered hidden states such as the ROIs identified in our first motivating example may be modeled via the following multinomial logit regression (Meligkotsidou and Dellaportas, 2011; Ascarza et al., 2018): for $k_1, k_2 = 1, \ldots, K$ and $t = 2, \ldots, T$,

$$q_{t,k_1k_2} = \mathbb{P}(z_t = k_2 | z_{t-1} = k_1, \boldsymbol{\rho}, \boldsymbol{w}_t) = \frac{\exp(\rho_{k_1k_2,0} + \boldsymbol{w}_t'\boldsymbol{\rho}_{k_1k_2})}{\sum_{k=1}^{K} \exp(\rho_{k_1k,0} + \boldsymbol{w}_t'\boldsymbol{\rho}_{k_1k})}, \tag{2.3}$$

where $\rho_{k_1k_2,0}$ denotes a state-specific intercept, $\boldsymbol{w}_t$ is an R_w-dimensional covariate vector, and $\boldsymbol{\rho}_{k_1k_2}$ is an R_w-dimensional vector of coefficients that can be interpreted as conditional log odds ratios. Specifically, a higher value of $\rho_{k_1k_2,j}$ indicates that a higher value of $w_{t,j}$ increases the likelihood of transitioning to state $z_t = k_2$ relative to state $z_t = K$ at time t, conditional on the state being $z_{t-1} = k_1$ at time $t-1$. For identifiability, we set $\rho_{kK,0} = 0$ and $\boldsymbol{\rho}_{kK} = 0$ for all k. We utilize a full design here to model the transition probability matrices by allowing each transition probability to have its own set of coefficients, which captures detailed effects of exogenous covariates on state transitions and summarizes comprehensive impacts of HSCs on children's attention shifts in our first application.

If the hidden state is ordinal, such as the latent needs for mobile Internet in our second application, q_{t,k_1k_2} can be modeled by the following continuation-ratio logit model (Ip et al., 2013; Song et al., 2017): for $k_1 = 1, \ldots, K$, $k_2 = 1, \ldots, K-1$, and $t = 2, \ldots, T$,

$$\log\left(\frac{\mathbb{P}(z_t = k_2 | z_{t-1} = k_1)}{\mathbb{P}(z_t > k_2 | z_{t-1} = k_1)}\right) = \log\left(\frac{q_{t,k_1k_2}}{q_{t,k_1k_2+1} + \cdots + q_{t,k_1K}}\right) = \rho_{k_1k_2,0} + \boldsymbol{w}_t'\boldsymbol{\rho}_{k_1}, \tag{2.4}$$

where the same vector of coefficients $\boldsymbol{\rho}_{k_1}$ for each logit is assumed to follow the common proportional odds assumption in regression models for ordinal responses (McCullagh, 1980; Ip et al., 2013; Song et al., 2017). The model forms logits in a sequential manner and is appropriate for discrete responses that have a rank order. We apply this transition model in our second real application to examine the transitions between states of latent needs for mobile Internet. Given the hidden state z_t, the conditional distribution of the observed $\boldsymbol{y}_t$ vector at time t is modeled through the following emission distribution:

$$(\boldsymbol{y}_t|z_t = k) \sim \mathcal{P}(\boldsymbol{y}_t|z_t = k, \boldsymbol{x}_t, \boldsymbol{\beta}), \tag{2.5}$$

where $\boldsymbol{x}_t$ is an R_x-dimensional covariate vector, and $\boldsymbol{\beta}$ represents the set of parameters in the emission distribution of dimension R_β. The distributional choice for emission distribution is flexible including normal (Meligkotsidou and Dellaportas, 2011), Gamma mixture (Heaps et al., 2015; Holsclaw et al., 2017), and categorical distributions (Ascarza et al., 2018). The structure of NHMMs is illustrated in Figure 2.1.

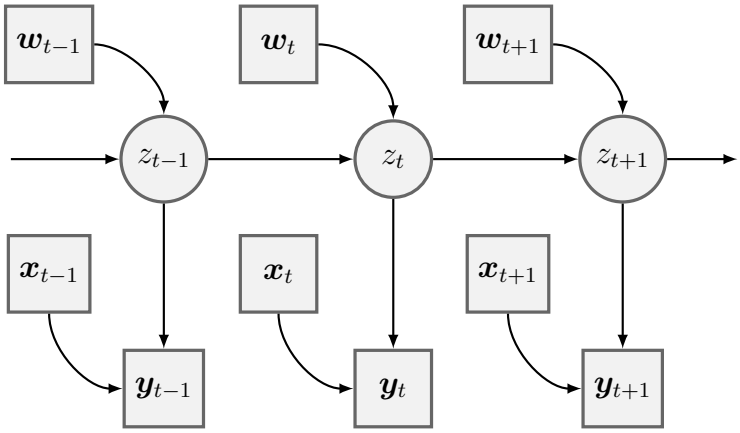


Figure 2.1: An illustration of the structure of NHMMs. $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are observed covariates and responses for the emission model, respectively. $\boldsymbol{w}_t$ are observed covariates for the transition model. z_t is the unobservable latent state variable.

2.3 Variational Bayesian Inference

In this section, we first briefly review aspects of the VB method, then propose a VB inference procedure for NHMMs with long observation sequences, and finally develop an efficient SVB procedure for NHMMs with ultra-long observation sequences.

2.3.1 Variational Bayes

VB is a Bayesian estimation approach that uses density functions from simple distribution families to approximate intractable posteriors (Jordan et al., 1999; Blei et al., 2017). Given a generic model $p(\boldsymbol{y}|\boldsymbol{\theta})$ with $\boldsymbol{y}$ denoting the observed data and $\boldsymbol{\theta}$ as unknown parameters, the aim of VB is to approximate the intractable posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ through a variational posterior distribution $\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ from a tractable variational distribution family $\widetilde{\mathcal{P}}$, where $\boldsymbol{\phi}$ is the set of variational parameters that govern the variational distribution. The distance between $\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is often measured by the Kullback-Leibler (KL) divergence

$$\begin{aligned}
\mathrm{KL}\left[\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\boldsymbol{y})\right] &= \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}}\left[\log \widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{y})\right] \\
&= \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}}\left[\log \widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{y},\boldsymbol{\theta})\right] + \log p(\boldsymbol{y}) \\
&= -\mathcal{L}(\boldsymbol{\phi}) + \log p(\boldsymbol{y}). \qquad (2.6)
\end{aligned}$$

Since $\mathrm{KL}\left[\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\boldsymbol{y})\right]$ is nonnegative, Equation (2.6) implies that $\mathcal{L}(\boldsymbol{\phi}) \leq \log p(\boldsymbol{y})$, $\forall \widetilde{p}_{\boldsymbol{\phi}}$. So $\mathcal{L}(\boldsymbol{\phi})$, being a lower bound for the log marginal likelihood, is called the evidence lower bound (ELBO) function. The optimal variational posterior can be obtained by maximizing $\mathcal{L}(\boldsymbol{\phi})$ over $\boldsymbol{\phi}$, which is performed via SGA (Robbins and Monro, 1951) in this study because the ELBO objective has no closed form in the setup of NHMMs. Letting $\nabla\mathcal{L}(\boldsymbol{\phi})$ denote the gradient of $\mathcal{L}(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$, after selecting an initial value $\boldsymbol{\phi}^{(0)}$,

$\boldsymbol{\phi}$ is updated via

$$\boldsymbol{\phi}^{(\tau+1)} = \boldsymbol{\phi}^{(\tau)} + \boldsymbol{\psi}_\tau \circ \nabla \mathcal{L}(\boldsymbol{\phi}^{(\tau)}), \tag{2.7}$$

where superscript (τ) denotes the τ-th iteration, operator $\circ$ denotes the Hadamard (element-wise) product, and $\{\boldsymbol{\psi}_\tau\}_{\tau \geq 0}$ contains a sequence of learning rates satisfying the Robbins-Monro conditions (Robbins and Monro, 1951). Each updating step involves determining values of the learning rates and the gradient.

Selecting appropriate learning rates in SGA helps improve its performance, which is typically done through a tuning procedure by hand (Hoffman et al., 2013). However, manual setting of learning rates before the analysis may cause the algorithm to converge slowly or even diverge. Therefore, adaptive methods adjusting the learning rates in each step are generally preferred (Kingma and Ba, 2015; Zeiler, 2012). In this study, we apply the well-received *Adam* approach for adaptive learning rates; *Adam* is efficient with limited tuning required, capable of handling noisy and/or spare gradients and non-stationary objectives, and particularly suitable for dealing with non-convex objective functions (Kingma and Ba, 2015). We proceed to briefly introduce the specification of the *Adam* optimization approach. Let $\phi_j^{(\tau)}$, $\psi_{\tau,j}$, and $\nabla l_j^{(\tau)}$ denote the j-th element of $\boldsymbol{\phi}^{(\tau)}$, $\boldsymbol{\psi}_\tau$, and $\nabla \mathcal{L}(\boldsymbol{\phi}^{(\tau)})$, respectively. The learning rate $\psi_{\tau,j}$ at the τth-iteration is computed as follows:

$$\psi_{\tau,j} = o_j \frac{\widehat{\chi}_{1,\tau,j}}{(\sqrt{\widehat{\chi}_{2,\tau,j}} + \epsilon_0)\nabla l_j^{(\tau)}},$$

where

$$\widehat{\chi}_{1,\tau,j} = \frac{\chi_{1,\tau,j}}{1-\epsilon_1^{\tau}},$$
$$\widehat{\chi}_{2,\tau,j} = \frac{\chi_{2,\tau,j}}{1-\epsilon_2^{\tau}},$$
$$\chi_{1,\tau,j} = \epsilon_1 \chi_{1,\tau-1,j} + (1-\epsilon_1)\nabla l_j^{(\tau)},$$
$$\chi_{2,\tau,j} = \epsilon_2 \chi_{2,\tau-1,j} + (1-\epsilon_2)\nabla l_j^{(\tau)^2},$$

and ϵ_0 ϵ_1, ϵ_2, and o_j are predefined hyperparameters. For this approach, we set starting values as $\chi_{1,0,j} = \chi_{2,0,j} = 0$ and hyperparameters as $\epsilon_0 = 10^{-8}$, $\epsilon_1 = 0.9$, $\epsilon_2 = 0.999$, $o_j = o_{1j}o_{2j}^{\tau}$, $o_{1j}, \in (10^{-4}, 10^{-2})$, and $o_{2j} \in (0.999, 0.9999)$, according to the suggestions in the literature (Shi et al., 2019; Kingma and Ba, 2015). For other details of the *Adam* approach, we refer readers to Kingma and Ba (2015).

For non-conjugate models such as NHMMs, the gradient $\nabla\mathcal{L}(\boldsymbol{\phi}^{(\tau)})$ cannot be computed analytically and thus is replaced by its estimate $\widehat{\nabla\mathcal{L}(\boldsymbol{\phi}^{(\tau)})}$ (Ranganath et al., 2014; Kucukelbir et al., 2017). To compute the gradient estimate, we utilize the reparameterization approach, which alleviates the variance control issue associated with the conventional Monte Carlo numerical integration method (Kingma and Welling, 2014; Ong et al., 2018). The reparameterization approach is applicable when $\boldsymbol{\theta}$ can be represented as $\boldsymbol{\theta} = t(\boldsymbol{\phi}, \boldsymbol{\zeta})$, where $\boldsymbol{\zeta}$ denotes a random vector with a known fixed distribution $f(\boldsymbol{\zeta})$. For instance, suppose $\boldsymbol{\theta}$ follows a multivariate normal variational distribution, then it can be reparameterized by the mean vector $\boldsymbol{\mu}$, the lower Cholesky factor $\boldsymbol{L}$ of its covariance matrix, and a standard normal random vector $\boldsymbol{\zeta}$ as $\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{\zeta}$. After reparameterization, the ELBO objective function can be rewritten as

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}}\left[\log p(\boldsymbol{y}, \boldsymbol{\theta}) - \log \widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\right] = \mathbb{E}_f\left[\log p(\boldsymbol{y}, t(\boldsymbol{\phi}, \boldsymbol{\zeta})) - \log \widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi}, \boldsymbol{\zeta}))\right],$$

and the gradient becomes

$$\nabla_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_f\left\{\frac{\mathrm{d}t(\boldsymbol{\phi},\boldsymbol{\zeta})}{\mathrm{d}\boldsymbol{\phi}}'\nabla_{\boldsymbol{\theta}}\left[\log p(\boldsymbol{y},t(\boldsymbol{\phi},\boldsymbol{\zeta})) - \log\widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi},\boldsymbol{\zeta}))\right] - \nabla_{\boldsymbol{\phi}}\log\widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi},\boldsymbol{\zeta}))\right\}. \tag{2.8}$$

The variance of the gradient estimate can be further reduced by dropping the last term in Equation (2.8) (Ong et al., 2018); the final formula for estimating gradient is given as:

$$\nabla_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_f\left\{\frac{\mathrm{d}t(\boldsymbol{\phi},\boldsymbol{\zeta})}{\mathrm{d}\boldsymbol{\phi}}'\nabla_{\boldsymbol{\theta}}\left[\log p(\boldsymbol{y},t(\boldsymbol{\phi},\boldsymbol{\zeta})) - \log\widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi},\boldsymbol{\zeta}))\right]\right\}, \tag{2.9}$$

which is computed with random samples generated from $f(\boldsymbol{\zeta})$.

The complexity and accuracy of the above VB method is significantly influenced by the choice of variational posterior family $\widetilde{\mathcal{P}}$ (Blei et al., 2017). In practice, a trade-off between approximation accuracy and computational complexity is sought. The mean-field family, which assumes complete independence for model parameters, is a common choice, but while it reduces computational complexity, the over-simplified form may lead to suboptimal approximations. In this study, we use a structured mean-field family that adds dependence among certain variational posteriors to better approximate the structures of true posteriors. The dependence is introduced based on the factor covariance structure developed by Ong et al. (2018). Numerical studies in latter sections show that the utilized posterior family performs satisfactorily.

2.3.2 Full-sequence VB for NHMMs with Long Sequences

Consider observation sequences $\{\boldsymbol{y}_{n,\mathcal{T}}, \boldsymbol{x}_{n,\mathcal{T}}, \boldsymbol{w}_{n,\mathcal{T}}\}$ for $n = 1,\ldots,N$ from N independent subjects. For notational simplicity, we assume the sequences have a common length T in the following discussion; extending the algorithm to handle unbalanced sequences is straightforward. Letting $\{z_{n,\mathcal{T}}\}_{n=1}^{N}$ denote the hidden state sequences, the complete-data

likelihood for NHMMs is

$$p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{w}) = \prod_{n=1}^{N} p(z_{n,1}) \prod_{t=1}^{T} p(\boldsymbol{y}_{n,t}|z_{n,t}, \boldsymbol{x}_{n,t}, \boldsymbol{\beta}) \prod_{t=2}^{T} p(z_{n,t}|z_{n,t-1}, \boldsymbol{w}_{n,t}, \boldsymbol{\rho}) p(\boldsymbol{\beta}) p(\boldsymbol{\rho}), \tag{2.10}$$

where $p(\boldsymbol{\beta})$ and $p(\boldsymbol{\rho})$ are priors, model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ are continuous, and $p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{w})$ is differentiable with respect to model parameters. Bayesian inference computes the posteriors of model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ as well as hidden state sequences $\{z_{n,\mathcal{T}}\}$, conditional on data and model setup. We develop a VB procedure to approximate the true posteriors.

We first specify the variational posteriors with the factorized form

$$\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{z}) = \widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta}) \widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho}) \widetilde{p}(\boldsymbol{z}),$$

where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_\beta, \boldsymbol{\phi}_\rho\}$ are the variational parameters. We allow dependence within each of $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, and $\boldsymbol{z}$ to achieve better approximation, and retain independence between $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, and $\boldsymbol{z}$ for computational tractability. The ELBO objective,

$$\mathcal{L} = \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}} \left\{ \mathbb{E}_{\widetilde{p}(\boldsymbol{z})} \left[\log p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{w}) - \log \widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta}) \widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho}) \widetilde{p}(\boldsymbol{z}) \right] \right\}, \tag{2.11}$$

is maximized by updating $\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})$, $\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})$, and $\widetilde{p}(\boldsymbol{z})$ iteratively.

The updating of $\widetilde{p}(\boldsymbol{z})$ involves calculating posteriors of $z_{n,t}$ and $(z_{n,t-1}, z_{n,t})$ with current values of variational parameters. The joint posterior of $\{z_{n,\mathcal{T}}\}$ is proportional to

$$\prod_{n=1}^{N} \exp \left\{ \sum_{t=1}^{T} \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})} \left[\log p(\boldsymbol{y}_{n,t}|z_{n,t}, \boldsymbol{x}_{n,t}, \boldsymbol{\beta}) \right] + \sum_{t=2}^{T} \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})} \left[\log p(z_{n,t}|z_{n,t-1}, \boldsymbol{w}_{n,t}, \boldsymbol{\rho}) \right] \right\}. \tag{2.12}$$

Based on the posterior, we compute the marginal posteriors $\widetilde{p}(z_{n,t})$ and $\widetilde{p}(z_{n,t-1}, z_{n,t})$ using the forward and backward probabilities of the Baum-Welch procedure (Baum et al., 1970).

The forward probabilities $a_{n,t}(k)$ and backward probabilities $b_{n,t}(k)$ are defined as: for $k = 1, \ldots, K$,

$$\begin{aligned} a_{n,t}(k) &= p(\boldsymbol{y}_{n,1:t}, z_{n,t} = k | \boldsymbol{x}_{n,1:t}, \boldsymbol{w}_{n,1:t}, \boldsymbol{\beta}, \boldsymbol{\rho}), \\ b_{n,t}(k) &= p(\boldsymbol{y}_{n,(t+1):T} | z_{n,t} = k, \boldsymbol{x}_{n,(t+1):T}, \boldsymbol{w}_{n,(t+1):T}, \boldsymbol{\beta}, \boldsymbol{\rho}). \end{aligned} \tag{2.13}$$

They can be computed iteratively using the following formulas:

$$\begin{aligned} \boldsymbol{a}'_{n,1} &= \boldsymbol{\pi}'_n \boldsymbol{D}_{n,1}, \\ \boldsymbol{a}'_{n,t+1} &= \boldsymbol{a}'_{n,t} \boldsymbol{E}_{n,t+1} \boldsymbol{D}_{n,t+1}, \\ \boldsymbol{b}_{n,T} &= \mathbf{1}, \\ \boldsymbol{b}_{n,t} &= \boldsymbol{E}_{n,t+1} \boldsymbol{D}_{n,t+1} \boldsymbol{b}_{n,t+1}, \end{aligned} \tag{2.14}$$

where $\boldsymbol{a}_{n,t} = (a_{n,t}(1), ..., a_{n,t}(K))'$, $\boldsymbol{b}_{n,t} = (b_{n,t}(1), ..., b_{n,t}(K))'$, $\mathbf{1}$ is a K-dimensional vector with all elements being 1, and $\boldsymbol{D}_{n,t}$ and $\boldsymbol{E}_{n,t}$ are diagonal and square matrices, respectively, as defined below. To prevent numerical underflow issues, $\boldsymbol{a}_{n,t}$ and $\boldsymbol{b}_{n,t}$ are normalized at each iteration. The diagonal elements of $\boldsymbol{D}_{n,t}$ are given as follows: for $k = 1, \ldots, K$,

$$d_{n,t,kk} = \exp\left\{ \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})} \left[\log p(\boldsymbol{y}_{n,t} | z_{n,t} = k, \boldsymbol{x}_{n,t}, \boldsymbol{\beta}) \right] \right\}. \tag{2.15}$$

Elements in $\boldsymbol{E}_{n,t}$ are given as follows: for $k_1, k_2 = 1, \ldots, K$

$$e_{n,t,k_1k_2} = \exp\left\{ \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})} \left[\log p(z_{n,t} = k_2 | z_{n,t-1} = k_1, \boldsymbol{w}_{n,t}, \boldsymbol{\rho}) \right] \right\}. \tag{2.16}$$

The above two expectations are computed numerically using Monte Carlo samples from $\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})$ and $\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})$. Note that $\boldsymbol{D}_{n,t}$ and $\boldsymbol{E}_{n,t}$ can be viewed as variational estimates of emission probabilities in Equation (2.5) and transition probability matrices in Equation (2.2)

respectively. Marginal posteriors $\widetilde{p}(z_{n,t})$ and $\widetilde{p}(z_{n,t-1}, z_{n,t})$ can be computed as follows:

$$\begin{aligned}\widetilde{p}(z_{n,t} = k) &\propto a_{n,t}(k) b_{n,t}(k) \\ &= \frac{a_{n,t}(k) b_{n,t}(k)}{\sum_j a_{n,t}(j) b_{n,t}(j)},\end{aligned} \tag{2.17}$$

and

$$\begin{aligned}\widetilde{p}(z_{n,t-1} = k_1, z_{n,t} = k_2) &\propto a_{n,t-1}(k_1) e_{n,t,k_1k_2} d_{n,t,k_2k_2} b_{n,t}(k_2) \\ &= \frac{a_{n,t-1}(k_1) e_{n,t,k_1k_2} d_{n,t,k_2k_2} b_{n,t}(k_2)}{\sum_{j_1} \sum_{j_2} a_{n,t-1}(j_1) e_{n,t,j_1j_2} d_{n,t,j_2j_2} b_{n,t}(j_2)}.\end{aligned} \tag{2.18}$$

The updating of variational parameters $\boldsymbol{\phi}_\rho$ and $\boldsymbol{\phi}_\beta$ is conducted through gradient ascent. We assume all model parameters are unconstrained in the following discussion and the constrained parameters can be converted to unconstrained ones; for instance, a nonnegative parameter $\sigma \geq 0$ can be represented by e^ς with $\varsigma \in \mathbb{R}$ (Kucukelbir et al., 2017). By converting parameters into the same space, the advanced Gaussian variational family with a factor covariance structure developed by Ong et al. (2018) can be applied to all parameters, which flexibly captures the complex dependence structure in NHMMs and attains model parsimony that enables efficient estimation. For the updating of $\boldsymbol{\phi}_\rho$, we discuss the transition model of multinomial logit regression with a full design given by Equation (2.3). As for the transition model of continuation-ratio logit regression given by Equation (2.4), the updating procedure is similar. We assign the following factorized variational posterior for $\boldsymbol{\rho}$:

$$\prod_{k_1=1}^{K} \prod_{k_2=1}^{K-1} \widetilde{p}_{\boldsymbol{\phi}_{\rho,k_1k_2}}(\boldsymbol{\rho}_{k_1k_2}), \tag{2.19}$$

where the intercept $\rho_{k_1k_2,0}$ is included into $\boldsymbol{\rho}_{k_1k_2}$ and inferred together, and $\widetilde{p}_{\boldsymbol{\phi}_{\rho,k_1k_2}}(\boldsymbol{\rho}_{k_1k_2})$ is a $(R_w + 1)$-dimensional multivariate normal distribution with a factor covariance structure

(Ong et al., 2018):

$$\widetilde{p}_{\boldsymbol{\phi}_{\rho,k_1k_2}}(\boldsymbol{\rho}_{k_1k_2}) = \mathcal{N}\left(\boldsymbol{\rho}_{k_1k_2}; \boldsymbol{\mu}_{\rho,k_1k_2}, \boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{G}'_{\rho,k_1k_2} + \boldsymbol{H}^2_{\rho,k_1k_2}\right). \tag{2.20}$$

Thus, $\boldsymbol{\phi}_{\rho,k_1k_2} = \{\boldsymbol{\mu}_{\rho,k_1k_2}, \boldsymbol{G}_{\rho,k_1k_2}, \boldsymbol{H}_{\rho,k_1k_2}\}$ are variational parameters to be estimated, where $\boldsymbol{\mu}_{\rho,k_1k_2}$ is the variational mean vector for $\boldsymbol{\rho}_{k_1k_2}$, $\boldsymbol{G}_{\rho,k_1k_2}$ is a $(R_w+1) \times r_{\rho,k_1k_2}$ matrix with $r_{\rho,k_1k_2} \leq R_w + 1$, r_{ρ,k_1k_2} denotes the number of factors used to approximate the correlation among elements in $\boldsymbol{\rho}_{k_1k_2}$, the upper triangle of $\boldsymbol{G}_{\rho,k_1k_2}$ is restricted to 0 for identification, and $\boldsymbol{H}_{\rho,k_1k_2}$ is a diagonal matrix with diagonal elements $\boldsymbol{h}_{\rho,k_1k_2} = (h_{\rho,k_1k_2,1}, ..., h_{\rho,k_1k_2,R_w+1})'$. As demonstrated in Ong et al. (2018), a small number of factors ($r_{\rho,k_1k_2} = 3$ or 4) already provides satisfactory approximation to high-dimensional posteriors. Indeed, approximation accuracy can be improved if we increase the number of factors r_{ρ,k_1k_2} used, which also raises the optimization difficulty; if we set $r_{\rho,k_1k_2} = R_w + 1$, we are using a multivariate normal distribution with a full covariance structure, which yields the closest approximation. Note that, in our applications, we generally set $r_{\rho,k_1k_2} = R_w + 1$ to obtain close approximation, considering that the dimensionality of posteriors is relatively low.

With the factorized variational posterior for $\boldsymbol{\rho}$, updating $\boldsymbol{\phi}_\rho$ reduces to the iterative update of $\boldsymbol{\phi}_{\rho,k_1k_2}$. To update $\boldsymbol{\phi}_{\rho,k_1k_2}$, we need to compute the gradient estimates, $\widehat{\nabla_{\boldsymbol{\mu}_{\rho,k_1k_2}}\mathcal{L}}$, $\widehat{\nabla_{\boldsymbol{G}_{\rho,k_1k_2}}\mathcal{L}}$, and $\widehat{\nabla_{\boldsymbol{h}_{\rho,k_1k_2}}\mathcal{L}}$, and then update $\boldsymbol{\mu}_{\rho,k_1k_2}$, $\boldsymbol{G}_{\rho,k_1k_2}$, and $\boldsymbol{H}_{\rho,k_1k_2}$ sequentially according to the formula (2.7). The reparametrization that represents $\boldsymbol{\rho}_{k_1k_2}$ as follows:

$$\boldsymbol{\rho}_{k_1k_2} = \boldsymbol{\mu}_{\rho,k_1k_2} + \boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{\zeta}_{\rho,k_1k_2,1} + \boldsymbol{h}_{\rho,k_1k_2} \circ \boldsymbol{\zeta}_{\rho,k_1k_2,2},$$

where $\boldsymbol{\zeta}_{\rho,k_1k_2,1}$ and $\boldsymbol{\zeta}_{\rho,k_1k_2,2}$ are independent standard normal random vectors of dimensions r_{ρ,k_1k_2} and $R_w + 1$, respectively. The gradient estimates are computed numerically using samples of $\boldsymbol{\zeta}_{\rho,k_1k_2,1}$ and $\boldsymbol{\zeta}_{\rho,k_1k_2,2}$.

Finally, the updating of $\boldsymbol{\phi}_\beta$ is similar to that of $\boldsymbol{\phi}_\rho$. The possible constrained parameters in the emission model can be converted to unconstrained ones; for instance, a nonnegative parameter $\sigma \geq 0$ can be represented by e^ς with $\varsigma \in \mathbb{R}$ (Kucukelbir et al., 2017). By converting parameters into the same space, the advanced Gaussian variational family with a factor covariance structure developed by Ong et al. (2018) can be applied to all parameters, which flexibly captures the complex dependence structure in NHMMs and attains model parsimony that enables efficient estimation. For the K sets of R_β-dimensional emission model parameters $\{\boldsymbol{\beta}_k\}$, we assign them with similar factorized multivariate normal variational posteriors as follows:

$$\prod_{k=1}^{K} \widetilde{p}(\boldsymbol{\beta}_k) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\beta}_k; \boldsymbol{\mu}_{\beta,k}, \boldsymbol{G}_{\beta,k}\boldsymbol{G}'_{\beta,k} + \boldsymbol{H}^2_{\beta,k}\right), \tag{2.21}$$

where $\boldsymbol{\mu}_{\beta,k}$, $\boldsymbol{G}_{\beta,k}$, and $\boldsymbol{H}_{\beta,k}$ are defined similarly. The number of factors is denoted as $r_{\beta,k}$. The variational parameters $\boldsymbol{\phi}_{\beta,k} = \{\boldsymbol{\mu}_{\beta,k}, \boldsymbol{G}_{\beta,k}, \boldsymbol{H}_{\beta,k}\}$ can be updated similarly as above.

The label-switching issue of the NHMM elicited by the invariance of the likelihood to a random permutation of the state labels is addressed by incorporating the permutation sampling idea from Frühwirth-Schnatter (2001). At each iteration, we adjust the state labels of emission variational parameters $\boldsymbol{\phi}_k$. The proposed VB method is guaranteed to converge to an local optimum; we terminate the algorithm after a sufficient number of iterations when convergence is attained (Blei et al., 2017; Ong et al., 2018; Hoffman et al., 2013). Detailed procedures and formulas of the VB method are summarized as follows:

Step 1: Specify the number of iterations for the algorithm, *Iter*. Initialize variational parameters, $\mu_{\rho,k_1k_2,0}$, $h_{\rho,k_1k_2,0}$, $\boldsymbol{\mu}_{\rho,k_1k_2}$, $\boldsymbol{G}_{\rho,k_1k_2}$, and $\boldsymbol{h}_{\rho,k_1k_2}$, for $k_1 = 1, ..., K$ and $k_2 = 1, ..., K-1$, in Equations (2.19) - (2.20) for the transition model. Randomly initialize variational parameters, $\boldsymbol{\mu}_{\beta,k}$, $\boldsymbol{G}_{\beta,k}$, and $\boldsymbol{H}_{\beta,k}$, for $k = 1, ..., K$, in Equation (2.21) for the emission model.

The procedure iteratively performs the following steps until *Iter* is reached.

Step 2: Randomly sample N_s subjects from the total N subjects without replacement.

Step 3: Update the posteriors for hidden states $z_{n,t}$ based on the current values of variational parameters and the sampled N_s sequences. Specifically, $\boldsymbol{D}_{n,t}$ and $\boldsymbol{E}_{n,t}$ are computed for the sampled m sequences following Equations (2.15) - (2.16), which are subsequently used to compute forward and backward probabilities following Equations (2.13) - (2.14). To address numerical underflow issues, we normalize the forward and backward probabilities for each time point t. The variational posteriors are finally computed following Equations (2.17) - (2.18).

Step 4: Update variational parameters, $\boldsymbol{\mu}_{\beta,k}$, $\boldsymbol{G}_{\beta,k}$, and $\boldsymbol{H}_{\beta,k}$, for the emission model using SGA. The adaptive learning rates are determined using the introduced *Adam* approach. The gradient with respect to each parameter vector is obtained based on the general formula given in Equation (2.9). Let $\mathcal{L}_{N_s}$ denote the ELBO (defined in Equation [2.11]) computed from the sampled N_s sequences. The specific gradient with respect to $\boldsymbol{\mu}_{\beta,k}$ is

$$\nabla_{\boldsymbol{\mu}_{\beta,k}}\mathcal{L}_{N_s} = \mathbb{E}_f\Big\{\nabla_{\boldsymbol{\beta}_k}\left[\log p_{N_s}(\boldsymbol{\mu}_{\beta,k} + \boldsymbol{G}_{\beta,k}\boldsymbol{\zeta}_{\beta,k,1} + \boldsymbol{h}_{\beta,k}\circ\boldsymbol{\zeta}_{\beta,k,2})\right] + \left(\boldsymbol{G}_{\beta,k}\boldsymbol{G}'_{\beta,k} + \boldsymbol{h}^2_{\beta,k}\right)^{-1}\left(\boldsymbol{G}_{\beta,k}\boldsymbol{\zeta}_{\beta,k,1} + \boldsymbol{h}_{\beta,k}\circ\boldsymbol{\zeta}_{\beta,k,2}\right)\Big\},$$

where p_{N_s} is the likelihood function computed from the sampled N_s sequences after marginalizing hidden states with posteriors computed from the last step, $\boldsymbol{h}_{\beta,k}$ is the diagonal elements of $\boldsymbol{H}_{\beta,k}$, and $\boldsymbol{\zeta}_{\beta,k,1}$ and $\boldsymbol{\zeta}_{\beta,k,2}$ are $r_{\beta,k}$ and R_β dimensional vectors with elements sampled from independent standard univariate normal distribution, respectively. Note that we only show the argument related to $\boldsymbol{\beta}_k$ for the likelihood function in the above formula

for notational simplicity. The gradient with respect to $\boldsymbol{G}_{\beta,k}$ is

$$\nabla_{\boldsymbol{G}_{\beta,k}}\mathcal{L}_{N_s} = \mathbb{E}_f\Big\{\nabla_{\boldsymbol{\beta}_k}\left[\log p_{N_s}(\boldsymbol{\mu}_{\beta,k} + \boldsymbol{G}_{\beta,k}\boldsymbol{\zeta}_{\beta,k,1} + \boldsymbol{h}_{\beta,k}\circ\boldsymbol{\zeta}_{\beta,k,2})\right]\boldsymbol{\zeta}'_{\beta,k,1} + \left(\boldsymbol{G}_{\beta,k}\boldsymbol{G}'_{\beta,k} + \boldsymbol{H}^2_{\beta,k}\right)^{-1}\left(\boldsymbol{G}_{\beta,k}\boldsymbol{\zeta}_{\beta,k,1} + \boldsymbol{h}_{\beta,k}\circ\boldsymbol{\zeta}_{\beta,k,2}\right)\boldsymbol{\zeta}'_{\beta,k,1}\Big\}.$$

The gradient with respect to $\boldsymbol{h}_{\beta,k}$ is

$$\nabla_{\boldsymbol{h}_{\beta,k}}\mathcal{L}_{N_s} = \mathbb{E}_f\Big\{\mathrm{diag}\Big(\nabla_{\boldsymbol{\beta}_k}\left[\log p_{N_s}(\boldsymbol{\mu}_{\beta,k} + \boldsymbol{G}_{\beta,k}\boldsymbol{\zeta}_{\beta,k,1} + \boldsymbol{h}_{\beta,k}\circ\boldsymbol{\zeta}_{\beta,k,2})\right]\boldsymbol{\zeta}'_{\beta,k,2} + \left(\boldsymbol{G}_{\beta,k}\boldsymbol{G}'_{\beta,k} + \boldsymbol{H}^2_{\beta,k}\right)^{-1}\left(\boldsymbol{G}_{\beta,k}\boldsymbol{\zeta}_{\beta,k,1} + \boldsymbol{h}_{\beta,k}\circ\boldsymbol{\zeta}_{\beta,k,2}\right)\boldsymbol{\zeta}'_{\beta,k,2}\Big)\Big\}.$$

The expectations are computed numerically by generating samples $\boldsymbol{\zeta}_{\beta,k,1}$ and $\boldsymbol{\zeta}_{\beta,k,2}$ from respective standard normal distributions. Adjust state labels using permutation sampler in Frühwirth-Schnatter (2001).

Step 5: Update variational parameters, $\mu_{\rho,k_1k_2,0}$, $h_{\rho,k_1k_2,0}$, $\boldsymbol{\mu}_{\rho,k_1k_2}$, $\boldsymbol{G}_{\rho,k_1k_2}$, and $\boldsymbol{h}_{\rho,k_1k_2}$, for the transition model using SGA. The adaptive learning rates are determined using the introduced *Adam* approach. The specific gradient with respect to $\mu_{\rho,k_1k_2,0}$ is

$$\nabla_{\mu_{\rho,k_1k_2,0}}\mathcal{L}_{N_s} = \mathbb{E}_f\left\{\nabla_{\rho_{k_1k_2,0}}\left[\log p_{N_s}(\mu_{\rho,k_1k_2,0} + h_{\rho,k_1k_2,0}\zeta_{\rho,k_1k_2,0})\right] + h^{-1}_{\rho,k_1k_2,0}\zeta_{\rho,k_1k_2,0}\right\},$$

where $\zeta_{\rho,k_1k_2,0}$ is sampled from a univariate standard normal distribution. The gradient with respect to $h_{\rho,k_1k_2,0}$ is

$$\nabla_{h_{\rho,k_1k_2,0}}\mathcal{L}_{N_s} = \mathbb{E}_f\left\{\nabla_{\rho_{k_1k_2,0}}\left[\log p_{N_s}(\mu_{\rho,k_1k_2,0} + h_{\rho,k_1k_2,0}\zeta_{\rho,k_1k_2,0})\right]\zeta_{\rho,k_1k_2,0} + h^{-1}_{\rho,k_1k_2,0}\zeta^2_{\rho,k_1k_2,0}\right\}.$$

The gradient with respect to $\boldsymbol{\mu}_{\rho,k_1k_2}$ is

$$\nabla_{\boldsymbol{\mu}_{\rho,k_1k_2}}\mathcal{L}_{N_s} = \mathbb{E}_f\Big\{\nabla_{\boldsymbol{\rho}_{k_1k_2}}\left[\log p_{N_s}(\boldsymbol{\mu}_{\rho,k_1k_2} + \boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{\zeta}_{\rho,k_1k_2,1} + \boldsymbol{h}_{\rho,k_1k_2}\circ\boldsymbol{\zeta}_{\rho,k_1k_2,2})\right] + \left(\boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{G}'_{\rho,k_1k_2} + \boldsymbol{h}^2_{\rho,k_1k_2}\right)^{-1}\left(\boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{\zeta}_{\rho,k_1k_2,1} + \boldsymbol{h}_{\rho,k_1k_2}\circ\boldsymbol{\zeta}_{\rho,k_1k_2,2}\right)\Big\}.$$

The gradient with respect to $\boldsymbol{G}_{\rho,k_1k_2}$ is

$$\nabla_{\boldsymbol{G}_{\rho,k_1k_2}}\mathcal{L}_{N_s} = \mathbb{E}_f\Big\{\nabla_{\boldsymbol{\rho}_{k_1k_2}}\left[\log p_{N_s}(\boldsymbol{\mu}_{\rho,k_1k_2} + \boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{\zeta}_{\rho,k_1k_2,1} + \boldsymbol{h}_{\rho,k_1k_2}\circ\boldsymbol{\zeta}_{\rho,k_1k_2,2})\right] \cdot \boldsymbol{\zeta}'_{\rho,k_1k_2,1} + \left(\boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{G}'_{\rho,k_1k_2} + \boldsymbol{h}^2_{\rho,k_1k_2}\right)^{-1}\left(\boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{\zeta}_{\rho,k_1k_2,1} + \boldsymbol{h}_{\rho,k_1k_2}\circ\boldsymbol{\zeta}_{\rho,k_1k_2,2}\right)\boldsymbol{\zeta}'_{\rho,k_1k_2,1}\Big\}.$$

The gradient with respect to $\boldsymbol{h}_{\rho,k_1k_2}$ is

$$\nabla_{\boldsymbol{h}_{\rho,k_1k_2}}\mathcal{L}_{N_s} = \mathbb{E}_f\Big\{\mathrm{diag}\Big(\nabla_{\boldsymbol{\rho}_{k_1k_2}}\left[\log p_{N_s}(\boldsymbol{\mu}_{\rho,k_1k_2} + \boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{\zeta}_{\rho,k_1k_2,1} + \boldsymbol{h}_{\rho,k_1k_2}\circ\boldsymbol{\zeta}_{\rho,k_1k_2,2})\right] \cdot \boldsymbol{\zeta}'_{\rho,k_1k_2,2} + \left(\boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{G}'_{\rho,k_1k_2} + \boldsymbol{h}^2_{\rho,k_1k_2}\right)^{-1}\left(\boldsymbol{G}_{\rho,k_1k_2}\boldsymbol{\zeta}_{\rho,k_1k_2,1} + \boldsymbol{h}_{\rho,k_1k_2}\circ\boldsymbol{\zeta}_{\rho,k_1k_2,2}\right)\boldsymbol{\zeta}'_{\rho,k_1k_2,2}\Big)\Big\}.$$

The expectations are computed numerically by generating samples $\zeta_{\rho,k_1k_2,0}$, $\boldsymbol{\zeta}_{\rho,k_1k_2,1}$, and $\boldsymbol{\zeta}_{\rho,k_1k_2,2}$ from respective standard normal distributions. Adjust state labels using permutation sampler in Frühwirth-Schnatter (2001).

Step 6: Compute the ELBO value $\mathcal{L}_{N_s}$ using current variational parameters.

The time complexity per update is of order $O(NT)$ because the entire dataset is used. The developed method is capable of efficiently handling long sequences of a length T up to 10^4.The time complexity of $O(NT)$ indicates that the VB method's efficiency faces challenges when dealing with massive datasets, that is, when N or T is ultra-large. The cases with large N can be handled easily via subsampling because of the independence among subjects. In specific, we may randomly sample N_s out of the N subjects at each iteration ($N_s = 1$ is sufficient for most applications) and update parameters based on the

subjects; this can be achieved using the stochastic variational inference procedure developed by Hoffman et al. (2013). The time complexity of the algorithm thus reduces significantly to $O(N_s T)$. The cases with ultra-large T, however, present major methodological challenges.

2.3.3 SVB for NHMMs with Ultra-long Sequences

Applying NHMMs to datasets with ultra-large T ($T > 10^4$) could not be simply achieved through subsampling, as simply sampling several subsequences from a full sequence breaks dependence in an HMM's hidden state sequence and thus may produce significant biases (Aicher et al., 2019). Recently, the subsequence method developed by Aicher et al. (2019), Foti et al. (2014), Ma et al. (2017), and Ye (2018) proposes to control for the bias through attaching sufficiently long buffers at both ends of each subsequence under the homogeneous HMM framework. Here we develop a new efficient SVB method to analyze NHMMs with ultra-long sequences; we highlight that, unlike homogeneous HMMs, NHMMs require careful consideration of local nonhomogeneity in the subsequence method.

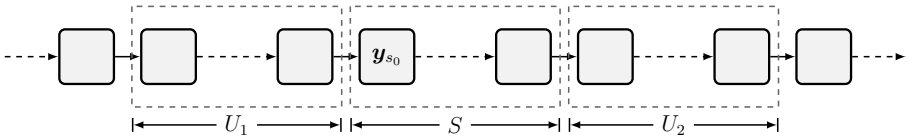


Figure 2.2: An illustration of a subsequence with two buffers. S is the length of the randomly sampled subsequence starting at time index s_0. U_1 and U_2 are buffers attached to the subsequence in order to control the estimation bias caused by breaking sequential dependence from subsampling.

An illustration of a subsequence and its corresponding buffers is shown in Figure 2.2, where S, U_1 and U_2 denote the lengths of the subsequence and the two buffers, respectively, and s_0 denotes the starting time index of the subsequence. Below, we show that direct subsampling from the sequence causes bias and how attaching buffers can help reduce the generated bias. Applying the forward-backward algorithm to the subsequence S directly leads to inaccurate specifications of the starting forward and backward probabilities (i.e.,

forward probabilities at s_0 and backward probabilities at $s_0 + S - 1$) as follows:

$$\boldsymbol{a}^{*'}_{n,s_0} = \boldsymbol{\pi}'_{n,s_0}\boldsymbol{D}_{n,s_0}, \quad \boldsymbol{b}^{*}_{n,s_0+S-1} = \mathbf{1}.$$

Note that the forward and backward probabilities are normalized in each iteration from the Baum-Welch procedure. The error distances between the normalized inaccurate specifications and the normalized true forward and backward probabilities are thus represented as follows:

$$\begin{aligned} \text{dist}(\bar{\boldsymbol{a}}^{*}_{n,s_0}, \bar{\boldsymbol{a}}_{n,s_0}) &= \left\| \bar{\boldsymbol{a}}^{*}_{n,s_0} - \bar{\boldsymbol{a}}_{n,s_0} \right\|, \\ \text{dist}(\bar{\boldsymbol{b}}^{*}_{n,s_0+S-1}, \bar{\boldsymbol{b}}_{n,s_0+S-1}) &= \left\| \bar{\boldsymbol{b}}^{*}_{n,s_0+S-1} - \bar{\boldsymbol{b}}_{n,s_0+S-1} \right\|, \end{aligned} \tag{2.22}$$

where $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{b}}$ denote normalized vectors, and $\|\cdot\|$ denotes the l^2 norm. The error distances can be substantial, and thus, may cause significant bias in statistical inference. The bias can be reduced by adding buffers since the error distances decay exponentially as the buffer length grows. Specifically, we can inaccurately specify the starting forward and backward probabilities at the outside endpoints (i.e., $s_0 - U_1$ and $s_0 + S + U_2 - 1$) of the buffers as

$$\boldsymbol{a}^{*'}_{n,s_0-U_1} = \boldsymbol{\pi}'_{n,s_0-U_1}\boldsymbol{D}_{n,s_0-U_1}, \quad \boldsymbol{b}^{*}_{n,s_0+S+U_2-1} = \mathbf{1}.$$

The memory decay property of HMMs guarantees that forward and backward probabilities converge exponentially from different starting values (Le Gland and Mevel, 2000a,b). That is, though the starting forward and backward probabilities are still specified inaccurately, they converge to the true probabilities at both endpoints of the subsequence as lengths of the buffers increase, so that the error distances in Equation (2.22) are reduced exponentially. Therefore, at each iteration of the proposed SVB procedure, we first randomly sample N_s subjects. From each of the N_s full sequences, M subsequences of equal length S are then

randomly sampled, with two buffers attached to each sampled subsequence determined subsequently. The forward-backward algorithm is performed on the buffered subsequences to control for the error distances of forward-backward probabilities. Finally, variational parameters are updated solely based on the subsequences without buffers using procedures developed above. Hence, we need to determine the number of subsequences (i.e., M) and lengths of each subsequence (i.e., S) and the attached buffers (i.e., $U_{n,m,1}$ and $U_{n,m,2}$).

Choosing large values for M and S tends to improve the estimation accuracy but increases the computation complexity considerably. A balance needs to be attained when determining M and S. Existing studies (Foti et al., 2014; Ma et al., 2017) mention that M and S should be set according to the transition patterns of the Markov chain in homogeneous HMMs. For instance, if a Markov chain transits frequently between states, a large S is preferable to achieve a sample well representing the information of transition dynamics. Note that the transition patterns of the Markov chain in an NHMM may vary over time due to possible complex nonhomogeneity, which requires that both M and S should be moderately large at least to reflect crucial features of the NHMM. From the numerical studies in the literature and our experiments, we find that the setting of $S = 10$ to 20 and $M = 10$ to 20 provided overall satisfactory performance. We further recommend a sensitivity analysis by altering M and S in applications to evaluate whether estimation results are sensitive to the choices.

The challenging part of the SVB method is to determine buffer lengths ($U_{n,m,1}$ and $U_{n,m,2}$) for each subsequence, so that the buffers are not only long enough to control the bias from subsampling but also sufficiently short to attain computational efficiency. Aicher et al. (2019) propose to use buffers of a fixed length for homogeneous HMMs. However, there is no guarantee that, with fixed length buffers, the bias can be reduced to the desired level for NHMMs. Another stream of research (Ye, 2018; Foti et al., 2014; Ma et al., 2017) proposes to determine precise buffer lengths based on the memory decay property of homogeneous HMMs so that the bias are controlled effectively. We here adopt the latter scheme and

extend it by considering local nonhomogeneity in NHMMs and proposing the use of LLEs to quantify local memory decay rates of NHMMs and estimate buffer lengths adaptively.

The aim of attaching buffers is to reduce the error distances between the inaccurately specified forward and backward probabilities and their true counterparts on the subsequence. The following result given in Collet and Leonardi (2014) provides a theoretical basis for the memory decay rates of HMMs. Define $\boldsymbol{M}_{n,t,t} = \boldsymbol{Q}_{n,t}\boldsymbol{P}_{n,t}$ and $\boldsymbol{M}_{n,t,s} = \boldsymbol{Q}_{n,s}\boldsymbol{P}_{n,s}\cdots\boldsymbol{Q}_{n,t}\boldsymbol{P}_{n,t}$ for $s < t$, where $\boldsymbol{Q}_{n,t}$ is the transition probability matrix in Equation (2.2), and $\boldsymbol{P}_{n,t}$ denotes a $K \times K$ diagonal matrix with the kth diagonal element being the emission probability $p_{n,t,kk} = p(\boldsymbol{y}_{n,t}|z_{n,t} = k, \boldsymbol{x}_{n,t}, \boldsymbol{\beta})$. Collet and Leonardi (2014) show that, assume $\boldsymbol{P}_{n,t}$ and $\boldsymbol{Q}_{n,t}$ are positive, there exists $\lambda_n < 0$ such that

$$\limsup_{t\to\infty} \frac{1}{t} \log \left\| \bar{\boldsymbol{a}}^*_{n,t} - \bar{\boldsymbol{a}}_{n,t} \right\| = \lambda_n, \quad a.s., \tag{2.23}$$

where $\lambda_n = \lambda_{n,2} - \lambda_{n,1}$, and $\lambda_{n,1}$ and $\lambda_{n,2}$ are the first two Lyapunov exponents of $\boldsymbol{M}_{n,t,1}$, as $t \to \infty$. Here we focus on the forward probability vectors $\boldsymbol{a}^*_{n,t}$ and $\boldsymbol{a}_{n,t}$, and the corresponding buffer length $U_{n,m,1}$ in the above theorem and the following procedure of determining buffer lengths; the theorem and the procedure apply similarly to backward probability vectors $\boldsymbol{b}^*_{n,t}$ and $\boldsymbol{b}_{n,t}$, and the corresponding buffer length $U_{n,m,2}$.

The first two components of the Lyapunov spectrum of the whole HMM system (i.e., $\lambda_{n,1}$ and $\lambda_{n,2}$ from Equation [2.23]) are commonly used in characterizing the long-run memory decay rate (i.e., over infinite time) of an HMM, and thus, are referred to as *Global Lyapunov Exponents (GLEs)*. For a homogeneous HMM with time-invariant transition probability matrix and emission distributions, this global decay rate $\exp(\lambda_n)$ roughly applies to buffer length estimations for any subsequences sampled from the full sequence; therefore, Ye (2018) proposes that to control the error distance at s_0 under a specified level δ, the

buffer length $U_{n,m,1}$ should be set as

$$U_{n,m,1} = [\log(\delta)/(\lambda_n)], \quad \forall m,$$

where the operator $[\cdot]$ indicates rounding to the nearest integer, and λ_n is estimated numerically. This method is effective for homogeneous HMMs but could be compromised by the nonhomogeneity of NHMMs, as demonstrated in the following example.

Example 1: *Consider a sequence of a length* $T = 10^5$ *from an NHMM with* $K = 3$ *states defined by Equations (2.1) - (2.3) and (2.5). For the transition model (2.3), the covariate vector* $\boldsymbol{w}_t = (w_{t,1}, w_{t,2})'$ *is set as* $(0.45, 0.45)'$ *for* $t = 1, \ldots, 10^3$ *and* $(-0.45, -0.45)'$ *for* $t = 1001, \ldots, 10^5$*, and the parameters are set as* $\rho_{11,0} = 2$*,* $\rho_{12,0} = -2$*,* $\rho_{21,0} = -2$*,* $\rho_{22,0} = 2$*,* $\rho_{31,0} = -2$*,* $\rho_{32,0} = -2$*,* $\boldsymbol{\rho}_{11} = (2, 2)'$*,* $\boldsymbol{\rho}_{12} = (-2, -2)'$*,* $\boldsymbol{\rho}_{21} = (-2, -2)'$*,* $\boldsymbol{\rho}_{22} = (2, 2)'$*,* $\boldsymbol{\rho}_{32} = (-2, -2)'$*, and* $\boldsymbol{\rho}_{32} = (-2, -2)'$*. The emission distribution is set as* $(\boldsymbol{y}_t | z_t = k) \sim \mathcal{N}(\mu_{yk}, \sigma^2_{yk})$*, for* $k = 1, 2, 3$*, where* $\mu_{y1} = -0.5$*,* $\mu_{y2} = 0$*,* $\mu_{y3} = 0.5$*,* $\sigma_{y1} = 1$*,* $\sigma_{y2} = 1$*, and* $\sigma_{y3} = 1$*.*

For the first part of the observation sequence (i.e., $t = 1, \ldots, 10^3$), the transition probability matrices are approximately diagonal, which significantly slow down the memory decay of the Markov chain; whereas for the second part (i.e., $t = 1001, \cdots, 10^5$), the transition probability matrices are with approximately identical rows, which leads to much faster convergence speed. Therefore, when the subsequence S is sampled from the first part of the observation sequence, long buffers are required to control the error distances; whereas much shorter buffers are needed when S is from the second part of the observation sequence. We analyze the data sequence using the method proposed by Ye (2018) and our method as developed below, respectively. For demonstration, we set $M = 1$ in this example.

The estimated buffer lengths for both methods are shown in Figure 2.3. It is evident that Ye (2018)'s method is heavily influenced by the first part of the observation sequence

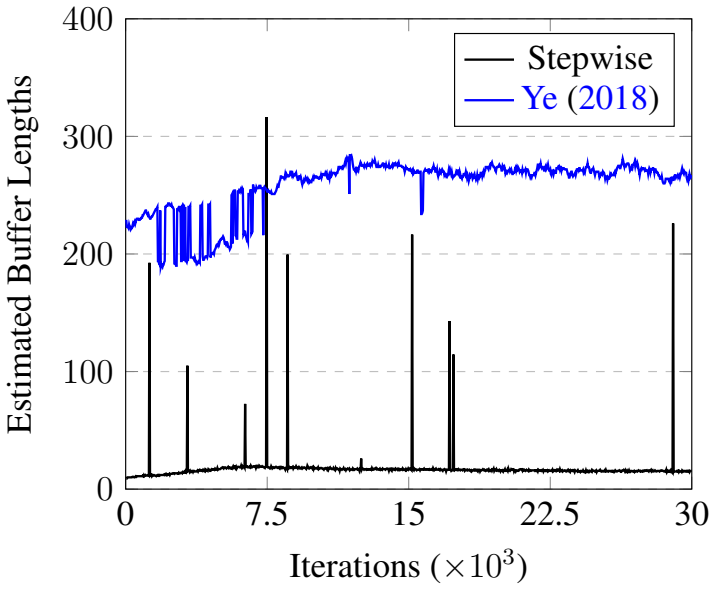


Figure 2.3: Estimated buffer lengths for the example based on the method in Ye (2018) and the proposed local stepwise method in this paper, respectively. The proposed stepwise approach returns required buffer lengths dependent on the local nonhomogeneity of the NHMM, while the approach proposed in Ye (2018) tends to yield buffer lengths longer than needed.

and produces unnecessarily long buffers in many iterations. On the contrary, our approach yields desired adaptive buffer lengths, which demonstrates the necessity of considering local nonhomogeneity of an NHMM.

This example highlights that the nonhomogeneity of an NHMM can significantly affect its local memory decay rates, and thus, the buffer lengths required in the subsampling procedure. Although GLEs characterize the long-run memory decay rate for a nonhomogeneous system, it struggles in capturing finite-time local memory decay patterns because the long-run limit nature of GLEs renders them independent of local dynamics of the system (Abarbanel et al., 1992). A precise measure that quantifies the local memory decay rates is thus required. In this paper, we propose the use of LLEs instead of GLEs to estimate buffer lengths for NHMMs. Specifically, LLEs are defined as follows (Abarbanel et al., 1992):

$$\lambda_{L,n}(t, t+T_0) = \lambda_{L,n,2}(t, t+T_0) - \lambda_{L,n,1}(t, t+T_0) = \frac{1}{T_0} \log \left\| \bar{\boldsymbol{a}}^*_{n,t+T_0} - \bar{\boldsymbol{a}}_{n,t+T_0} \right\|,$$

where $\lambda_{L,n,1}(t, t+T_0)$ and $\lambda_{L,n,2}(t, t+T_0)$ are the first two LLEs of the subsequence from t to $t+T_0$ with initial state distributions being $\bar{\boldsymbol{a}}^*_{n,t}$ and $\bar{\boldsymbol{a}}_{n,t}$. LLEs characterize the finite-time local average memory decay rates of the subsequence from time t to $t+T_0$, and are useful in determining buffer lengths for NHMMs. In fact, LLEs can be viewed as the finite-time version of GLEs on the subsequence from t to $t+T_0$, and they converge to GLEs as $T_0 \to \infty$.

Since LLEs vary when t or T_0 changes, we propose a stepwise method to determine the buffer length, $U_{n,m,1}$. The stepwise method starts from the shortest possible buffer (i.e., $U_{n,m,1} = 1$) adjacent to the sampled subsequence S and evaluate whether the buffer is sufficiently long to control the error distance; if so, then $U_{n,m,1}$ is determined, and if not, we increase the buffer length by one and repeat the above process. The evaluation compares the current $U_{n,m,1}$ and $[\log(\delta)/\lambda_{L,n}(s_0 - U_{n,m,1}, s_0)]$; the buffer is sufficiently long if $U_{n,m,1} \geq [\log(\delta)/\lambda_{L,n}(s_0 - U_{n,m,1}, s_0)]$. The remaining task is to estimate $\lambda_{L,n}(t, t+T_0)$.

Abarbanel et al. (1992) show that $\lambda_{L,n}(t, t+T_0)$ can be estimated by the following formula:

$$\lambda_{L,n}(t, t+T_0) = \frac{1}{T_0} \log \left\| \boldsymbol{J}_{n,t+T_0}(\boldsymbol{v}_{n,t+T_0-1}) \cdots \boldsymbol{J}_{n,t+1}(\boldsymbol{v}_{n,t}) \right\|, \tag{2.24}$$

where $\boldsymbol{v}_{n,t}$ is a $(K-1) \times 1$ vector from a mapping of normalized forward probabilities to an unconstrained real space, and $\boldsymbol{J}_{n,t+1}(\boldsymbol{v}_{n,t})$ is the $(K-1) \times (K-1)$ Jacobian matrix that characterizes the contraction and expansion properties of $\boldsymbol{v}_{n,t+1}$. We follow the idea given in Ye (2018) and treat the forward probabilities as a random dynamical system (RDS) and derive formulas and expressions for $\boldsymbol{v}_{n,t}$ and $\boldsymbol{J}_{n,t+1}(\boldsymbol{v}_{n,t})$. We treat the unnormalized forward probabilities $\{\boldsymbol{a}_{n,t+T_0} = \boldsymbol{M}_{n,t+T_0,t+1}\boldsymbol{a}_{n,t}\}$ as a random dynamical system (RDS); the normalized forward probabilities $\{\bar{\boldsymbol{a}}_{n,t+T_0}\}$ are thus an induced RDS satisfying the

following constraint:

$$\sum_{k=1}^{K} \bar{a}_{n,t+T_0}(k) = 1.$$

We refer readers to Arnold (1998) and Ye (2018) for a detailed introduction of RDS. A transformation can be applied to $\bar{\boldsymbol{a}}_{n,t+T_0}$ as:

$$\begin{aligned}\bar{\boldsymbol{a}}_{n,t+T_0} &= (\bar{a}_{n,t+T_0}(1), ..., \bar{a}_{n,t+T_0}(K))' \\ \rightarrow \boldsymbol{v}_{n,t+T_0} &= \left(\log\left(\frac{\bar{a}_{n,t+T_0}(1)}{\bar{a}_{n,t+T_0}(K)}\right), ..., \log\left(\frac{\bar{a}_{n,t+T_0}(K-1)}{\bar{a}_{n,t+T_0}(K)}\right)\right)', \end{aligned} \tag{2.25}$$

with the inverse mapping:

$$\begin{aligned}\boldsymbol{v}_{n,t+T_0} &= (v_{n,t+T_0}(1), ..., v_{n,t+T_0}(K-1))' \\ \rightarrow \bar{\boldsymbol{a}}_{n,t+T_0} &= \left(\frac{\exp(v_{n,t+T_0}(1))}{\sum_{k=1}^{K-1}\exp(v_{n,t+T_0}(k))+1}, ..., \frac{1}{\sum_{k=1}^{K-1}\exp(v_{n,t+T_0}(k))+1}\right)'. \end{aligned} \tag{2.26}$$

The transformation preserves the Lyapunov spectrum. Here $\boldsymbol{v}_{n,t+T_0}$ contains $K-1$ unconstrained entries and can be written as

$$\boldsymbol{v}_{n,t+T_0} = \boldsymbol{v}_{0,n,t+T_0} + \boldsymbol{F}_{n,t+T_0}(\boldsymbol{v}_{n,t+T_0-1}), \tag{2.27}$$

where

$$\begin{aligned}\boldsymbol{v}_{0,n,t+T_0} = \Bigg(&\log\left(\frac{p(\boldsymbol{y}_{n,t+T_0}|z_{n,t+T_0}=1, \boldsymbol{x}_{n,t+T_0}, \boldsymbol{\beta})}{p(\boldsymbol{y}_{n,t+T_0}|z_{n,t+T_0}=K, \boldsymbol{x}_{n,t+T_0}, \boldsymbol{\beta})}\right), ..., \\ &\log\left(\frac{p(\boldsymbol{y}_{n,t+T_0}|z_{n,t+T_0}=K-1, \boldsymbol{x}_{n,t+T_0}, \boldsymbol{\beta})}{p(\boldsymbol{y}_{n,t+T_0}|z_{n,t+T_0}=K, \boldsymbol{x}_{n,t+T_0}, \boldsymbol{\beta})}\right)\Bigg)', \end{aligned} \tag{2.28}$$

$$\boldsymbol{F}_{n,t+T_0}(\boldsymbol{v}_{n,t+T_0-1})$$
$$= \left(\log\left(\frac{\exp(\boldsymbol{v}_{n,t+T_0-1})'\boldsymbol{Q}_{n,t+T_0,\cdot 1}}{\exp(\boldsymbol{v}_{n,t+T_0-1})'\boldsymbol{Q}_{n,t+T_0,\cdot K}}\right), \cdots, \log\left(\frac{\exp(\boldsymbol{v}_{n,t+T_0-1})'\boldsymbol{Q}_{n,t+T_0,\cdot K-1}}{\exp(\boldsymbol{v}_{n,t+T_0-1})'\boldsymbol{Q}_{n,t+T_0,\cdot K}}\right)\right)', \tag{2.29}$$

and $\boldsymbol{Q}_{n,t+T_0,\cdot k}$ denotes the vector of the k-th column of transition matrix $\boldsymbol{Q}_{n,t+T_0}$. The contraction and expansion properties (as measured by LLEs) of $\{\boldsymbol{v}_{n,t+T_0}\}$ can be characterized by the $(K-1)\times(K-1)$ Jacobian matrix

$$\boldsymbol{J}_{n,t+T_0}(\boldsymbol{v}_{n,t+T_0-1}) = \nabla \boldsymbol{F}_{n,t+T_0}(\boldsymbol{v}_{n,t+T_0-1}), \tag{2.30}$$

where the ij-th element in the matrix is expressed as follows: for $i, j = 1, ..., K-1$

$$J_{n,t+T_0,ij}(\boldsymbol{v}_{n,t+T_0-1}) = \frac{\exp(v_{n,t+T_0-1}(j))q_{n,t+T_0,ji}}{\exp(\boldsymbol{v}_{n,t+T_0-1})'\boldsymbol{Q}_{n,t+T_0,\cdot i}} - \frac{\exp(v_{n,t+T_0-1}(j))q_{n,t+T_0,jK}}{\exp(\boldsymbol{v}_{n,t+T_0-1})'\boldsymbol{Q}_{n,t+T_0,\cdot K}}. \tag{2.31}$$

The Jacobian matrices can be used in estimating $\lambda_{L,n}(t, t+T_0)$ according to Equation (2.24).

For computation, instead of the QR decomposition approach given in Abarbanel et al. (1992), we adapt the efficient method developed in Ye (2018) to estimate $\lambda_{L,n}(t, t+T_0)$ by sequentially multiplying the Jacobian matrices to an initialized unit vector and renormalizing the obtained vector at each step. This estimation scheme facilitates the implementation of the proposed stepwise method; when we increase the buffer length by one, we only need to multiply another Jacobian matrix for the estimation of the new LLE.

The developed SVB can be efficiently performed on NHMMs with ultra-long sequences. At each iteration, N_s subjects are sampled; M subsequences of equal length S are sampled from each of the N_s full sequences; two buffers are determined adaptively according to the proposed local stepwise method for each subsequence; the forward-backward algorithm is

then performed on the buffered subsequences; finally, variational parameters are updated solely based on the subsequences without buffers. We first introduce the detailed procedure of determining the length $U_{n,m,1}$ for the forward buffer, which applies to the determination of the length $U_{n,m,2}$ for the backward buffer. The full procedure of the SVB method is presented subsequently.

The buffer length $U_{n,m,1}$ of a subsequence is determined though the following steps:

Step 1: Suppose the starting point is s_0 for the sampled subsequence of a length S from the subject n. Specify a threshold δ for the error distance, which is set as 10^{-8} in this study. Initiate the buffer length $U_{n,m,1} = 1$. Initiate a unit working vector $\boldsymbol{\varpi}_1$ of a length $K-1$ and a scalar $\varphi_2 = 0$.

The forward buffer starts from $s_0 - 1$ towards the direction of $s_0 - 2$. The following steps are repeated until the termination criterion is reached.

Step 2: Compute $\boldsymbol{v}_{n,s_0-U_{n,m,1}}$ following Equation (2.25) with current values of normalized forward probabilities.

Step 3: Compute $\boldsymbol{J}_{n,s_0-U_{n,m,1}}(\boldsymbol{v}_{n,s_0-U_{n,m,1}+1})$ via Equations (2.30) - (2.31) with entries of $\boldsymbol{E}_{n,s_0-U_{n,m,1}}$, the variational estimate of $\boldsymbol{Q}_{n,s_0-U_{n,m,1}}$.

Step 4: Update the working vector $\boldsymbol{\varpi}_1$ through the following formula:

$$\boldsymbol{\varpi}_1 = \boldsymbol{J}_{n,s_0-U_{n,m,1}}(\boldsymbol{v}_{n,s_0-U_{n,m,1}+1})\boldsymbol{\varpi}_1.$$

Update the scalar ϖ_2 through the following formula:

$$\varpi_2 = \varpi_2 + \log \|\boldsymbol{\varpi}_1\|.$$

Step 5: Compute $\lambda_{L,n}(s_0 - U_{n,m,1}, s_0)$ by averaging ϖ_2 though the following formula:

$$\lambda_{L,n}(s_0 - U_{n,m,1}, s_0) = \frac{\varpi_2}{U_{n,m,1}}.$$

Step 6: Compute $[\log(\delta)/\lambda_{L,n}(s_0 - U_{n,m,1}, s_0)]$. Terminate the algorithm and return $U_{n,m,1}$ if $U_{n,m,1} \geq [\log(\delta)/\lambda_{L,n}(s_0 - U_{n,m,1}, s_0)]$, the buffer has reached the starting point of the entire sequence, or the buffer length $U_{n,m,1}$ is regarded too long (e.g., $U_{n,m,1}$ reaches 500). Otherwise, renormalize $\boldsymbol{\varpi}_1$, set $U_{n,m,1} = U_{n,m,1} + 1$, and repeat steps 2 - 6.

The full procedure of the SVB method (details similar to those in the VB procedure are omitted) is given as follows:

Step 1: Specify the number of iterations for the algorithm, *Iter*. Initialize variational parameters and forward and backward probabilities.

The procedure iteratively performs the following steps until *Iter* is reached.

Step 2: Randomly sample N_s subjects from the total N subjects without replacement.

Step 3: Randomly sample M subsequences of a length S from each of the sampled N_s sequences. Determine buffers for each subsequence using the above procedure.

Step 4: Update forward and backward probabilities and the posteriors for hidden states based on the buffered subsequences.

Step 5: Update variational parameters for the emission and transition models based on the subsequences without buffers.

Step 6: Compute the ELBO value $\mathcal{L}_{N_s}$ using current variational parameters.

2.4 Simulation Studies

We conduct two simulation studies to assess the empirical performance of proposed methods. In Simulation 1, we evaluate the finite sample performance of the developed methods under different sample sizes. We consider three approaches as benchmarks, including an SVB method without buffers, a conventional Bayesian MCMC method, and a conventional frequentist MLE method. In Simulation 2, we further compare our methods to the recently developed Bayesian method for NHMMs via Pólya-Gamma data augmentation (PGMCMC, Holsclaw et al., 2017). The PGMCMC method is specifically designed for NHMMs with unordered hidden states governed by reduced-design transition probability matrices, a special case of our full-design transition model given by Equation (2.3).

2.4.1 Simulation 1

In this simulation, we consider an NHMM defined by Equations (2.1)-(2.3) and (2.5) with $K = 3$ unordered hidden states. Detailed settings of this simulation are given as follows. For the transition model given by Equation (2.3), we set true values of $\boldsymbol{\rho}$ as $\rho_{11,0} = 0.5$, $\rho_{12,0} = -0.5$, $\rho_{21,0} = -0.5$, $\rho_{22,0} = 0.5$, $\rho_{31,0} = -0.5$, $\rho_{32,0} = -0.5$, $\boldsymbol{\rho}_{11} = (0.5)'$, $\boldsymbol{\rho}_{12} = (-0.5)'$, $\boldsymbol{\rho}_{21} = (-0.5)'$, $\boldsymbol{\rho}_{22} = (0.5)'$, $\boldsymbol{\rho}_{31} = (-0.5)'$, and $\boldsymbol{\rho}_{32} = (-0.5)'$. We let $\boldsymbol{w}_{n,t} = (w_{n,t,1})'$, where $w_{n,t,1}$ is generated as:

$$
\begin{aligned}
w^*_{n,t,1} &\sim \mathcal{N}(4\sin(0.002t), 1), \\
w_{n,t,1} &= w^*_{n,t,1}\mathbb{I}(w^*_{n,t,1} \geq 0) + 0.1 w^*_{n,t,1}\mathbb{I}(w^*_{n,t,1} < 0),
\end{aligned}
$$

and $\mathbb{I}(\cdot)$ denotes an indicator function. The local memory decay rates of the considered NHMM are influenced by the covariate. In specific, the Markov chain exhibits a lower memory decay rate as $w_{n,t,1}$ grows more positive. The emission distributions of the

NHMM are taken as $(y_{n,t}|z_{n,t} = k, \boldsymbol{\beta}, \boldsymbol{x}_{n,t}) \sim \mathcal{N}(\boldsymbol{x}'_{n,t}\boldsymbol{\beta}_k, \sigma_k^2)$, for $k = 1, 2, 3$, where $\boldsymbol{x}_{n,t} = (x_{n,t,1}, x_{n,t,2})'$, $x_{n,t,1} = 1$, $x_{n,t,2}$ is generated from a standard normal distribution, and the true values of the parameters in $\boldsymbol{\beta}$ are set as $\boldsymbol{\beta}_1 = (-5, -1)'$, $\boldsymbol{\beta}_2 = (-1, 1)'$, $\boldsymbol{\beta}_3 = (3, 1)'$, and $\sigma_k = 1$ for $k = 1, 2, 3$. We consider 12 different scenarios by setting $N = 10, 50, 100$ and $T = 10, 10^2, 10^3, 10^4$. For each scenario, we generate 100 data sets for replications.

In this simulation, we consider three approaches for comparison. The first approach is an SVB method without buffers. The only difference of the approach with the developed SVB method is that buffers are excluded. This benchmark is used to examine the effectiveness of attaching adaptive buffers to control for bias in the developed SVB method. Two conventional approaches, including a Bayesian MCMC approach and a frequentist MLE method, are also used as baselines. The MCMC approach is commonly used for analyzing NHMMs (e.g., Heaps et al., 2015; Kang et al., 2019; Spezia, 2006); it utilizes Gibbs sampler, Metropolis-Hastings algorithm, and forward-backward algorithm to sample from the exact posteriors. The approach is expected to produce more accurate estimates by generating exact samples from the target posteriors than the proposed variational Bayesian methods, but at the cost of being computationally intensive or even infeasible for massive datasets.

The frequentist MLE approach is another conventional method for NHMMs (e.g., Hughes et al., 1999; Kani et al., 2018). The expectation-maximization algorithm is the standard choice which treats the hidden states as missing data. The E-step uses forward-backward algorithm to determine the forward and backward probabilities, and the M-step performs numerical optimizations on the unknown parameters. In applications with NHMMs, MLE methods are less preferred compared to the Bayesian methods because Bayesian methods can incorporate prior information, directly provide interval estimations for the parameters, and produce reliable results for complex models such as NHMMs. Nonetheless, we consider the MLE method here to thoroughly evaluate the performance of the proposed methods. In total, we consider five methods in this simulation study, that is,

VB, SVB, SVB without buffers, MCMC method, and MLE method.

We assign normal priors for $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ in the four Bayesian methods (i.e., MCMC, VB, SVB, and SVB without buffers) as:

$$p(\boldsymbol{\beta}) \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta,0}), \;\; p(\boldsymbol{\rho}) \sim \mathcal{N}(\boldsymbol{\rho}_0, \boldsymbol{\Sigma}_{\rho,0}),$$

where the hyperparameters are set as $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_{\beta,0} = 10^2\boldsymbol{I}$, $\boldsymbol{\rho}_0 = \mathbf{0}$, and $\boldsymbol{\Sigma}_{\rho,0} = 10^2\boldsymbol{I}$ so that the priors are diffuse. For the VB and SVB methods, we apply $r_{\rho,k_1k_2} = 2$ and $r_{\beta,k} = 3$ factors in variational posteriors for $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$, respectively. Moreover, we sample $N_s = 1$ subject out of the N subjects randomly at each iteration for the VB methods; as mentioned previously, this stochastic consideration reduces computational cost significantly and attains acceptable estimation accuracy. For the SVB methods, the number of subsequences M is set as 10 and the length of each subsequence S is set as 10. Further sensitivity analyses show that varying M and S provides similar estimation results. All five approaches are implemented using Python. Related computer codes are provided as a supplement for this paper. Several test runs show that the MCMC method converges within 5000 iterations; we obtain estimation results for this method using 5000 samples after discarding the first 5000 burn-in iterations. Convergence for MLE, VB, and SVB methods generally occurs within 2000 iterations; we thus use a conservative 10000 iterations for these methods. Each of the five methods is applied to analyze the generated data sets on a computer with an 8 GB memory and a 2.70 GHz CPU; the estimation results are assessed using computational time, bias and root mean square errors (RMSE). Table 2.1 summarizes the results for this simulation. Note that to make the table concise, we report the average RMSE and bias of the parameters in the emission model and transition model, respectively; the detailed bias and RMSE of each specific parameter are available upon request.

The results show that the MCMC approach (first panel in Table 2.1) performs the best in recovering the parameters in both the emission model and transition model with different

sample sizes. The main obstacle for the MCMC approach is the high computational cost; it is evident that the computation time for one replication increases to over 8500 seconds when the total sample size NT reaches 10^4. Further MCMC replications are not completed due to our limited computational resources. The MLE approach (second panel in Table 2.1) performs not as well as MCMC in terms of estimation accuracy; its computational efficiency is also limited. The proposed VB and SVB methods (third and fourth panels in Table 2.1) provide comparable estimation accuracy to that of the MCMC and MLE methods, especially for datasets with relatively long sequences (e.g., $T \geq 10^3$). The efficiency gain for the two proposed methods is significant; the VB method is able to handle datasets with long sequences within a reasonable time frame, and the SVB method attains high efficiency in analyzing datasets with ultra-long sequences. As expected, we note that the SVB method is slightly less accurate than the VB method because SVB only uses several buffered short subsequences sampled from an ultra-long sequence. The significant bias and RMSE associated with the estimations from the method of SVB without buffers (fifth panel in Table 2.1) highlight the effectiveness of using adaptive buffers to control for bias due to subsampling. Overall, the MCMC approach is preferable when N and T are small to moderate, the VB approach is preferable when sequences are long, and the SVB approach is preferable when sequences are ultra-long. Therefore, we apply the VB approach in our first data application to analyze an NHMM with long sequences of children's eye-tracking scan-paths, and use the SVB approach in the second data application to analyze an NHMM with ultra-long sequences of customers' mobile Internet usage records.

Table 2.1: Estimation results in Simulation 1.

		MCMC			MLE			VB			SVB			SVB w/o Buffer		
N	T	BIAS	RMSE	Time*	BIAS	RMSE	Time	BIAS	RMSE	Time	BIAS	RMSE	Time	BIAS	RMSE	Time
Parameters in emission model, β																
10	10^1	0.0849	0.1846	117.79	0.0920	0.2894	48.47	0.2325	0.5063	139.30	-	-	-	-	-	-
	10^2	0.0045	0.1226	991.24	0.0058	0.1228	300.11	-0.0467	0.3842	224.38	-	-	-	-	-	-
	10^3	0.0022	0.0705	8639.88	-0.0015	0.0720	2526.32	0.0032	0.1569	767.75	0.0056	0.1529	1001.21	0.3269	0.4933	245.71
	10^4	-	-	-	-	-	-	0.0046	0.1373	6134.78	0.0012	0.0703	1376.63	0.3345	0.4578	249.30
50	10^1	-0.0093	0.1341	593.94	0.0157	0.1670	195.18	-0.2242	0.4651	140.47	-	-	-	-	-	-
	10^2	0.0026	0.0805	6011.44	0.0036	0.1202	1521.44	-0.0235	0.3161	231.23	-	-	-	-	-	-
	10^3	-	-	-	-	-	-	0.0027	0.1691	774.81	0.0032	0.1066	1004.82	0.3434	0.5184	246.39
	10^4	-	-	-	-	-	-	0.0008	0.1021	6376.39	-0.0005	0.0787	1358.54	0.3529	0.4785	245.45
100	10^1	0.0026	0.0821	1140.92	-0.0020	0.1248	378.25	0.2226	0.4283	162.25	-	-	-	-	-	-
	10^2	-0.0019	0.0577	8540.19	0.0025	0.0830	3080.02	0.0325	0.3109	236.85	-	-	-	-	-	-
	10^3	-	-	-	-	-	-	0.0050	0.1938	790.37	0.0003	0.1033	1033.26	0.3295	0.5725	246.00
	10^4	-	-	-	-	-	-	-0.0007	0.0924	6403.50	0.0072	0.0770	1370.99	0.3516	0.4823	246.62
Parameters in transition model, ρ																
10	10^1	-0.1442	0.3218		-0.2780	0.4110		-0.2985	0.5452		-	-		-	-	
	10^2	-0.0577	0.1807		0.0917	0.1961		0.0921	0.3744		-	-		-	-	
	10^3	0.0069	0.0956		-0.0090	0.1367		-0.0155	0.2605		-0.0891	0.2949		-0.3263	0.5737	
	10^4	-	-		-	-		-0.0061	0.1679		-0.0628	0.1305		-0.2395	0.3796	
50	10^1	-0.0981	0.1695		-0.1649	0.3523		-0.3012	0.4934		-	-		-	-	
	10^2	0.0216	0.1302		-0.0736	0.1824		-0.0506	0.3212		-	-		-	-	
	10^3	-	-		-	-		-0.0100	0.2346		-0.0847	0.2337		-0.3419	0.6036	
	10^4	-	-		-	-		0.0086	0.1141		-0.0574	0.1138		-0.2482	0.4126	
100	10^1	-0.0607	0.1275		-0.0645	0.1791		-0.2995	0.5263		-	-		-	-	
	10^2	-0.0192	0.0830		0.0122	0.1217		-0.0818	0.2969		-	-		-	-	
	10^3	-	-		-	-		-0.0081	0.2437		-0.0821	0.2307		-0.3232	0.5766	
	10^4	-	-		-	-		0.0077	0.1116		-0.0544	0.1032		-0.2428	0.4023	

*Average computation time in seconds for one replication.

2.4.2 Simulation 2

In this simulation, we compare the proposed methods to the PGMCMC approach in Holsclaw et al. (2017). The PGMCMC approach utilizes the method of Pólya-Gamma data augmentation to induce conjugacy for parameters in the transition model so that an efficient and closed-form Gibbs sampler can be obtained. The closed-form Gibbs sampler enables the PGMCMC approach to handle NHMMs with long sequences of a length up to 10^4. Here we use this recently developed method as another baseline.

The PGMCMC approach is developed for NHMMs with $N = 1$ sequence of multivariate temporal process $\boldsymbol{y}_{\mathcal{T}}$. The corresponding emission model is defined separately on each element of $\boldsymbol{y}_t$ from y_{t1} to y_{tR_y}; and $\boldsymbol{y}_t$ forms a multivariate emission model similarly to the one given in Equation (2.5). The corresponding transition model, on the other hand, is in a reduced form given as: for $k_1, k_2 = 1, \ldots, K$,

$$q_{t,k_1k_2} = \frac{\exp(\rho_{k_1k_2,0} + \boldsymbol{w}_t'\boldsymbol{\rho}_{k_2})}{\sum_{j=1}^{K} \exp(\rho_{k_1j,0} + \boldsymbol{w}_t'\boldsymbol{\rho}_j)},$$

where $\rho_{kK,0}$ and $\boldsymbol{\rho}_{kK}$ are set as $\mathbf{0}$ for all k for identification purposes. This reduced form model has one set of regression coefficients for the probability of entering each state k_2 and is a special case of our full-design transition model given in Equation (2.3).

In this simulation study, we consider an NHMM as described above with $K = 3$ hidden states. For the transition model, we have $\boldsymbol{w}_t = (w_{t,1}, w_{t,2})'$, where $w_{t,1}$ is generated from a standard normal distribution and $w_{t,2}$ is generated from a Bernoulli distribution with a probability of 0.5 to be 1. The true values of parameters are set as $\rho_{11,0} = 2$, $\xi_{12,0} = 2$, $\xi_{21,0} = -2$, $\xi_{22,0} = 2$, $\xi_{31} = -2$, $\xi_{32} = -2$, $\boldsymbol{\rho}_1 = (-1, 1)'$, and $\boldsymbol{\rho}_2 = (0.5, -1)'$. For the emission model, we consider two cases. The first case is a Gaussian emission model with

the dimensionality of $\boldsymbol{y}_t$ being $R_y = 2$; the model is given as follows:

$$(y_{tj}|z_t = k, \boldsymbol{\beta}) \sim \mathcal{N}(\mu_{jk}, \sigma^2_{jk}),$$

where the true values of parameters in $\boldsymbol{\beta}$ are set as $\mu_{11} = -5$, $\mu_{12} = -1$, $\mu_{13} = 3$, $\mu_{21} = -4$, $\mu_{22} = 0$, $\mu_{23} = 4$, and $\sigma_{jk} = 1$ for $j = 1, 2, k = 1, 2, 3$. The second case is an exponential emission model with $R_y = 2$; the model is given as follows:

$$(y_{tj}|z_t = k, \boldsymbol{\beta}) \sim Exp(\mu_{jk}),$$

where $Exp(\cdot)$ denotes an exponential distribution and the true values of parameters in $\boldsymbol{\beta}$ are set as $\mu_{11} = 2.7$, $\mu_{12} = 1$, $\mu_{13} = 0.4$, $\mu_{21} = 4.5$, $\mu_{22} = 1.6$, and $\mu_{23} = 0.6$. We consider 4 different scenarios with $T = 5 \times 10^3, 10^4, 5 \times 10^4, 10^5$ for evaluation. For each scenario, we generate 100 data sets for replications.

Similar diffuse priors as in Simulation 1 are used. The settings and implementations of VB and SVB methods are also similar to those described in Simulation 1. The PGMCMC approach is implemented using the R-package `NHMM` provided by Holsclaw et al. (2017), and the default settings therein are used in this study. Several test runs show that the PGMCMC approach converges within 5000 iterations. We therefore use a burn-in phase of 5000 iterations and obtain the estimation results based on another 5000 iterations. In simulation 2, we use 10000 iterations for the VB and SVB methods. Table 2.2 summarizes results for this simulation.

The results show that the proposed VB and SVB methods provide comparable estimation accuracy and superior computational efficiency compared to PGMCMC. The computational gain of SVB is again significant with slight sacrifice of estimation accuracy. This simulation study further confirms the satisfactory performance of the proposed methods.

Table 2.2: Estimation results in Simulation 2.

	PGMCMC			VB			SVB		
T	BIAS	RMSE	Time*	BIAS	RMSE	Time*	BIAS	RMSE	Time*
Gaussian emission case, parameters β									
5×10^3	0.0283	0.1630	8653.22	0.0029	0.0110	2407.13	0.0035	0.0110	433.02
10^4	0.0145	0.0260	34178.44	0.0014	0.0073	5140.68	0.0042	0.0087	473.76
5×10^4	-	-	-	-	-	-	0.0009	0.0066	511.56
10^5	-	-	-	-	-	-	0.0012	0.0054	549.78
Gaussian emission case, parameters ρ									
5×10^3	0.0515	0.0932		0.0165	0.0533		0.0180	0.0670	
10^4	0.0427	0.0775		0.0086	0.0380		0.0116	0.0405	
5×10^4	-	-	-	-	-	-	0.0073	0.0251	
10^5	-	-	-	-	-	-	0.0038	0.0182	
Exponential emission case, parameters β									
5×10^3	0.0637	0.1481	8511.60	0.0133	0.0191	3024.41	0.0128	0.0545	431.34
10^4	0.0215	0.0797	32832.20	0.0071	0.0103	5943.06	0.0038	0.0135	441.84
5×10^4	-	-	-	-	-	-	0.0040	0.0092	432.18
10^5	-	-	-	-	-	-	0.0030	0.0075	446.46
Exponential emission case, parameters ρ									
5×10^3	0.0507	0.0832		0.0537	0.1039		0.0798	0.1487	
10^4	0.0397	0.0671		0.0409	0.0706		0.0688	0.1017	
5×10^4	-	-	-	-	-	-	0.0535	0.0686	
10^5	-	-	-	-	-	-	0.0372	0.0461	

*Average computation time in seconds for one replication.

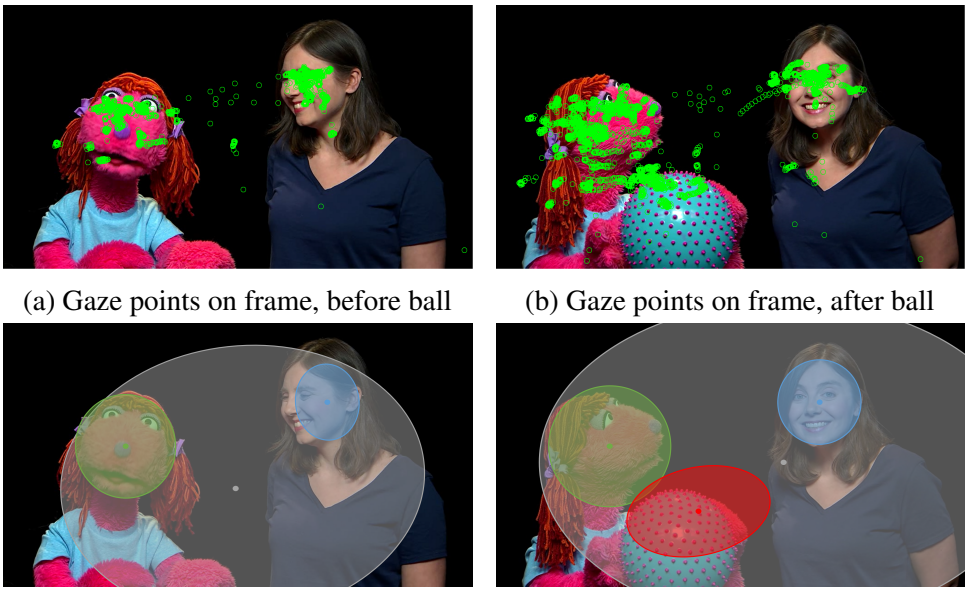
2.5 Analysis of Eye-tracking Scan-path Data

In this section, we present an analysis of the first motivating example. This analysis used an NHMM to delineate how children with ASD watch social-communicative scenes when interacting partners in the scenes vary by social salience (i.e., puppet v.s. person) and to reveal how HSCs (e.g., speech cue) affect selective social attention of the children. We applied the proposed VB method to perform statistical inference for the NHMM.

2.5.1 Scan-path Data

A group of ASD children ($N = 39$; Median age = 49.44 months) were eye-tracked while they were watching a designed video clip. The featured video is 86-second-long, depicting two interacting partners, a puppet and an actress (i.e., a person), engaging in a playful conversation. In the video, the puppet and the person took turns speaking and playing with a ball; the ball was introduced at around the 23-second mark of the video, segmenting the video into two parts naturally. Typical frames from the video clip are shown in Figure 2.4. The setting of this study is suitable to examine the relative social salience of the puppet and the person based on the following considerations. First, of a similar size to the actress in the scene, the puppet mimicked human behaviors with moving limbs, head and mouth, stable eyes, and a female voice performed by another actress, ensuring the puppet is human-like. Second, two partners in the video separated the left and right halves of the scene without any overlap and their locations remained largely unchanged, facilitating the detection and interpretation of ROIs. Third, introduction of the ball enabled us to evaluate the distracting effect of the ball on children's attention to the puppet and the person. Finally, HSCs such as speech and face direction of the puppet and the person were clearly present in the video, allowing us to examine the impacts of HSCs on children's attention shift across ROIs. The study was implemented on an eye-tracking platform. The video was displayed on a computer monitor with the resolution of 1680×1050 pixels that comprised 43.1 degrees of visual angle. Coordinates of each child's gaze points were recorded by an *SR EyeLink 1000 Plus 500 Hz* eye-tracker at a sampling rate of 30 Hz (the sampling rate of the eye-tracker matches with the frame rate of the video), which results in an observation scan-path of a length $T \approx 2.6 \times 10^3$ (see Figure 2.4 for gaze points of a child).

We pooled the gaze points of the entire group together and plotted them in Figure 2.5, where the monitor frame area is ($0 \leq x \leq 1680$, $0 \leq y \leq 1050$). The distribution of gaze points is dispersed with no obvious clustering patterns and a considerable amount of gaze



(a) Gaze points on frame, before ball

(b) Gaze points on frame, after ball

(c) ROIs on frame, before ball

(d) ROIs on frame, after ball

Figure 2.4: Typical frames of the video in the eye-tracking study for children with autism spectrum disorders (ASD). (a) a frame from before the ball is introduced; (b) a frame from after the ball is introduced; the gaze points are from a randomly selected ASD child throughout the video clip. (c) and (d) include data-driven ROIs (i.e., hidden states) given by the NHMM modeling. Except for the first hidden state of a point at (0,1050), the second, third, fourth, and fifth ROIs are denoted in grey, blue, green, and red colors, which can be interpreted as background, person face, puppet face, and ball, respectively.

points are located outside the monitor frame area. A gaze point outside the monitor frame area suggests the eye-tracked participant not paying attention to the video content at the moment. One technical issue of the eye-tracker is that it is only able to track the precise location of an outside-of-frame gaze point when the point is not far from the frame; if the gaze point is too far from the frame, the default program of the eye-tracker automatically uses the coordinate of upper left corner of the frame, $(0, 1050)$, as a surrogate (account for 16.2% of all gaze points in our sample). Such surrogate observations were classified into a separate state in NHMM, which should draw our special attention because such observations indicate that children were highly distracted at the moment.

HSCs in the video are mainly speech and face directions of the two partners. Speech was present for 93% of the video clip, with remaining time of the video filled with naturally

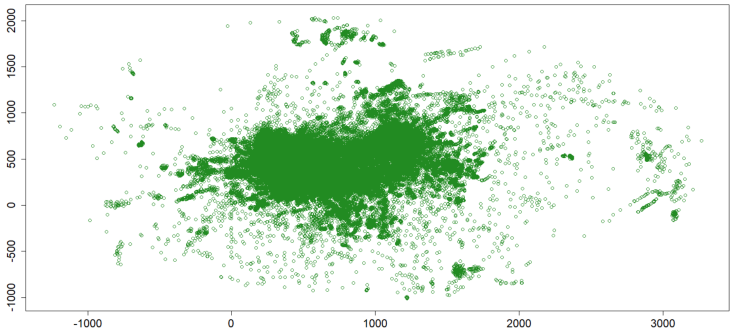


Figure 2.5: Pooled gaze points of all participants in the eye-tracking study.

occurring, transitional silences between conversation partners. The speech of the puppet and the actress accounted for 64% and 36% of the overall speech in the video, respectively; the two partners took turns and did not speak simultaneously. As for the face directions, before the ball was introduced, the partners either spoke to the camera or looked at each other; they also looked at the ball after the ball was introduced. The HSCs in the video were coded frame-by-frame according to the two partners' speech and face direction to form time-varying covariates, which may influence children's attention shift between ROIs. In total, six binary indicators variables (given in Table 2.3) were defined to specify combinations of puppet speech, person speech, puppet face direction, and person face direction, where no speech and person and puppet looking at camera (i.e., looking at the participants) are set as reference categories.

Specifically, recall that $\boldsymbol{w}_t$ denotes the covariate vector in the transition model at time t, then $\boldsymbol{w}_t = (\texttt{Ppt\_Spk}_t, \texttt{Per\_Spk}_t, \texttt{Ppt\_Per}_t, \texttt{Ppt\_Bal}_t, \texttt{Per\_Ppt}_t, \texttt{Per\_Bal}_t)'$, where $\texttt{Ppt\_Spk}_t$, $\texttt{Per\_Spk}_t$, $\texttt{Ppt\_Per}_t$, $\texttt{Ppt\_Bal}_t$, $\texttt{Per\_Ppt}_t$, and $\texttt{Per\_Bal}_t$ are binary indicators taking the value of 1 when the corresponding scenario occurs in frame t.

2.5.2 Model Specification and Inference

In this study, we proposed an NHMM with an emission model that uses a bivariate normal density to describe the distribution of the coordinates of each gaze point conditional on

Table 2.3: Codings of higher saliency cues in the video of the eye-tracking study for children with ASD.

Factor	Value	Variable	Interpretation
Speech	0	Baseline	No Speech
	1	$\texttt{Ppt\_Spk}_t$	Puppet Speaking
	2	$\texttt{Per\_Spk}_t$	Person Speaking
Puppet Face Direction	0	Baseline	Puppet Looking at Camera
	1	$\texttt{Ppt\_Per}_t$	Puppet Looking at Person
	2	$\texttt{Ppt\_Bal}_t$	Puppet Looking at Ball
Person Face Direction	0	Baseline	Person Looking at Camera
	1	$\texttt{Per\_Ppt}_t$	Person Looking at Puppet
	2	$\texttt{Per\_Bal}_t$	Person Looking at Ball

the hidden state and a transition model that considers the impact of HSCs on transition probabilities between hidden states. Since we directly modeled children's eye movements, the hidden states in our model are associated with the ROIs on which children's attention is focused. ROIs can be interpreted depending on their specific locations and do not follow a rank order. The transition between hidden states thus reflects children's attention shift across ROIs; and the nonhomogeneous setting in the transition model examines the effects of HSCs on children's attention shift.

The proposed NHMM is defined by Equations (2.1) - (2.3) and (2.5) with the number of hidden states K to be determined. The specific emission model is given as: for $n = 1, \ldots, N$,

$$(\boldsymbol{y}_{nt}|z_{nt} = 1) \sim \mathbb{I}(\boldsymbol{y}_{nt} = \boldsymbol{\mu}_1),$$
$$(\boldsymbol{y}_{nt}|z_{nt} = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \text{for } k = 2, ..., K,$$

where the covariance matrices can be represented as

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_{k1}^2 & \alpha_k \sigma_{k1} \sigma_{k2} \\ \alpha_k \sigma_{k1} \sigma_{k2} & \sigma_{k2}^2 \end{pmatrix}.$$

The mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$ in the emission model characterize the center and the spread of the corresponding state. As mentioned above, the emission distribution of the first state is assumed to be a constant distribution with a point mass at the coordinate $(0, 1050)$ to model the surrogates for the gaze points far away from the frame. This separate state indicates that children were highly distracted at the moment. The multinomial logit model given in Equation (2.3) was adopted for modeling the transition probabilities.

One feature of the designed video is that the ball was introduced at around the 23-second mark of the video. The ball itself is likely to represent an ROI, indicating the inclusion of the ball may change the number of the hidden states. We therefore split each scan-path into two segments according to the inclusion of the ball and applied the proposed NHMM twice to analyze the scan-paths recorded before the ball was introduced ($T \approx 0.8 \times 10^3$) and after the ball was introduced ($T \approx 1.8 \times 10^3$), respectively. The obtained two sets of results may provide clues to the distracting effect of the ball on children's attention. Note that before the ball was introduced, the covariate vector in the transition model should be $\boldsymbol{w}_t = (\texttt{Ppt\_Spk}_t, \texttt{Per\_Spk}_t, \texttt{Ppt\_Per}_t, \texttt{Per\_Ppt}_t)'$.

The developed VB method was utilized to perform statistical inference on the NHMM. Prior distributions similar to that used in the simulation study were used for the Bayesian inference. We applied $r_{\rho,k_1k_2} = 5$ ($r_{\rho,k_1k_2} = 7$ for the analysis of scan-paths after the ball is introduced) and $r_{\beta,k} = 5$ factors in variational posteriors for $\boldsymbol{\rho}$ and $\boldsymbol{\beta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, respectively. Note that we used full covariance matrices because the dimensionality in this scenario is relatively low. Researchers can use substantially smaller number of factors

compared to parameter dimensions when facing high-dimensional situations. We sampled $N_s = 1$ subject out of the $N = 39$ subjects randomly at each iteration; as demonstrated in the simulation study, this stochastic method reduces computational cost significantly and attains satisfactory estimation accuracy. To obtain appropriate initial values of parameters, we followed the suggestion in Song et al. (2017) by conducting a preliminary analysis using the MCMC method on a small proportion of the data. Test runs showed that the VB method converges within 2000 iterations. We terminated the VB algorithm after 5000 iterations (see Figure 2.8 in the Appendix for the plot of the corresponding ELBO values). The number of hidden states, K, was determined using the widely applicable information criterion (WAIC, Watanabe, 2010). We varied K from 2 to 8 and computed WAIC for each candidate model, with results shown in Table 2.4.

Table 2.4: WAIC values of the NHMM model with different number K of hidden states in the analysis of eye-tracking scan-path data. "Before Ball" and "After Ball" denote the scenarios of before and after the inclusion of the ball in the video, respectively.

K	Before Ball	After Ball
2	9165.66	25463.88
3	8729.95	21021.13
4	**8179.22**	20940.01
5	8826.18	**19860.86**
6	8853.79	20232.38
7	8984.94	20254.24
8	9487.99	22597.82

2.5.3 Results

We summarized results in this section. NHMMs with $K = 4$ and $K = 5$ hidden states exhibit the best fit to the scan-paths before and after the introduction of the ball, respectively (shown in Table 2.4). This suggests that the introduction of the ball increases the number of ROIs by one and the ball itself represents an ROI to the children. The estimation results for parameters in the emission model are given in Table 2.5.

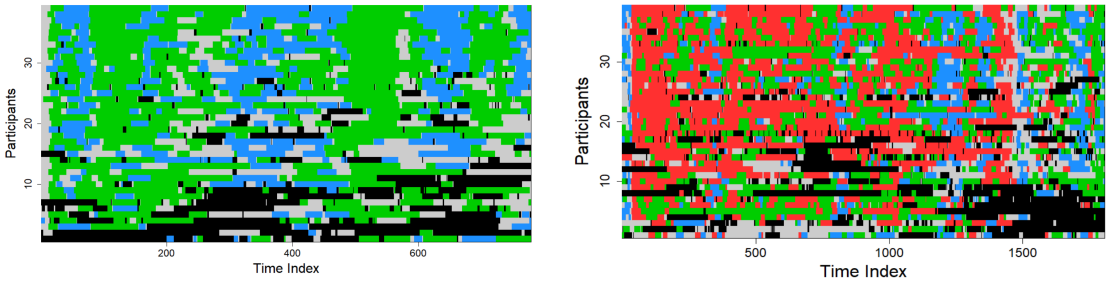
Table 2.5: Bayesian estimates (posterior standard deviations in the parentheses) for the parameters in the emission model in the analysis of eye-tracking scan-path data. Parameters in $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ indicate the center locations and area spreads of ROIs, respectively. "Before Ball" and "After Ball" denote the scenarios of before and after the inclusion of the ball in the video, respectively.

States (k)	Parameters	Before Ball	After Ball	Parameters	Before Ball	After Ball
Background (2)	μ_{21}	792.51 (1.4774)	1016.57 (1.4650)	σ_{21}	397.99 (1.1922)	634.18 (1.0673)
	μ_{22}	312.01 (1.8068)	448.85 (1.1941)	σ_{22}	314.53 (1.0717)	446.63 (1.2474)
				α_2	0.2216 (0.0134)	0.0895 (0.0073)
Person Face (3)	μ_{31}	1129.23 (0.5139)	1129.24 (0.5756)	σ_{31}	59.08 (0.3414)	77.19 (0.5002)
	μ_{32}	699.26 (0.4903)	691.03 (0.9804)	σ_{32}	64.72 (0.2812)	77.40 (0.7768)
				α_3	0.1182 (0.0421)	0.1714 (0.0125)
Puppet Face (4)	μ_{41}	447.44 (1.2733)	377.04 (0.5760)	σ_{41}	91.70 (0.4293)	108.78 (0.2411)
	μ_{42}	551.70 (1.8640)	527.54 (1.1750)	σ_{42}	99.18 (0.9299)	105.35 (0.3937)
				α_4	-0.0043 (0.0104)	0.0513 (0.0256)
Ball (5)	μ_{51}	-	619.53 (0.7351)	σ_{51}	-	132.13 (0.8658)
	μ_{52}	-	274.61 (1.2062)	σ_{52}	-	80.53 (0.3793)
				α_5	-	0.2529 (0.0196)

The estimated ROIs are depicted in Figure 2.4. By comparing the two sets of results in Table 2.5 and Figure 2.4, we noted that ROIs display larger spread after the inclusion of the ball. Larger ROI spreads might indicate less concentrated attention on related objects, providing evidence for the distracting effect of the ball on children's attention. According to the locations and spread areas of the ROIs, they were interpreted as background, person face, puppet face, and ball (see Figure 2.4). The ROIs of person face and puppet face indicate children's attention to the two partners in the video and are the main interest of the

current study.

The data-driven ROIs detected through the NHMM offer an alternative to the conventional method of pre-defining ROIs manually. For example, previous research generally isolates the body parts of partners in the video as separate ROIs (e.g., Chawarska et al., 2012; Shic et al., 2019), which seems not be supported by the current analysis. NHMM classified the body parts into the background region according to signals of children's scan-paths, suggesting the body parts of the partners may not be salient objects in the scene.



(a) Hidden states sequences, Before Ball (b) Hidden states sequences, After Ball

Figure 2.6: Estimated hidden state sequences for children with ASD in the eye-tracking study. The five hidden states of a point at (0,1050), background, person face, puppet face, and ball, are denoted in black, grey, blue, green, and red colors.

The estimated state sequences $z_{n,\mathcal{T}}$, shown in Figure 2.6, indicate children's attention shifts across ROIs and provide fruitful information on children's eye movement patterns, which help reveal the relative salience of the two partners. Specifically, the attention to the puppet ROI exhibits a higher degree of consistency among the children than that of the attention to the person ROI. This phenomenon is particularly noticeable in the first subfigure of Figure 2.6. The column-wise consistency of the green hidden states indicates that the puppet ROI may draw systematic attention from the children, which is not the case for the person ROI; this provides evidence of the higher saliency of the puppet relative to the person.

Parameters in the transition model reflect the effects of HSCs on children's attention

shifts across ROIs. For example, the influences of speech cues on children's attention to puppet face can be obtained from the coefficients related to the transition probabilities from other ROIs to the ROI of puppet face. The full estimation results for parameters in the transition model are given in the Appendix (Tables 2.10 - 2.11).

Table 2.6: Bayesian estimates (posterior standard deviations in the parentheses) for the parameters in the transition model that reflect the effects of speech cues on attention shift to the two partners in the video.

Stimuli	From*	To*	Before Ball	After Ball	Stimuli	From*	To*	Before Ball	After Ball
Person Speech	1	3	-0.6570 (0.6839)	-0.8066 (1.4790)	Puppet Speech	1	4	-0.2497 (0.4738)	-0.6349 (1.0550)
	2	3	0.9172 (0.5046)	0.7941 (1.1395)		2	4	0.8417 (0.5259)	0.1981 (0.8562)
	3	3	0.9697** (0.3440)	1.0117 (0.6536)		3	4	0.9048** (0.2642)	0.3469 (1.0873)
	4	3	0.7766 (0.5225)	0.1424 (0.8967)		4	4	1.0180** (0.3164)	1.0971** (0.3941)
	5	3	-	0.0043 (0.5992)		5	4	-	0.1989 (0.4754)

* State 1: $(0, 1050)$, 2: Background, 3: Person Face, 4: Puppet Face, 5: Ball.
** Zero is not contained in the 95% credibility interval.

The estimated parameters that reflect the effects of speech cues on attention shifts to the two partners are summarized in Table 2.6. The results show that speech cues may increase the saliency of the speaker and draw children's attention to the speaker. One advantage of the NHMM modeling is that it reveals the mechanisms of how the speech cues help draw children's attention. Specifically, before the introduction of the ball, person speech helps to maintain a child's attention to the person, conditional on that the child is paying attention to the person; puppet speech not only helps to maintain a child's attention to the puppet, conditional on that the child is paying attention to the puppet, but also increases the likelihood that the child shifts attention to the puppet, conditional on that the child is paying attention to the person. After the inclusion of the ball, only puppet speech is effective in maintaining a child's attention to the puppet, conditional on that the child is paying attention to the puppet. These results demonstrate the higher saliency of the puppet relative to the person with effects of speech cues and that children's attention to the puppet might be less distracted by the ball.

The above findings were further confirmed by the estimated transition matrices in different scenarios of speech cues given in Table 2.7. We found that the puppet seemed to

Table 2.7: Transition matrices in different scenarios of speech cues in the analysis of eye-tracking scan-path data.

Case 1: Before Ball

	Scenario 1: No Speech				Scenario 2: Person Speech Only			
From\To*	1	2	3	4	1	2	3	4
1	0.9118	0.0367	0.0228	0.0287	0.9396	0.0325	0.0119	0.0160
2	0.0739	0.8849	0.0185	0.0227	0.0546	0.8826	0.0352	0.0275
3	0.0498	0.0152	0.9225	0.0125	0.0201	0.0108	0.9628	0.0063
4	0.0465	0.0103	0.0084	0.9348	0.0212	0.0090	0.0092	0.9606

	Scenario 3: Puppet Speech Only			
From\To	1	2	3	4
1	0.9409	0.0251	0.0110	0.0231
2	0.0416	0.9190	0.0098	0.0296
3	0.0330	0.0229	0.9237	0.0204
4	0.0175	0.0066	0.0015	0.9743

Case 2: After Ball

	Scenario 1: No Speech					Scenario 2: Person Speech Only				
From\To	1	2	3	4	5	1	2	3	4	5
1	0.9005	0.0321	0.0155	0.0231	0.0288	0.9380	0.0257	0.0071	0.0148	0.0144
2	0.2908	0.6067	0.0301	0.0351	0.0372	0.1428	0.8078	0.0252	0.0091	0.0151
3	0.1496	0.0201	0.7891	0.0211	0.0201	0.0576	0.0074	0.9241	0.0043	0.0065
4	0.1470	0.0137	0.0156	0.7960	0.0277	0.0664	0.0052	0.0080	0.9074	0.0130
5	0.1593	0.0188	0.0210	0.0233	0.7776	0.0807	0.0090	0.0107	0.0100	0.8895

	Scenario 3: Puppet Speech Only				
From\To	1	2	3	4	5
1	0.9392	0.0244	0.0059	0.0128	0.0177
2	0.1687	0.7688	0.0107	0.0249	0.0269
3	0.0688	0.0109	0.8971	0.0137	0.0095
4	0.0567	0.0042	0.0040	0.9196	0.0154
5	0.0859	0.0082	0.0104	0.0153	0.8802

* State 1: $(0, 1050)$, 2: background, 3: person face, 4: puppet face, 5: ball.
Note: Reference categories that person and puppet looking at camera are used here.

be more salient than the person in most scenarios of speech cues (scenarios 1 and 3 in both cases in Table 2.7). Specifically, children were more likely to shift attention from the ROIs of the point $(0, 1050)$ and background to the puppet ROI than to the person ROI. They were also less likely to shift attention to other ROIs from the puppet ROI than from the person

ROI. Moreover, even in the scenario that only the person was delivering speech (scenario 2 in both cases in Table 2.7), the puppet ROI was more likely to draw children's attention from the point $(0, 1050)$ ROI than the person ROI.

In this analysis, the VB analysis of NHMM determined data-driven ROIs, revealed the state sequences associated with children's eye movements, depicted children's attention shift between ROIs, evaluated the impacts of HSCs on children's attention shifts, and explored the transition matrices under different scenarios of speech cues. These modeling achievements are difficult to obtain using the conventional methods in autism research. Our obtained results provide evidence on the higher saliency of the puppet relative to the person in the designed social-communicative scenes, which may help to understand the mechanism behind the advantageous performance of humanoid representations in teaching children with ASD and should be considered in the development of intervention schemes for ASD.

2.6 Analysis of Mobile Internet Usage Data

This section presents an analysis of the second motivating example. This analysis proposed an NHMM to model customers' mobile Internet usage behaviors, the underlying latent needs for mobile Internet, and the influences of customers' conventional telecom behaviors on the latent needs. The SVB method was used to analyze the ultra-long sequences of customers' telecom records. We also demonstrated the capability of the utilized NHMM framework in forecasting of customers' future mobile Internet usage, based on which companies can adjust their CRM strategy and achieve better mobile Internet capacity planning.

2.6.1 Mobile Internet Usage Data

The mobile Internet usage dataset contains records of whether each of the 82 customers made/received calls, sent/received texts, and used mobile Internet at a frequency of every

five minutes for 10 months from September 2013 to June 2014, resulting in $N = 82$ observation sequences of a length $T = 87552$. The dataset also contains price of mobile Internet data at each time period. Records of whether customers used mobile Internet $y_{n,t}$ are considered as the binary dependent variable, where $y_{n,t} = 1$ if customer n used mobile Internet at time t. Possible covariates in the NHMM are:

$\texttt{Int\_Chg}_t \geq 0$ denotes the price of mobile Internet data at time t;

$\texttt{Call}_{n,t} \in \{0, 1\}$, $\texttt{Call}_{n,t} = 1$ if customer n made or received phone calls at time t;

$\texttt{SMS}_{n,t} \in \{0, 1\}$, $\texttt{SMS}_{n,t} = 1$ if customer n sent or received text messages at time t.

We expected the price of mobile Internet data at each period may directly affect customers' mobile Internet usage at that period, and customers' conventional telecom behaviors such as calls and texts may affect their latent needs for mobile Internet which govern customers' decisions of using mobile Internet. The instant effect of the price of mobile Internet data is intuitive, as customers tend to be price-sensitive. Customers' conventional telecom behaviors are assumed to affect their states of latent needs for mobile Internet because the effects tend to be in the longer term considering that the observation time window is short in this study. The communication behaviors are likely to create a regime shift in a customer's mobile Internet use pattern by transitioning the customer to a different state of latent need for mobile Internet. Preliminary analysis suggests that conventional telecom behaviors may be associated with customers' higher latent demands for mobile Internet. We found that if a customer made or received calls or texts in the previous period, there was a increase of the likelihood of using mobile Internet data in the period after (shown in see Figure 2.7).

2.6.2 Model Specification and Inference

We considered a K-state NHMM defined by Equations (2.1)-(2.2) and (2.4)-(2.5) with an emission model of a logistic regression to describe the probability of using mobile Internet at each period and examine the instant effect of mobile Internet data price on the probability

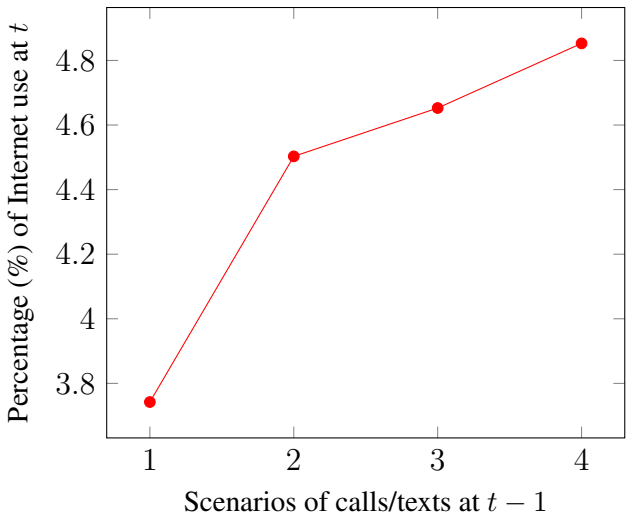


Figure 2.7: Customers' mobile Internet usage under different scenarios of calls/texts in the previous period. Scenarios 1: No call/text message in the previous period; 2: Only made/received calls; 3: Only sent/received texts; and 4: Both calls & texts.

conditional on the current hidden state. The specific form of the emission model is: for $n = 1, \ldots, N$, $k = 1, \ldots, K$ and $t = 1, \ldots, T$,

$$\mathbb{P}(y_{n,t} = 1 | z_{n,t} = k) = \frac{\exp(\beta_{k0} + \beta_{k1} \texttt{Int\_Chg}_t)}{1 + \exp(\beta_{k0} + \beta_{k1} \texttt{Int\_Chg}_t)}, \tag{2.32}$$

where β_{k1} captures the instant effect of mobile Internet data price. Since the hidden states govern customers' decisions of using mobile Internet, these states can be interpreted as latent needs for mobile Internet and have a natural order from weak to strong. Therefore, the transition model with continuation-ratio logits given in Equation (2.4) was used and the nonhomogeneous setting therein revealed the impacts of conventional telecom behaviors on customers' states of latent needs for mobile Internet. The specific form of the transition model is given as follows: for $n = 1, \ldots, N$, $k_1 = 1, \ldots, K$, $k_2 = 1, \ldots, K-1$ and

$t = 2, \ldots, T$,

$$\log\left(\frac{\mathbb{P}(z_{n,t} = k_2 | z_{n,t-1} = k_1)}{\mathbb{P}(z_{n,t} > k_2 | z_{n,t-1} = k_1)}\right) = \log\left(\frac{q_{n,t,k_1k_2}}{q_{n,t,k_1k_2+1} + \cdots + q_{n,t,k_1K}}\right)$$
$$= \rho_{k_1k_2,0} + \rho_{k_1,1}\texttt{Call}_{n,t-1} + \rho_{k_1,2}\texttt{SMS}_{n,t-1}.$$

The proposed SVB method was utilized to conduct statistical inference on the proposed NHMM. The prior distributions similar to that used in the simulation study were adopted. We applied $r_{\rho,k_1k_2} = 3$ and $r_{\beta,k} = 2$ factors in variational posteriors for $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$, respectively. We sampled $N_s = 1$ subject out of the $N = 82$ customers and $M = 10$ subsequences of a length $S = 20$ from the full sequence randomly at each iteration. We analyzed a small proportion of the data to obtain good initial values for parameters estimation (Song et al., 2017). Test runs showed that the SVB method converged within 4000 iterations. The algorithm was thus terminated after a conservative 10000 iterations (see Figure 2.9 in the Appendix for the plot of ELBO values). The number of hidden states, K, was determined through WAIC. We varied K from 1 to 5 and computed WAIC for each candidate model.

2.6.3 Results

The obtained results are summarized in this section. The WAIC values for candidate models are 17187.03, 16157.54, 15669.34, 15809.98 and 15884.97 (K = 1 to 5). The NHMM with $K = 3$ hidden states provides the best overall fit to data. The average probabilities of using mobile Internet conditional on hidden states 1 to 3 were calculated as 0.0315, 0.1037 and 0.2092, suggesting the states can be interpreted as the states of weak, moderate, and strong latent needs for mobile Internet, respectively.

The estimates of parameters are given in Table 2.8. For the emission model, the effects of mobile Internet data price on mobile Internet use are negative, which is consistent with the fact that customers are generally sensitive to price. Interestingly, our NHMM modeling

Table 2.8: Bayesian estimates (posterior standard deviations in the parentheses) for the parameters in the NHMM in the analysis of mobile Internet usage data.

	Covariates	Parameters	States $(k)^*$		
			1	2	3
Emission	-	β_{k0}	-3.4047^{**} (0.0074)	-2.4706^{**} (0.0163)	-1.8920^{**} (0.0083)
	Int_Chg	β_{k1}	-3.3272^{**} (1.5742)	-2.3374^{**} (0.6752)	-1.8970 (2.2948)
Transition	-	$\rho_{k1,0}$	3.7546^{**} (0.0029)	1.3522^{**} (0.0152)	-4.3117^{**} (0.0101)
	-	$\rho_{k2,0}$	2.7517^{**} (0.0397)	0.1622^{**} (0.0168)	-5.0030^{**} (0.0225)
	Call	ρ_{k1}	-0.9476^{**} (0.0226)	-1.8282^{**} (0.0117)	-1.7834^{**} (0.0433)
	SMS	ρ_{k2}	-0.2671^{**} (0.0546)	-0.2637^{**} (0.0475)	-0.4397^{**} (0.0471)

* State 1: weak need for mobile Internet, 2: moderate need, 3: strong need.
** Zero is not contained in the 95% credibility interval.

revealed that the negative effects of mobile Internet price are significant only under the hidden states of weak to moderate latent needs for mobile Internet. When customers are in the state of strong needs, they tend to ignore the influence of mobile Internet charge.

For the transition model, we found that conventional telecom behaviors have negative effects on the probability of transitioning to a state of weaker needs for mobile Internet. These results indicate that conventional telecom behaviors may motivate customers to transition to states of stronger latent needs for mobile Internet, which agrees with the findings in the preliminary analysis. The relationships between customers' conventional telecom behaviors and states of latent needs for mobile Internet can be further revealed through the estimated transition matrices (shown in Table 2.9).

We found that the impacts of phone calls on mobile Internet use seemed to be stronger than that of text messages by comparing scenarios 1 - 3 in Table 2.9. Specifically, calls substantially increase the probabilities of transitioning from any state to the state of strong need for mobile Internet (i.e., state 3). That is, calls may be more effective in stimulating customers' latent needs for mobile Internet, compared to texts. These results are reason-

Table 2.9: Transition matrices in different scenarios of communication behaviors in the analysis of mobile Internet usage data.

	Scenario 1: No Activity			Scenario 2: Calls Only		
From \ To	1	2	3	1	2	3
1	0.9771	0.0215	0.0014	0.9431	0.0489	0.0080
2	0.7945	0.1111	0.0944	0.3832	0.0980	0.5188
3	0.0132	0.0066	0.9802	0.0022	0.0011	0.9967
	Scenario 3: Texts Only			Scenario 4: Calls & Texts		
From \ To	1	2	3	1	2	3
1	0.9703	0.0274	0.0023	0.9269	0.0602	0.0129
2	0.7481	0.1196	0.1323	0.3231	0.0858	0.5911
3	0.0086	0.0043	0.9872	0.0015	0.0007	0.9978

able considering that phone calls may be more effective than text messages in affecting customers' psychological status (e.g., to a status of more willing to play online games on their smartphones) or delivering information that should be processed using mobile Internet. We also noted the combination of calls and texts is most effective in increasing customer's likelihood of using mobile Internet (scenario 4 in Table 2.9).

The above analysis not only revealed that the usage of conventional telecom services could positively stimulate that of mobile Internet services, but also delineated the impacts of customers' conventional telecom behaviors on their states of latent needs for mobile Internet. These results provided reference for telecom companies in understanding their business and customers. In the next section, we further demonstrated the potential of the proposed NHMM framework together with the developed SVB method in CRM by examining its out-of-sample forecasting of customers' mobile Internet use behaviors.

2.6.4 Out-of-sample Forecasting

We examined the out-of-sample forecasting performance of the proposed NHMM by comparing with three benchmark models. Model 1 is a classic model of logistic regression

which removes the latent dynamics offered by the HMM framework. $y_{n,t}$ was regarded as the response variable and all possible covariates including $y_{n,t-1}$, $\texttt{Int\_Chg}_{n,t}$, $\texttt{Call}_{n,t-1}$ and $\texttt{SMS}_{n,t-1}$ were included. Model 2 is a homogeneous HMM which retains the emission model given by Equation (2.32) of the proposed NHMM and has time-invariant transition probabilities. Model 3 is a homogeneous HMM which considers all possible covariates including $\texttt{Int\_Chg}_{n,t}$, $\texttt{Call}_{n,t-1}$, and $\texttt{SMS}_{n,t-1}$ in the emission model. Models 2 and 3 were examined to assess the usefulness of allowing for nonhomogeneity in modeling dynamics of hidden states. Model 4 is the proposed NHMM. For a fair comparison, three hidden states were considered for Models 2 and 3. All the three benchmark models were estimated by adapting the proposed SVB method. We used the first 75% observations of each customer for calibration and the remaining 25% observations for validation. We used the predictive log score (PLS, Meligkotsidou and Dellaportas, 2011; Gneiting and Raftery, 2007; Holsclaw et al., 2017) as the criterion for assessment. PLS attains a high score when the model returns large predictive probability for values that occur in the validation set. A higher PLS thus indicates a superior predictive ability. Specifically, PLS was calculated as follows:

$$\text{PLS} = N^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T-T^*} \log \hat{p}(y_{n,T^*+t} | y_{n,\leq T^*+t-1}),$$

where

$$\hat{p}(y_{n,T^*+t} | y_{n,\leq T^*+t-1}) =$$
$$\int_{\boldsymbol{\beta},\boldsymbol{\rho},z_{n,\leq T^*+t-1}} p(y_{n,T^*+t} | y_{n,\leq T^*+t-1}, \boldsymbol{\beta}, \boldsymbol{\rho}, z_{n,\leq T^*+t-1}) \hat{p}(\boldsymbol{\beta}) \hat{p}(\boldsymbol{\rho}) \hat{p}(z_{n,T^*+t-1}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{\rho} \mathrm{d} z_{n,\leq T^*+t-1},$$

T^* denotes the length of calibration period and $\hat{p}(\boldsymbol{\beta})$, $\hat{p}(\boldsymbol{\rho})$, and $\hat{p}(z_{n,T^*+t-1})$ are obtained variational posteriors. The PLS values for Model 1 to Model 4 were calculated as -79572.69, -78040.53, -77874.69, and -75668.61, respectively, indicating the proposed NHMM had the best out-of-sample forecasting performance. The results showcase the potential of

the proposed NHMM framework in CRM. With the system of monitoring and forecasting customers' mobile Internet use behaviors, companies can effectively plan their mobile network capacity in order to provide better services to customers. For instance, if companies forecast that many customers may use the mobile Internet, they may release more capacity for their mobile network so that customers can enjoy high mobile Internet speed. Another interesting finding is that the superior performance of the proposed NHMM relative to Model 3 provides empirical evidence that customers' conventional telecom behaviors may indeed affect their states of latent needs for mobile Internet.

2.7 Discussion

In this paper, we build a framework of variational Bayesian inference consists of two methods for NHMMs with long and ultra-long observation sequences. The proposed VB method works on full sequences. It utilizes a structured Gaussian variational family with a factor covariance structure to approximate the target posteriors and combines the forward-backward algorithm and SGA to update the unknown quantities. The proposed VB method is efficient and handles long sequences in NHMMs. The computational efficiency of the VB method can be further improved through subsampling, leading to the SVB method. The SVB method uses buffers to reduce the bias caused by working on the subsequences directly. We demonstrate that the local nonhomogeneity of NHMMs is crucial in determining the desired buffer lengths. LLEs, which quantify the finite-time local memory decay rates of NHMMs, are proposed to estimate buffer lengths adaptively. An efficient method by treating the unnormalized forward and backward probabilities as RDSs is given to estimate LLEs. The proposed SVB method handles ultra-long sequences for NHMMs efficiently.

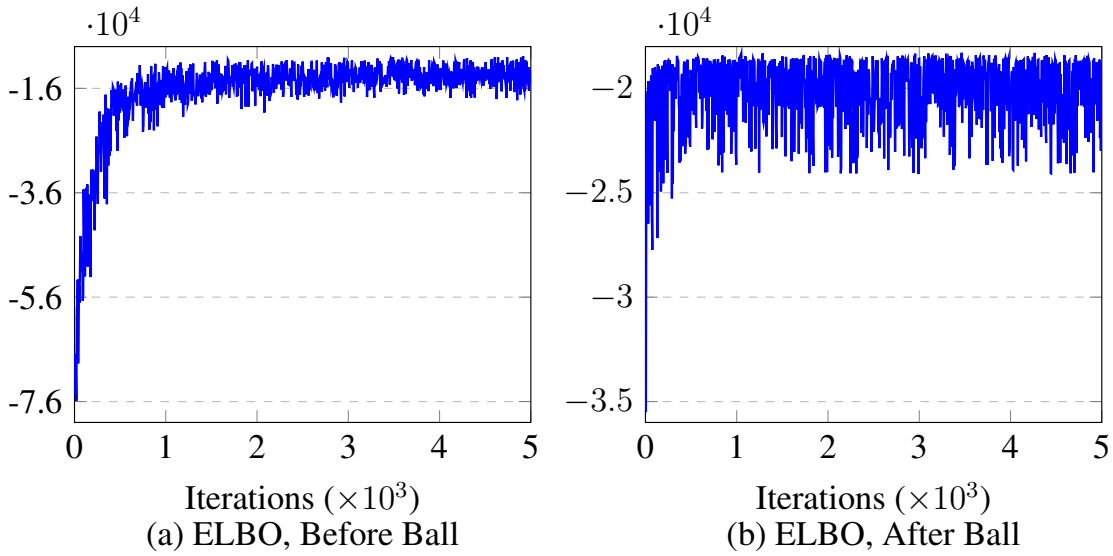
The developed methods are demonstrated useful in two real applications. The first application uses an NHMM together with the VB method to model long eye-tracking scan-paths from children with ASD. The NHMM framework utilizes both the spatial

and temporal information of the scan-paths to determine data-driven ROIs, uncover the underlying hidden state sequences, depict children's attention shift between ROIs, and assess the impacts of HSCs on the attention shift. The results provide evidence on the higher saliency of the puppet relative to the person in the designed social-communicative scenes, which partly explains the advantageous performance of humanoid representations in teaching ASD children. The second application uses an NHMM estimated by the SVB method to model ultra-long telecom records of customers. The analysis focuses on customers' mobile Internet use behaviors by revealing the underlying states of latent needs for mobile Internet, assessing the influences of the conventional telecom behaviors on the latent needs, and evaluating the forecasting ability of the NHMM framework. The results show an overall complementary relationship between the conventional telecom services and mobile Internet services, detailed impacts of different conventional telecom behaviors on mobile Internet use behaviors, and a satisfactory predictive ability of the proposed NHMM framework.

The present study can be extended in several directions. First, the currently considered NHMMs do not consider heterogeneity among subjects. Random effects can be included in the emission and the transition models to account for possible heterogeneity in the emission and hidden processes (Altman, 2007; Ip et al., 2013; Song et al., 2017). Another type of heterogeneity in the number of states across subjects is recently noticed, ignoring which may lead to model misspecification and erroneous interpretations (Padilla et al., 2020). A mixture of HMMs model is proposed to account for such heterogeneity. Extending the current model framework as well as the developed variational methods by considering comprehensive heterogeneity in the longitudinal setting can enhance its flexibility and analytic power. Second, in the developed variational methods, we use a structured variational family of Gaussian distributions with a factor covariance structure. Other structured variational families such as mixtures of Gaussian and Dirichlet processes (Blei and Jordan, 2006) can be considered in different applications. Third, the idea of using LLEs to account for the

nonhomogeneity of NHMMs in this paper could also be considered for other dynamical systems such as state space models in general. Finally, extra data can be collected to strengthen our real data analyses. For instance, if children's neuroimaging data can be simultaneously obtained in our first data example, we may jointly analyze the sequences of neural activities and eye movements to generate more insights. In our second data example, we model customers' decisions to use mobile Internet because the number of users on the network is the major challenge to companies' network capacity. We may further obtain and model customers' expenses on mobile Internet data which are directly related to companies' profits. Customers' other behavioral data may also be obtained and linked with their mobile Internet use behaviors.

2.8 Appendix



(a) ELBO, Before Ball

(b) ELBO, After Ball

Figure 2.8: Plots of ELBO values for the NHMM in the analysis of eye-tracking scan-path data. Convergence is generally achieved within 2000 iterations.

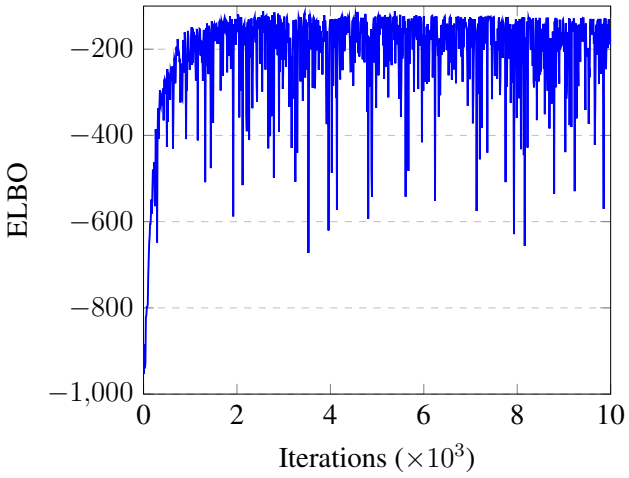


Figure 2.9: Plot of ELBO values for the NHMM in the analysis of mobile Internet usage data. Convergence is achieved around 2000 iterations.

Table 2.10: Bayesian estimates for the parameters in the transition model in the analysis of the eye-tracking scan-path data under the scenario of "Before Ball".

State 1			State 2			State 3			State 4		
Par	Est	SE	Par	Est	SE	Par	Est	SE	Par	Est	SE
$\rho_{12,1}$	-3.2122^{**}	0.1158	$\rho_{22,1}$	2.4821^{**}	0.3173	$\rho_{32,1}$	-1.1875^{**}	0.4322	$\rho_{42,1}$	-1.5048^{**}	0.2597
$\rho_{12,2}$	-0.4132	0.2222	$\rho_{22,2}$	0.6134^{**}	0.1556	$\rho_{32,2}$	0.8220^{**}	0.4162	$\rho_{42,2}$	0.5345	0.2992
$\rho_{12,3}$	-0.1530	0.2085	$\rho_{22,3}$	0.3012	0.3630	$\rho_{32,3}$	0.5674	0.4403	$\rho_{42,3}$	0.6488^{**}	0.3090
$\rho_{12,4}$	-0.1253	0.4573	$\rho_{22,4}$	0.4054	0.4105	$\rho_{32,4}$	0.6759	0.3578	$\rho_{42,4}$	1.1905^{**}	0.2061
$\rho_{12,5}$	-0.1297	0.4680	$\rho_{22,5}$	0.6136	0.5184	$\rho_{32,5}$	0.8371^{**}	0.3438	$\rho_{42,5}$	0.6795^{**}	0.1795
$\rho_{13,1}$	-3.6893^{**}	0.2542	$\rho_{23,1}$	-1.3845^{**}	0.4388	$\rho_{33,1}$	2.9185^{**}	0.3941	$\rho_{43,1}$	-1.7157^{**}	0.3722
$\rho_{13,2}$	-0.7638	1.0175	$\rho_{23,2}$	-0.0588	0.3894	$\rho_{33,2}$	0.4126	0.3187	$\rho_{43,2}$	-0.7110	0.4315
$\rho_{13,3}$	-0.6570	0.6839	$\rho_{23,3}$	0.9172	0.5046	$\rho_{33,3}$	0.9697^{**}	0.3440	$\rho_{43,3}$	0.7766	0.5225
$\rho_{13,4}$	-0.5917	0.5233	$\rho_{23,4}$	0.5863^{**}	0.1645	$\rho_{33,4}$	0.7548	0.4730	$\rho_{43,4}$	0.5023	0.5145
$\rho_{13,5}$	-0.5931	0.9390	$\rho_{23,5}$	0.1553	0.4119	$\rho_{33,5}$	0.4983	0.4165	$\rho_{43,5}$	0.1398	0.4457
$\rho_{14,1}$	-3.4578^{**}	0.4028	$\rho_{24,1}$	-1.1828	0.6622	$\rho_{34,1}$	-1.3863	0.4893	$\rho_{44,1}$	3.0008^{**}	0.3223
$\rho_{14,2}$	-0.2497	0.4738	$\rho_{24,2}$	0.8417	0.5259	$\rho_{34,2}$	0.9048^{**}	0.2642	$\rho_{44,2}$	1.0180^{**}	0.3164
$\rho_{14,3}$	-0.6138	0.3621	$\rho_{24,3}$	0.4988	0.5456	$\rho_{34,3}$	0.2216	0.3192	$\rho_{44,3}$	0.8114^{**}	0.3307
$\rho_{14,4}$	-0.4909^{**}	0.2308	$\rho_{24,4}$	0.4784^{**}	0.1939	$\rho_{34,4}$	0.4771	0.5634	$\rho_{44,4}$	0.3145	0.2569
$\rho_{14,5}$	-0.2985	0.4882	$\rho_{24,5}$	0.5843	0.4695	$\rho_{34,5}$	0.6428^{**}	0.3284	$\rho_{44,5}$	0.7225^{**}	0.1967

Par: Parameter, Est: Estimate, SE: Standard Error.
State 1: $(0, 1050)$, State 2: Background, State 3: Person Face, State 4: Puppet Face.

Table 2.11: Bayesian estimates for the parameters in the transition model in the analysis of the eye-tracking scan-path data under the scenario of "After Ball".

State 1			State 2			State 3			State 4			State 5		
Par	Est	SE	Par	Est	SE	Par	Est	SE	Par	Est	SE	Par	Est	SE
$\rho_{12,1}$	-3.3349^{**}	0.5398	$\rho_{22,1}$	0.7353^{**}	0.1860	$\rho_{32,1}$	-2.0075^{**}	0.3138	$\rho_{42,1}$	-2.3696^{**}	0.9434	$\rho_{52,1}$	-2.1391^{**}	0.7130
$\rho_{12,2}$	-0.3162	0.7169	$\rho_{22,2}$	0.7815	0.4149	$\rho_{32,2}$	0.1628	0.9267	$\rho_{42,2}$	-0.2225	0.8555	$\rho_{52,2}$	-0.2082	0.6756
$\rho_{12,3}$	-0.2621	0.9475	$\rho_{22,3}$	0.9978	0.6259	$\rho_{32,3}$	-0.0382	1.0870	$\rho_{42,3}$	-0.1746	1.4472	$\rho_{52,3}$	-0.0545	1.0119
$\rho_{12,4}$	-0.2189	0.7708	$\rho_{22,4}$	0.7632	0.5732	$\rho_{32,4}$	0.1961	1.1344	$\rho_{42,4}$	0.1618	1.4359	$\rho_{52,4}$	-0.0426	1.1367
$\rho_{12,5}$	-0.2551	1.0507	$\rho_{22,5}$	0.9676	0.7697	$\rho_{32,5}$	-0.1704	1.5235	$\rho_{42,5}$	-0.2305	1.6012	$\rho_{52,5}$	-0.1016	1.2319
$\rho_{12,6}$	-0.2960	0.7137	$\rho_{22,6}$	0.9132^{**}	0.4124	$\rho_{32,6}$	0.3293	1.2444	$\rho_{42,6}$	-0.1492	1.1066	$\rho_{52,6}$	-0.4555	1.1047
$\rho_{12,7}$	-0.2159	1.2296	$\rho_{22,7}$	1.0035	0.8700	$\rho_{32,7}$	-0.4026	1.5212	$\rho_{42,7}$	-0.4144	1.6052	$\rho_{52,7}$	0.0640	1.4009
$\rho_{13,1}$	-4.0622^{**}	0.9388	$\rho_{23,1}$	-2.2671^{**}	0.9872	$\rho_{33,1}$	1.6633^{**}	0.5427	$\rho_{43,1}$	-2.2451^{**}	0.5196	$\rho_{53,1}$	-2.0258^{**}	0.1910
$\rho_{13,2}$	-1.0077	1.3651	$\rho_{23,2}$	-0.4896	1.0335	$\rho_{33,2}$	0.9047	0.5753	$\rho_{43,2}$	-0.3974	0.8406	$\rho_{53,2}$	-0.0887	0.9363
$\rho_{13,3}$	-0.8066	1.4790	$\rho_{23,3}$	0.7941	1.1395	$\rho_{33,3}$	1.0117	0.6536	$\rho_{43,3}$	0.1424	0.8967	$\rho_{53,3}$	0.0043	0.5992
$\rho_{13,4}$	-0.6826	1.0184	$\rho_{23,4}$	0.4254	1.4259	$\rho_{33,4}$	0.9203	0.5813	$\rho_{43,4}$	0.1473	0.7578	$\rho_{53,4}$	0.0300	0.9428
$\rho_{13,5}$	-0.9594	1.1946	$\rho_{23,5}$	-0.3164	1.2986	$\rho_{33,5}$	0.8479	0.6923	$\rho_{43,5}$	-0.3240	1.4741	$\rho_{53,5}$	-0.5981	1.1012
$\rho_{13,6}$	-0.8599	1.2008	$\rho_{23,6}$	-0.1562	1.1282	$\rho_{33,6}$	0.8206	0.6523	$\rho_{43,6}$	-0.2349	1.0752	$\rho_{53,6}$	-0.4559	0.9159
$\rho_{13,7}$	-0.8588	1.5440	$\rho_{23,7}$	-0.5517	1.8278	$\rho_{33,7}$	0.7005	1.0382	$\rho_{43,7}$	-0.5656	1.8128	$\rho_{53,7}$	-0.7380	1.5912
$\rho_{14,1}$	-3.6642^{**}	0.8289	$\rho_{24,1}$	-2.1134^{**}	0.9386	$\rho_{34,1}$	-1.9576^{**}	0.8808	$\rho_{44,1}$	1.6894^{**}	0.4284	$\rho_{54,1}$	-1.9231^{**}	0.5938
$\rho_{14,2}$	-0.6349	1.0550	$\rho_{24,2}$	0.1981	0.8562	$\rho_{34,2}$	0.3469	1.0873	$\rho_{44,2}$	1.0971^{**}	0.3941	$\rho_{54,2}$	0.1989	0.4754
$\rho_{14,3}$	-0.4841	1.1881	$\rho_{24,3}$	-0.6417	1.3152	$\rho_{34,3}$	-0.6336	1.5455	$\rho_{44,3}$	0.9254	0.5429	$\rho_{54,3}$	-0.1618	0.8401
$\rho_{14,4}$	-0.5764	1.1854	$\rho_{24,4}$	0.1316	0.7660	$\rho_{34,4}$	0.0568	1.4159	$\rho_{44,4}$	0.6826	0.6402	$\rho_{54,4}$	0.1629	1.0329
$\rho_{14,5}$	-0.6875	1.1445	$\rho_{24,5}$	-0.2703	1.1001	$\rho_{34,5}$	-0.2541	1.5277	$\rho_{44,5}$	0.6760	0.6627	$\rho_{54,5}$	-0.0672	1.0520
$\rho_{14,6}$	-0.6079	1.0785	$\rho_{24,6}$	0.0791	1.0883	$\rho_{34,6}$	0.4415	1.2429	$\rho_{44,6}$	1.0183	0.5753	$\rho_{54,6}$	0.1762	0.8691
$\rho_{14,7}$	-0.5697	1.3993	$\rho_{24,7}$	-0.5864	1.5599	$\rho_{34,7}$	-0.4967	1.7206	$\rho_{44,7}$	0.7770	0.7365	$\rho_{54,7}$	-0.0629	1.0408
$\rho_{15,1}$	-3.4419^{**}	0.6108	$\rho_{25,1}$	-2.0551^{**}	0.9359	$\rho_{35,1}$	-2.0068^{**}	0.5626	$\rho_{45,1}$	-1.6674^{**}	0.6968	$\rho_{55,1}$	1.5852^{**}	0.4024
$\rho_{15,2}$	-0.5287	0.7055	$\rho_{25,2}$	0.2186	0.8595	$\rho_{35,2}$	0.0221	0.8538	$\rho_{45,2}$	0.3635	0.8259	$\rho_{55,2}$	0.7417^{**}	0.2976
$\rho_{15,3}$	-0.7379	0.7392	$\rho_{25,3}$	-0.1946	1.2550	$\rho_{35,3}$	-0.1710	1.2318	$\rho_{45,3}$	0.0345	1.1261	$\rho_{55,3}$	0.8143	0.5642
$\rho_{15,4}$	-0.6159	0.9357	$\rho_{25,4}$	0.2419	1.0535	$\rho_{35,4}$	-0.0521	1.1977	$\rho_{45,4}$	0.2377	0.9805	$\rho_{55,4}$	0.8447	0.5690
$\rho_{15,5}$	-0.4543	1.1111	$\rho_{25,5}$	0.1500	1.1053	$\rho_{35,5}$	-0.4369	1.5831	$\rho_{45,5}$	0.6259	0.8937	$\rho_{55,5}$	1.1271	0.7300
$\rho_{15,6}$	-0.6652	1.0318	$\rho_{25,6}$	0.1395	0.9829	$\rho_{35,6}$	-0.0606	1.1915	$\rho_{45,6}$	0.4016	0.7560	$\rho_{55,6}$	1.1269	0.6030
$\rho_{15,7}$	-0.4647	1.2746	$\rho_{25,7}$	0.0684	1.4895	$\rho_{35,7}$	-0.4142	1.9991	$\rho_{45,7}$	0.1091	1.2577	$\rho_{55,7}$	1.0315	0.8546

Par: Parameter, Est: Estimate, SE: Standard Error.
State 1: $(0, 1050)$, State 2: Background, State 3: Person Face, State 4: Puppet Face, State 5: Ball.

Chapter 3

A Bayesian approach for estimating the partial potential impact fraction with exposure measurement error under a main study/internal validation design[1]

Abstract

The partial potential impact fraction (pPIF) describes the proportion of disease cases that can be prevented if the distribution of modifiable continuous exposures is shifted in a population, while other risk factors are not modified. It is a useful quantity for evaluating the burden of disease in epidemiologic and public health studies. When exposures are measured with error, the pPIF estimates may be biased, which necessitates methods to correct for the exposure measurement error. Motivated by the Health Professionals Follow-up Study (HPFS), we develop a Bayesian approach to adjust for exposure measurement error when estimating the pPIF under the main study/internal validation study design. We adopt the reclassification approach that leverages the strength of the main study/internal validation study design, and clarify transportability assumptions for valid inference. We

[1] Co-authored with Joseph Chang, Donna Spiegelman, and Fan Li

assess the finite-sample performance of both the point and credible interval estimators via extensive simulations, and apply the proposed approach in the HPFS to estimate the pPIF for colorectal cancer (CRC) incidence under interventions exploring shifting the distributions of red meat, alcohol, and/or folate intake.

3.1 Introduction

The potential impact fraction (PIF), or sometimes called the population impact fraction, refers to the proportion of cases of a disease that would be prevented if the exposure or risk factor distributions were to be modified among a target population. The concept was first introduced as the generalized impact fraction in Walter (1980) and Morgenstern and Bursic (1982) and represents a useful measure that evaluates the burden of disease in epidemiologic and health studies. For instance, our motivating example involves estimating the proportion of colorectal cancer (CRC) cases that might be prevented when the distributions of several modifiable risk factors are shifted among participants in the Health Professionals Follow-up Study (HPFS) (Platz et al., 2000).

The HPFS is a prospective cohort study that started in 1986, and a total of 51,530 male health professionals were enrolled by responding to a baseline questionnaire (Rimm et al., 1991). Participants responded to follow-up questionnaires every two years on topics including dietary intake and health status. The accuracy of responses in the food frequency questionnaires was validated with dietary records in a sub-sample of 127 participants (Rimm et al., 1992). A comparison between the dietary records and participants' responses in the validation study revealed that red meat intake, alcohol intake, and folate intake were subject to measurement error. These errors will likely distort the association estimates between key risk factors and CRC, resulting in misleading estimates of the disease burden (Carroll et al., 2006).

The PIF is a function of the conditional disease probability model and the prevalence

of exposures (Drescher and Becher, 1997). Therefore, accurate estimation of the PIF necessitates valid measurement of exposures as well as a correctly specified outcome model relating disease to exposures and additional covariates. When continuous exposures are measured with error, estimation of the conditional disease probability model that ignores this error typically leads to biased regression parameter estimates (Goldberg, 1975; Copeland et al., 1977; Hsieh and Walter, 1988), which will likely distort the subsequent PIF estimates. Bias in the PIF estimate can be anticipated from the known bias in estimating the population attributable risk (PAR), which measures the fraction of diseases prevented if the exposure were to be eliminated. Hsieh and Walter (1988) showed that the PAR will be underestimated in the presence of misclassification of a binary exposure. Unlike the bias in the single-exposure PAR, which can only be towards the null, Wong et al. (2018) recently demonstrated that the bias in the PAR can be in either direction with two misclassified binary exposures. In addition, these studies found that the magnitude of the bias is most dependent upon the sensitivity of the exposure being eliminated.

Although there is a growing literature on addressing bias in the PAR estimates when the discrete exposures are misclassified (Wong et al., 2020), related methods to correct for measurement error bias in continuous exposures for estimating the PIF are not available. Motivated by the analysis of the HPFS, we develop an estimation strategy for the PIF that operates on mis-measured continuous exposures. We focus on addressing non-differential measurement error in multiple continuous exposures under a main study/internal validation study (MS/IVS) design, where validation data are obtained from participants who are also part of the main study, as is the case with the HPFS. In particular, the measurement error in continuous exposures is non-differential when it is independent of the outcome conditional on error-prone exposures and additional covariates (Yi et al., 2015). On the other hand, measurement error is differential when, for example, sensitivity and specificity of a binary exposure differ between the disease cases and the controls and may arise through dichotomizing a continuous exposure which is subject to non-differential measurement

error (Johnson et al., 2014; Dalen et al., 2009).

In the HPFS, we seek to quantify the partial PIF (pPIF), defined as the fraction of the preventable CRC cases when the distributions of three modifiable risk exposures, red meat intake, alcohol intake, and folate, are shifted, while maintaining the levels of other non-modifiable exposures at their original values. In parallel to the partial PAR (pPAR) and the PAR, the definition of the pPIF extends that of the PIF in a multi-factorial disease setting when not all risk factors are modifiable. Thus, it could be a more interpretable measure of the impact of interventions to be completed, or to what extent the disease burden can be reduced. The PIF and pPIF are analogous to the attributable fraction (AF) and adjusted attributable fraction (aAF) introduced by Eide and Heuch (2001).

Even in the absence of exposure measurement error, inference for the pPIF is non-trivial since it involves integrating over the original and the modified exposure distributions, and a simple closed-form variance expression is not available. Graham (2000) adopted a Bayesian approach for estimating generalized population attributable fraction without measurement error; an important sampling scheme is considered for posterior computation but relies on the choice of an adequate proposal distribution. With a single binary exposure, Pirikahu et al. (2016) developed an efficient Bayesian approach to calculate posterior credible intervals for estimating the PAR without misclassification. They showed via simulations that the posterior credible interval often maintains closer to nominal coverage percentage compared to the delta method or bootstrap resampling. In this article, we develop a computationally tractable Bayesian approach for estimating the pPIF when exposures are measured with error. We model the disease-exposure relationship using logistic regression, and through the Pólya-gamma data augmentation (Polson et al., 2013), we propose an efficient posterior sampling scheme that solely depends on closed-form complete conditionals. The proposed Bayesian algorithm also alleviates the need to derive an asymptotic variance expression for inference, as the variance and interval estimates can be conveniently obtained from the posterior samples.

The remainder of this article is organized into five sections. In Section 3.2, we provide details on the definition of pPIF, the model specifications and assumptions. In Section 3.3, we discuss the proposed Bayesian approach for estimation and inference. An extensive set of simulation studies are presented in Section 3.4, followed by an application to HPFS in Section 3.5. Section 3.6 concludes with a short discussion.

3.2 Models and Assumptions

3.2.1 Partial Potential Impact Fraction

We define the PIF following Eide and Heuch (2001). Specifically, we denote Y as the binary disease status ($Y = 1$ if the disease occurs and $Y = 0$ otherwise), and $\boldsymbol{X}$ as the collection of risk factors or exposures contributing to the occurrence of the disease. In the absence of additional risk factors, the PIF is defined as

$$\text{PIF} = 1 - \frac{\mathbb{P}(Y_* = 1)}{\mathbb{P}(Y = 1)}, \tag{3.1}$$

where $\mathbb{P}(Y = 1)$ and $\mathbb{P}(Y_* = 1)$ denote the disease probabilities before and after the exposure distributions are shifted, respectively. As the PIF increases, the intervention modifying the exposure distributions is increasingly more effective in reducing disease occurrence. In the continuous exposure setting, define

$$\mathbb{P}(Y = 1) = \int \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\beta}) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$
$$\mathbb{P}(Y_* = 1) = \int \mathbb{P}(Y = 1|\boldsymbol{X}^* = \boldsymbol{x}^*; \boldsymbol{\beta}) f^*_{\boldsymbol{X}^*}(\boldsymbol{x}^*) \mathrm{d}\boldsymbol{x}^*,$$

where $\mathbb{P}(Y = 1|\boldsymbol{X}; \boldsymbol{\beta})$ denotes the conditional disease probability model parametrized by $\boldsymbol{\beta}$, $f_{\boldsymbol{X}}$ is the density of exposure $\boldsymbol{X}$, and $f^*_{\boldsymbol{X}^*}$ is the density of the modified exposure $\boldsymbol{X}^*$.

The PIF can thus be written as

$$\text{PIF}(f_{\boldsymbol{X}}, f^*_{\boldsymbol{X}^*}, \boldsymbol{\beta}) = 1 - \frac{\int \mathbb{P}(Y=1|\boldsymbol{X}^*=\boldsymbol{x}^*;\boldsymbol{\beta}) f^*_{\boldsymbol{X}^*}(\boldsymbol{x}^*)\mathrm{d}\boldsymbol{x}^*}{\int \mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\beta}) f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}, \tag{3.2}$$

which depends on density functions $f_{\boldsymbol{X}}$, $f^*_{\boldsymbol{X}^*}$ and parameter $\boldsymbol{\beta}$.

In a multi-factorial disease setting, the burden of disease can be more realistically assessed by the partial PIF (pPIF), which we define below. Let $\boldsymbol{X} = (\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)})$, where $\boldsymbol{X}_{(1)}$ denotes modifiable exposures targeted by the intervention of interest and $\boldsymbol{X}_{(2)}$ denotes other non-modifiable risk factors. Now, $f_{\boldsymbol{X}}$ is the joint density of exposures, $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$, with $\boldsymbol{X}^* = (\boldsymbol{X}^*_{(1)}, \boldsymbol{X}_{(2)})$ and $f^*_{\boldsymbol{X}^*} = f^*_{\boldsymbol{X}^*}(\boldsymbol{x}^*_{(1)}, \boldsymbol{x}_{(2)})$ as the modified joint density. The pPIF is defined as

$$\text{pPIF}(f_{\boldsymbol{X}}, f^*_{\boldsymbol{X}^*}, \boldsymbol{\beta}) = 1 - \frac{\int\int \mathbb{P}(Y=1|\boldsymbol{X}^*_{(1)}=\boldsymbol{x}^*_{(1)}, \boldsymbol{X}_{(2)}=\boldsymbol{x}_{(2)};\boldsymbol{\beta}) f^*_{\boldsymbol{X}^*}(\boldsymbol{x}^*_{(1)}, \boldsymbol{x}_{(2)})\mathrm{d}\boldsymbol{x}^*_{(1)}\mathrm{d}\boldsymbol{x}_{(2)}}{\int\int \mathbb{P}(Y=1|\boldsymbol{X}_{(1)}=\boldsymbol{x}_{(1)}, \boldsymbol{X}_{(2)}=\boldsymbol{x}_{(2)};\boldsymbol{\beta}) f_{\boldsymbol{X}}(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)})\mathrm{d}\boldsymbol{x}_{(1)}\mathrm{d}\boldsymbol{x}_{(2)}}, \tag{3.3}$$

Throughout we assume the intervention functions through a change in the exposure density rather than a change in the disease probability model (Drescher and Becher, 1997). This latter scenario may be more likely to occur when the intervention does not change the exposure distribution but affects unobserved factors underlying the conditional disease probability model. For example, a vaccine program may not affect the prevalence of certain exposures but reduces the risk of getting the disease in the population, in which case the assumption of an exposure density shift may be questionable (Barendregt and Veerman, 2010).

Equation (3.3) reveals that specification of the conditional disease probability model as well as the joint density are critical elements for accurate estimation of the pPIF. For the

conditional disease probability model, the following log-binomial model may be assumed:

$$f_1(Y|\boldsymbol{X},\boldsymbol{\beta}) = \left(e^{\beta_0+\boldsymbol{\beta}_1'\boldsymbol{X}}\right)^Y \left(1-e^{\beta_0+\boldsymbol{\beta}_1'\boldsymbol{X}}\right)^{1-Y}, \tag{3.4}$$

where f_1 denotes the probability function of Y, $\boldsymbol{X}$, and unknown parameter $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$. An attractive feature of the log-binomial model is that it directly estimates the parameter of interest, the risk ratio (Spiegelman et al., 2007; Barendregt and Veerman, 2010). Potential limitations of this model include that the estimated probability is unbounded unless constrained optimization methods are used, which is typically not the case, and there may be unsatisfactory convergence of the model fit with multiple risk factors (Zou, 2004). An alternative specification of the disease probability model is through the logistic regression (Deubner et al., 1980):

$$f_1(Y|\boldsymbol{X},\boldsymbol{\beta}) = \frac{e^{Y(\beta_0+\boldsymbol{\beta}_1'\boldsymbol{X})}}{1+e^{(\beta_0+\boldsymbol{\beta}_1'\boldsymbol{X})}}. \tag{3.5}$$

While the estimation of the log-binomial model requires $e^{\beta_0+\boldsymbol{\beta}_1'\boldsymbol{X}} \in [0,1]$ as a boundary condition, such a condition is not required in logistic regression because the predicted probabilities are naturally bounded. Under the rare disease assumption, typically with the outcome probability $\leq 10\%$, the adjusted odds ratio obtained from the logistic model approximates the adjusted risk ratio in the log-binomial model (Zhang and Yu, 1998). Because the logistic model does not require a boundary condition for the model parameters and simplifies the computation with a closed-form posterior sampling algorithm for estimating the pPIF, we focus on this model in ensuing sections.

3.2.2 Measurement Error Models

When the exposures, $\boldsymbol{X}$, are measured without error, the estimation of $\boldsymbol{\beta}$ and the pPIF can proceed directly with likelihood-based or Bayesian inference, provided $f_1(Y|\boldsymbol{X},\boldsymbol{\beta})$

and $f_{\boldsymbol{X}}$ are correctly specified (Graham, 2000). In epidemiologic studies, exposures are often susceptible to measurement error (for continuous exposures) or misclassification (for categorical exposures). When the exposures are subject to non-differential measurement error, we refer to the mis-measured exposures as surrogates, denoted by $\boldsymbol{Z}$. In the HPFS, the surrogate exposures include the dietary intake measured by the food frequency questionnaire (Rimm et al., 1992). As we show in the simulation study in Section 3.4, failure to account for measurement error in surrogate exposures leads to substantial bias in estimating the pPIF.

Suppose the vector of true exposures, $\boldsymbol{X} = (X_1, \ldots, X_K)$, is of dimension K. The vector of surrogate exposure values, $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$, thus corresponds to $\boldsymbol{X}$ on an element-wise basis. When X_k is not mis-measured, we let $Z_k = X_k$ for $k = 1, \ldots, K$. Hereafter, we assume all elements of surrogate exposures, $\boldsymbol{Z}$, and true exposures, $\boldsymbol{X}$, are continuous. The model describing the relationship between the surrogate and true exposures, or the measurement error model, can be specified in at least two ways. The first strategy describes the measurement error process, where both the marginal true exposure density, $f_{\boldsymbol{X}}$, and the conditional density of the surrogate exposures, $f(\boldsymbol{Z}|\boldsymbol{X})$, the measurement error model, must be specified. In general, parametric or non-parametric approaches can be used to estimate these model parameters (Sinha et al., 2010). On the other hand, Spiegelman et al. (2000) considered a reclassification model that requires the specification of the conditional density of the true exposures given the surrogates. In the context of misclassification, this strategy attempts to reclassify the surrogate categorical exposures into the correct categories, and can be easily extended to continuous exposures, where the conditional probability model is replaced by the conditional density model $f(\boldsymbol{X}|\boldsymbol{Z})$. Under non-differential measurement error (Buonaccorsi, 2010), the joint likelihood of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$

can thus be written as

$$f(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) = f(\boldsymbol{X}|\boldsymbol{Z}) f(\boldsymbol{Y}|\boldsymbol{X}) f(\boldsymbol{Z}) \propto f(\boldsymbol{X}|\boldsymbol{Z}) f(\boldsymbol{Y}|\boldsymbol{X}),$$

which holds because $\boldsymbol{Z}$ is fully observed and $f(\boldsymbol{Z})$ does not contain additional information for parameters of interest.

We adopt the reclassification modeling approach because it is more convenient to model a single conditional density $f(\boldsymbol{X}|\boldsymbol{Z})$. The reclassification process is generically represented as follows:

$$(\boldsymbol{X}|\boldsymbol{Z}) \sim f_2(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\Theta}) = f_2(X_1, \ldots, X_K|Z_1, \ldots, Z_K, \boldsymbol{\Theta}) = \mathcal{P}(\boldsymbol{\Theta}), \tag{3.6}$$

where $\mathcal{P}(\boldsymbol{\Theta})$ is the assumed multivariate distribution with parameter $\boldsymbol{\Theta}$. Note that when $K_1 \leq K$ exposures are mis-measured, we can write $\boldsymbol{X} = (\widetilde{\boldsymbol{X}}, \widetilde{\widetilde{\boldsymbol{X}}})$, where $\widetilde{\boldsymbol{X}}$ contains true values of all K_1 modifiable exposures and/or non-modifiable risk factors that are measured with error, and $\widetilde{\widetilde{\boldsymbol{X}}}$ includes values of all correctly measured exposures and/or risk factors. That is, $\widetilde{\widetilde{\boldsymbol{X}}} \subseteq \boldsymbol{Z}$ by definition. In particular, when all modifiable exposures and non-modifiable risk factors are mis-measured, we have $\boldsymbol{X} = \widetilde{\boldsymbol{X}}$ and $\widetilde{\widetilde{\boldsymbol{X}}} = \emptyset$. This notation allows us to write the reclassification model in (3.6) as:

$$\begin{aligned} & f_2(X_1, \ldots, X_{K_1}|Z_1, \ldots, Z_{K_1}, X_{K_1+1}, \ldots, X_K, \boldsymbol{\Theta}) \\ =& f_2(X_1, \ldots, X_{K_1}|Z_1, \ldots, Z_{K_1}, Z_{K_1+1}, \ldots, Z_K, \boldsymbol{\Theta}) = f_2(\widetilde{\boldsymbol{X}}|\boldsymbol{Z}, \boldsymbol{\Theta}). \end{aligned} \tag{3.7}$$

Spiegelman et al. (2000) suggest that, in the case of dietary intakes, the multivariate normal model provides a reasonable way to characterize the reclassification process. We follow this approach and assume $\mathcal{P}(\boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{Z}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$, where $\boldsymbol{\alpha}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ represent the intercept vector, coefficient matrix and covariance matrix, respectively.

In the MS/IVS design, where mis-measured exposures are validated in a random sample taken from the main study, it is reasonable to assume a *reclassification transportability* condition, where the reclassification process is the same in both the main and validation studies. Importantly, this transportability condition differs from the transportability conditions introduced in Wong et al. (2020). In Wong et al. (2020), *single transportability* holds when, as in simple random sampling into the IVS, the error process $f(\boldsymbol{Z}|\boldsymbol{X})$ is the same in both the main study and the validation study. When the distribution of exposures, $f(\boldsymbol{X})$, is reasonably assumed to be the same between the main study and validation study, they further defined the *double transportability* condition. From Bayes' rule,

$$f(\boldsymbol{X}|\boldsymbol{Z}) = \frac{f(\boldsymbol{Z}|\boldsymbol{X})f(\boldsymbol{X})}{f(\boldsymbol{Z})} = \frac{f(\boldsymbol{Z}|\boldsymbol{X})f(\boldsymbol{X})}{\int f(\boldsymbol{Z}|\boldsymbol{X})f(\boldsymbol{X})\mathrm{d}\boldsymbol{X}}. \tag{3.8}$$

While the single transportability is not sufficient to ensure reclassification transportability, the reclassification transportability is necessary for double transportability. Further, reclassification transportability does not necessarily imply single or double transportability, though double transportablity ensures reclassification transportability. For example, under reclassification transportability, there could exist cases where $f(\boldsymbol{Z})$ is not transportable (e.g. when sampling into the validation study depends on the observed $\boldsymbol{Z}$), implying that the measurement error process is not transportable, further suggesting that even single transportability may not hold (see Web Figure 1 for a Venn diagram). While the reclassification transportability condition is not nested in the conditions studied in Wong et al. (2020), when the internal validation study is a simple random sample of the main study, double transportability holds, which directly implies reclassification transportability. Our approach assumes reclassification transportability, and to provide explicit connections to other assumptions used in the literature (Wong et al., 2020), we summarize the relationship between the three transportability conditions in Proposition 1, and provide a proof in section 3.7.1 of the Appendix.

Proposition 1. Double transportability holds if and only if both single transportability and reclassification transportability hold. Single transportability and reclassification transportability are distinct conditions with neither implying the other.

3.3 Bayesian Estimation and Inference

3.3.1 Likelihood and Prior Specification

Under the MS/IVS design, we have data on disease outcome and surrogate exposures for all main study participants. In addition, the true exposures are measured among internal validation study participants. Let N_m denote the number of participants in the main study who do not have their mis-measured exposures validated, and N_v the size of the validation study, with $N = N_m + N_v$ as the total sample size. Let $\boldsymbol{X}_i \in \mathbb{R}^K$ denote the vector of correctly measured exposures. Following previous definitions, $\boldsymbol{X}_i = (\widetilde{\boldsymbol{X}}_i', \widetilde{\widetilde{\boldsymbol{X}}}_i')'$, where $\widetilde{\boldsymbol{X}}_i$ is the K_1-dimensional vector of mis-measured risk factors, and $\widetilde{\widetilde{\boldsymbol{X}}}_i$ contains the exposures that are correctly measured. $\boldsymbol{Z}_i \in \mathbb{R}^K$ denote vectors of surrogate exposures corresponding on an element-wise basis to $\boldsymbol{X}_i$. In particular, $\boldsymbol{Z}_i$ is fully observed for all N participants, while $\widetilde{\boldsymbol{X}}_i$ is only observed in the validation study. Write $Y_i \in \{0, 1\}$ as the binary outcome, and without loss of generality, we label validation study participants by $i = N_m + 1, \ldots, N_m + N_v$.

Under the reclassification transportability condition, estimation of the pPIF requires the estimation of parameters in both the conditional disease probability model as well as the reclassification model. While it is possible to extend the likelihood-based inference to the pPIF estimation for continuous mis-measured exposures, we take a Bayesian approach here because it circumvents the necessity of deriving the complex asymptotic variance of the pPIF defined as the ratio of integrals, and have shown improved finite-sample operating characteristics for estimating attributable fractions even in the absence of measurement

error (Pirikahu et al., 2016). Treating the variables $\widetilde{\boldsymbol{X}}_i$ for participants i in the main study as latent variables, the complete-data likelihood can be written as

$$\prod_{i=1}^{N} f_1(Y_i|\boldsymbol{X}_i, \boldsymbol{\beta}) f_2(\widetilde{\boldsymbol{X}}_i|\boldsymbol{Z}_i, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{\boldsymbol{x}}). \tag{3.9}$$

These latent variables will be updated in each cycle of the MCMC algorithm presented in Section 3.3.2. In the absence of external knowledge, we assign diffuse conjugate multivariate normal priors for regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, vec($\boldsymbol{\Gamma}$) as well as $\{\widetilde{\boldsymbol{X}}_i\}_{i=1,\ldots,N_m}$, represented by $p(\boldsymbol{\beta})$, $p(\boldsymbol{\alpha})$, $p(\text{vec}(\boldsymbol{\Gamma}))$ and $p(\{\widetilde{\boldsymbol{X}}_i\}_{i=1,\ldots,N_m})$. We assign the conjugate and non-informative inverse Wishart prior for covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, denoted by $p(\boldsymbol{\Sigma}_{\boldsymbol{x}})$.

3.3.2 Posterior Computation

Posterior inference proceeds via data augmentation from the Pólya-gamma representation of the logistic function (Polson et al., 2013). Specifically, we follow Polson et al. (2013) and write the logistic function as a scale mixture of Gaussian densities

$$\frac{\left(e^{\psi}\right)^a}{1+e^{\psi}} = \frac{1}{2} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) \mathrm{d}\omega, \tag{3.10}$$

where $\kappa = a - 1/2$ and $p(\omega)$ is the standard Pólya-Gamma distribution $\mathcal{PG}(1, 0)$. Because the conditional distribution $p(\omega|\psi) = e^{-\omega\psi^2/2} p(\omega) / \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) \mathrm{d}\omega$ is again the Pólya-Gamma distribution $\mathcal{PG}(1, \psi)$, augmenting the data by introducing observation-specific latent variables ω_i admits a closed-form derivation of the complete conditionals and facilitates efficient posterior computation. Based on likelihood (3.9) and these prior

specifications, we can write the joint posterior as

$$\begin{aligned}
&\pi(\{\widetilde{\boldsymbol{X}}_i\}_{i=1,\ldots,N_m}, \{\omega_i\}_{1,\ldots,N}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \\
\propto &\prod_{i=1}^{N} \widetilde{f}_1(Y_i|\boldsymbol{X}_i, \omega_i, \boldsymbol{\beta}) f_2(\widetilde{\boldsymbol{X}}_i|\boldsymbol{Z}_i, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \\
&\times p(\omega_i)p(\boldsymbol{\beta})p(\boldsymbol{\alpha})p(\text{vec}(\boldsymbol{\Gamma}))p(\boldsymbol{\Sigma}_{\boldsymbol{x}})p(\{\widetilde{\boldsymbol{X}}_i\}_{i=1,\ldots,N_m})
\end{aligned} \tag{3.11}$$

where $\widetilde{f}_1(Y_i|\boldsymbol{X}_i, \omega_i, \boldsymbol{\beta})$ is the Pólya-Gamma representation of $f_1(Y_i|\boldsymbol{X}_i, \boldsymbol{\beta})$ according to (3.10).

For posterior computation, we propose the following efficient Markov Chain Monte Carlo (MCMC) algorithm based on closed-form complete conditional distributions. After assigning initial values to all model parameters, the algorithm iterates between the following five steps until convergence:

Step 1: Sample Pólya-Gamma variables $\{\omega_i\}_{i=1,\ldots,N}$

Using the Pólya-Gamma data augmentation method by Polson et al. (2013), we take advantage of the following identity:

$$\frac{\left(e^{\psi}\right)^a}{\left(1+e^{\psi}\right)^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) \mathrm{d}\omega,$$

where $\kappa = a - b/2$ and $\omega \sim \mathcal{PG}(b, 0)$. $\mathcal{PG}(b, 0)$ is the Pólya-Gamma distribution with parameters $(b, 0)$. And the conditional distribution

$$p(\omega|\psi) = \frac{e^{-\omega\psi^2/2} p(\omega)}{\int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) \mathrm{d}\omega},$$

which is also in the Pólya-Gamma class such that $\omega|\psi \sim \mathcal{PG}(b, \psi)$. Sample $\{\omega_i\}_{i=1,\ldots,N}$

from

$$\omega_i|\boldsymbol{X}_i, \beta_0, \boldsymbol{\beta}_1 \sim \mathcal{PG}(1, \beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta}_1).$$

Step 2: Sample MS True Values of Mis-measured Exposures $\{\widetilde{\boldsymbol{X}}_i\}_{i=1,\ldots,N_m}$

The full conditional distribution of $\{\widetilde{\boldsymbol{X}}_i\}_{i=1,\ldots,N_m}$ is proportional to

$$f_1(Y_i|\boldsymbol{X}_i, \boldsymbol{\beta}) f_2(\widetilde{\boldsymbol{X}}_i|\boldsymbol{Z}_i, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{\boldsymbol{x}}),$$

which is then proportional to

$$\frac{\left[\exp\left(\beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta}_1\right)\right]^{Y_i}}{1 + \exp\left(\beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta}_1\right)} \times \exp\left[-\frac{1}{2}(\widetilde{\boldsymbol{X}}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma}\boldsymbol{Z}_i)'\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\widetilde{\boldsymbol{X}}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma}\boldsymbol{Z}_i)\right].$$

The full conditional distribution for $\widetilde{\boldsymbol{X}}_i$ is a multivariate normal $\mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{M}}_i\right)$, where

$$\begin{aligned}
\widetilde{\boldsymbol{M}}_i &= \left(\omega_i \widetilde{\boldsymbol{\beta}}_1 \widetilde{\boldsymbol{\beta}}_1' + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\right)^{-1}, \\
\widetilde{\boldsymbol{\mu}}_i &= \widetilde{\boldsymbol{M}}_i \left\{ \left[Y_i - 1/2 - \omega_i(\beta_0 \widetilde{\widetilde{\boldsymbol{X}}}_i' \widetilde{\widetilde{\boldsymbol{\beta}}}_1) \right] \widetilde{\boldsymbol{\beta}}_1 + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\left(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{Z}_i\right) \right\}.
\end{aligned}$$

Here $\boldsymbol{\beta}_1 = (\widetilde{\boldsymbol{\beta}}_1', \widetilde{\widetilde{\boldsymbol{\beta}}}_1')'$, where $\widetilde{\boldsymbol{\beta}}_1 \in \mathbb{R}^{K_1}$ and $\widetilde{\widetilde{\boldsymbol{\beta}}}_1 \in \mathbb{R}^{K-K_1}$ are parameters corresponding to $\widetilde{\boldsymbol{X}}_i$ and $\widetilde{\widetilde{\boldsymbol{X}}}_i$, respectively.

Step 3: Sample $\boldsymbol{\beta}$

Notation-wise, let $\mathbf{X}$ and $\mathbf{Z}$ denote matrices with each row being $\boldsymbol{X}_i'$ and $\boldsymbol{Z}_i'$ for $i = 1, \ldots, N$ respectively. We then have,

$$\mathbf{W} = \begin{pmatrix} \mathbf{1}_N & \mathbf{X} \end{pmatrix},$$

with the i-th row of $\mathbf{W}$, $\boldsymbol{W}_i = (1, \boldsymbol{X}_i')$ for $i = 1, \ldots, N$. Define $\mathbf{Y} = (Y_1, \ldots, Y_N)$. We

assume a multivariate normal prior $\mathcal{N}(\boldsymbol{b}_0, \boldsymbol{B}_0)$ for $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$. We can then sample $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta}|\mathbf{Y}, \mathbf{W}, \boldsymbol{\Omega} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_\beta, \widetilde{\boldsymbol{M}}_\beta),$$

where

$$\widetilde{\boldsymbol{M}}_\beta = \left(\mathbf{W}'\boldsymbol{\Omega}\mathbf{W} + \boldsymbol{B}_0^{-1}\right)^{-1},$$
$$\widetilde{\boldsymbol{\mu}}_\beta = \widetilde{\boldsymbol{M}}_\beta \left(\mathbf{W}'\boldsymbol{\kappa} + \boldsymbol{B}_0^{-1}\boldsymbol{b}_0\right),$$

with diagonal matrix

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_N \end{pmatrix}$$

and vector

$$\boldsymbol{\kappa} = (Y_1 - 1/2, \ldots, Y_N - 1/2).$$

Step 4: Sample $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$

$\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$ can be jointly sampled due to the model structure. Let $\boldsymbol{\Gamma}^* = (\boldsymbol{\alpha}, \boldsymbol{\Gamma})$ and $\boldsymbol{\Gamma}_k^* = (\alpha_k, \boldsymbol{\Gamma}_k)$, where $\boldsymbol{\Gamma}_k^*$ and $\boldsymbol{\Gamma}_k$ denote the k-th row of $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\Gamma}$ respectively. We separately sample $\boldsymbol{\Gamma}_k^*$ for $k = 1, \ldots, K$. Let

$$\widetilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{1}_N & \mathbf{Z} \end{pmatrix},$$

with the i-th row $\widetilde{\boldsymbol{Z}}_i = (1, \boldsymbol{Z}_i')$. Denote the (i, j)-th entry of $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ by s^{ij}. Assume a multivariate normal prior $\mathcal{N}(\boldsymbol{\gamma}_0^*, \boldsymbol{\Gamma}_0^*)$ for $\boldsymbol{\Gamma}_k^*$. The full conditional for $\boldsymbol{\Gamma}_k^*$ is also a multivariate

normal $\mathcal{N}\left(\widetilde{\boldsymbol{\mu}}^*_{\Gamma,k}, \widetilde{\boldsymbol{M}}^*_{\Gamma,k}\right)$, where

$$\widetilde{\boldsymbol{M}}^*_{\Gamma,k} = \left(s^{kk}\sum_{i=1}^{N}\widetilde{\boldsymbol{Z}}_i'\widetilde{\boldsymbol{Z}}_i + \boldsymbol{\Gamma}_0^{*-1}\right)^{-1},$$
$$\widetilde{\boldsymbol{\mu}}^*_{\Gamma,k} = \widetilde{\boldsymbol{M}}^*_{\Gamma,k}\left[\sum_{i=1}^{N}\left(s^{kk}X_{ik} + \sum_{j\neq k}s^{jk}(X_{ij} - \boldsymbol{\Gamma}^*_j\widetilde{\boldsymbol{Z}}_i')\right)\widetilde{\boldsymbol{Z}}_i' + \boldsymbol{\Gamma}_0^{*-1}\boldsymbol{\gamma}_0^*\right].$$

Step 5: Sample $\boldsymbol{\Sigma_x}$

For $\boldsymbol{\Sigma_x}$, we assume an inverse Wishart prior $\mathcal{IW}\left(\nu_0, \boldsymbol{S}_0^{-1}\right)$, where $\boldsymbol{S}_0$ is a precision matrix, following notations in Hoff (2009). The full conditional for $\boldsymbol{\Sigma_x}$ is $\mathcal{IW}\left(\widetilde{\nu}, \widetilde{\boldsymbol{S}}^{-1}\right)$, where

$$\widetilde{\nu} = \nu_0 + N,$$
$$\widetilde{\boldsymbol{S}} = \boldsymbol{S}_0 + \sum_{i=1}^{N}\left(\boldsymbol{X}_i - \boldsymbol{\Gamma}^*\boldsymbol{Z}_i'\right)\left(\boldsymbol{X}_i - \boldsymbol{\Gamma}^*\boldsymbol{Z}_i'\right)'.$$

The convergence of the algorithm can be monitored by standard Bayesian diagnostics, such as trace plots and the Geweke's z-test. For the estimation of the pPIF according to a pre-specified distributional shift of the true exposures, we additionally obtain the pPIF estimate for each update of the Gibbs sampler. For instance, suppose we are interested in estimating the pPIF after modifying the distributions of a subset of $\boldsymbol{X}$. At a given iteration in the MCMC procedure, we sample $\widetilde{\boldsymbol{X}_i}$ for $i = 1, \ldots, N_m$ and perform the modification on the empirical sample, $\{\boldsymbol{X}_i\}_{i=1,\ldots,N}$ to obtain modified sample $\{\boldsymbol{X}_i^*\}_{i=1,\ldots,N}$. In the motivating HPFS example, one modification could be increasing the daily intake of folate of all participants by 0.5 grams, which is implemented by adding 0.5 to the element corresponding to folate intake in $\boldsymbol{X}_i$ for all $i = 1, \ldots, N$. The pPIF estimate for this update is computed by approximating the integrals in (3.3) through averaging over the empirical joint distribution of $\{\boldsymbol{X}_i\}_{i=1,\ldots,N}$ and $\{\boldsymbol{X}_i^*\}_{i=1,\ldots,N}$. The point estimate and the $100(1-\epsilon)\%$ credible interval for the pPIF are obtained as the mean, $100\epsilon/2$-th and $100(1-\epsilon/2)$-th quantiles of the corresponding posterior samples.

3.4 Simulation Studies

We investigated the accuracy of the proposed Bayesian approach for estimating the pPIF in the presence of exposure measurement error. We set the main study size $N_m = 10000$ and varied the internal validation study size $N_v \in \{100, 250, 500, 1000\}$, representing 1%, 2.5%, 5%, and 10% of the main study. The internal validation study was sampled randomly from the main study. We assumed two targeted modifiable exposures, $\widetilde{\boldsymbol{X}}_i = (X_{i1}, X_{i2})'$, that are measured with error, and a correctly measured non-modifiable exposure, X_{i3}. We simulated the surrogate exposures $(Z_{i1}, Z_{i2})' \sim \mathcal{N}((-0.2, 0.8)', \boldsymbol{\Sigma_z})$, where both diagonal elements of $\boldsymbol{\Sigma_z}$ are 1 and the off-diagonal element is 0.2. The correctly measured non-modifiable exposure X_{i3} was generated from the standard normal distribution. Conditional on $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})'$, where $Z_{i3} = X_{i3}$, we used the multivariate normal reclassification model to generate true exposures as $\widetilde{\boldsymbol{X}}_i \sim \mathcal{N}(\widetilde{\boldsymbol{\Gamma}}(1, \boldsymbol{Z}_i')', \boldsymbol{\Sigma_x})$, and set $\boldsymbol{\Sigma_x} = \boldsymbol{\Sigma_z}$ for simplicity. We specified three sets of values for $\widetilde{\boldsymbol{\Gamma}}$ to reflect low, moderate, and high amount of measurement error in $\widetilde{\boldsymbol{X}}_i$, and summarized these scenarios in Table 3.1.

Table 3.1: Low, moderate, and high degree of measurement errors, their corresponding true reclassification model regression parameters ($\boldsymbol{\Gamma}^*$), and resulting correlations between true (X_1 & X_2) and surrogate (Z_1 & Z_2) exposures.

Measurement Error	$\boldsymbol{\Gamma}^*$	Correlations
Low	$\begin{pmatrix} -0.2 & 1.5 & 0.8 & -0.7 \\ 0.1 & 0.8 & 1.2 & -0.5 \end{pmatrix}$	$\begin{array}{c} \\ X_1 \\ X_2 \end{array} \begin{pmatrix} Z_1 & Z_2 \\ 0.7545 & 0.5022 \\ 0.5432 & 0.7060 \end{pmatrix}$
Moderate	$\begin{pmatrix} -0.2 & 0.6 & 0.2 & -0.7 \\ 0.1 & 0.2 & 0.4 & -0.5 \end{pmatrix}$	$\begin{array}{c} \\ X_1 \\ X_2 \end{array} \begin{pmatrix} Z_1 & Z_2 \\ 0.4632 & 0.2275 \\ 0.2298 & 0.3627 \end{pmatrix}$
High	$\begin{pmatrix} -0.2 & 0.4 & 0.1 & -0.7 \\ 0.1 & 0.1 & 0.25 & -0.5 \end{pmatrix}$	$\begin{array}{c} \\ X_1 \\ X_2 \end{array} \begin{pmatrix} Z_1 & Z_2 \\ 0.3267 & 0.1371 \\ 0.1309 & 0.2320 \end{pmatrix}$

In particular, different choices of $\widetilde{\mathbf{\Gamma}}$ lead to different element-wise marginal correlations between (X_{i1}, X_{i2}) and (Z_{i1}, Z_{i2}), which was specified around 0.75, 0.5 and 0.25 to indicate low, moderate and high amounts of measurement errors. Finally, the conditional disease probability model was taken to be logistic with

$$\mathbb{P}(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3})}, \tag{3.12}$$

where $(\beta_1, \beta_2, \beta_3) = (1, 0.5, 0.5)$, and β_0 was varied to determine the baseline prevalence of disease. Specifically, we investigated cases with $\beta_0 = -2$ and $\beta_0 = -4$ to represent the common and rare disease scenarios, giving baseline disease prevalences around 10% and 2%, respectively. In summary, we studied a factorial design with four sample sizes, three amounts of measurement error, and two baseline disease prevalences, totalling 24 scenarios.

For each scenario, we estimated three pPIFs defined by pPIF_{X_1}, pPIF_{X_2} and pPIF_{X_1,X_2}, which correspond to modifying the distributions of X_1, X_2, and (X_1, X_2). Specifically, pPIF_{X_1} is defined in (3.3) when only X_1 has a location shift with minus half standard deviation and pPIF_{X_2} is defined likewise when only X_2 has a location shift with minus half standard deviation. The quantity pPIF_{X_1,X_2} is defined when both X_1 and X_2 have the same location shift equal to minus half standard deviation.

In each scenario, we compared the proposed Bayesian estimator with an uncorrected estimator that ignores measurement error, two regression calibration (RC) estimators, and a Bayesian estimator that only uses internal validation study data (IVS only) as a benchmark. The uncorrected estimator fits a Bayesian logistic disease model based on the mis-measured risk factors without measurement error correction, and serves to quantify the measurement error bias. Assuming reclassification transportability, the first RC estimator ($RC_{M/I}$) fits the reclassification model using the internal validation study data, and then predicts the unobserved true exposures in the main study using the estimated reclassification model

(Carroll et al., 2006). The coefficients in the disease probability model are then estimated using the predicted true exposures in the main study, and the pPIFs are estimated using predicted true exposures in the main study together with true exposures in the internal validation study. Similarly, the second RC estimator (RC_I) first fits the reclassification model using the internal validation study data, predicts the unobserved true exposures in the main study, and then estimates the conditional disease probability using predicted true exposures in the main study. In contrast to $RC_{M/I}$, RC_I assumes double transportability and only uses observed true exposures in the internal validation study to estimate pPIFs. Procedures fo the two RC estimators are summarized as follows:

Procedure for $RC_{M/I}$ (assume Reclassification Transportability)

1. Estimate reclassification model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ using validation study data, $\{\boldsymbol{X}_i\}_{i=N_m+1,\ldots,N}$ and $\{\boldsymbol{Z}_i\}_{i=N_m+1,\ldots,N}$, to obtain estimates $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ via the maximum likelihood estimator (MLE);

2. Impute true exposures in the main study by expected values $\widehat{\boldsymbol{X}_i} = \mathbb{E}(\boldsymbol{X}_i|\boldsymbol{Z}_i) = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\Gamma}}(1, \boldsymbol{Z}_i')'$ for $i = 1, \ldots, N_m$;

3. Estimate conditional disease probability model parameter $\boldsymbol{\beta}$ using $\{Y_i\}_{i=1,\ldots,N_m}$ and $\{\widehat{\boldsymbol{X}_i}\}_{i=1,\ldots,N_m}$ via MLE, to obtain $\widehat{\boldsymbol{\beta}}$;

4. Estimate the pPIF using estimated conditional disease probability model parameter $\widehat{\boldsymbol{\beta}}$ and empirical samples of $\{\widehat{\boldsymbol{X}_i}\}_{i=1,\ldots,N_m}$ and $\{\boldsymbol{X}_i\}_{i=N_m+1,\ldots,N}$.

Procedure for RC_I (assume Double Transportability)

1. Estimate reclassification model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ using validation study data, $\{\boldsymbol{X}_i\}_{i=N_m+1,\ldots,N}$ and $\{\boldsymbol{Z}_i\}_{i=N_m+1,\ldots,N}$, to obtain estimates $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ via the maximum likelihood estimator (MLE);

2. Impute true exposures in the main study by expected values $\widehat{\boldsymbol{X}_i} = \mathbb{E}(\boldsymbol{X}_i|\boldsymbol{Z}_i) = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\Gamma}}(1, \boldsymbol{Z}_i')'$ for $i = 1, \ldots, N_m$;

3. Estimate conditional disease probability model parameter $\boldsymbol{\beta}$ using $\{Y_i\}_{i=1,\ldots,N_m}$ and $\{\widehat{\boldsymbol{X}_i}\}_{i=1,\ldots,N_m}$ via MLE, to obtain $\widehat{\boldsymbol{\beta}}$;

4. Estimate the pPIF using estimated conditional disease probability model parameter $\widehat{\boldsymbol{\beta}}$ and empirical samples of $\{\boldsymbol{X}_i\}_{i=N_m+1,\ldots,N}$.

Finally, the IVS only estimator estimates the conditional disease probability model and the pPIF using internal validation study only. The procedure is considered a benchmark for bias because all variables are free of measurement error in the internal validation study. Standard errors and confidence intervals for the RC and IVS only approaches are computed via nonparametric bootstrap with 1000 replicates.

For the proposed Bayesian, uncorrected, and the IVS only estimators, we use a non-informative $\mathcal{N}(0, 10^3)$ prior for each element of $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\Gamma}}$, as well as a $\mathcal{IW}(3, I)$ prior for $\boldsymbol{\Sigma_x}$, and run one chain of length 10000 (with the first 5000 iterations as burn-in). We compare the estimators in terms of relative bias (Bias $\%$), the Monte Carlo standard error (MCSE), average of the estimated standard error (ASE), root mean squared error (RMSE), and frequentist coverage percentage of the credible (confidence) interval (Coverage $\%$). The relative bias is defined as the ratio of bias over the true pPIF. We use the MCSE to quantify the efficiency of the estimator, and compare the ASE with MCSE to assess the accuracy of the standard error estimate. The simulation results are summarized from 1000 data replications.

Table 3.2: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 100, 250$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 3.4, *Bayes* = proposed. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease) and $\beta_0 = -4$ (rare disease), multivariate normal true reclassification model, and the moderate measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators									
			$\beta_0 = -2$					$\beta_0 = -4$				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*	*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
100	pPIF_{X_1}	BIAS	-27.77	-9.98	-20.53	2.86	1.16	-26.36	-35.48	-19.12	-5.12	3.57
		MCSE	0.009	0.097	0.064	0.067	0.049	0.018	0.231	0.088	0.203	0.105
		ASE	0.008	0.265	0.093	0.068	0.042	0.018	0.841	0.120	0.279	0.068
		RMSE	0.082	0.103	0.085	0.068	0.050	0.097	0.274	0.111	0.206	0.112
		Coverage	0.00	96.50	92.90	**94.50**	89.30	0.00	97.20	**96.20**	91.50	79.80
	pPIF_{X_2}	BIAS	-18.74	-37.02	-26.87	-1.86	1.60	-18.41	-102.62	-26.74	-39.82	-7.32
		MCSE	0.009	0.157	0.099	0.081	0.067	0.023	0.341	0.131	0.260	0.171
		ASE	0.009	0.351	0.136	0.085	0.061	0.021	1.124	0.183	0.398	0.109
		RMSE	0.030	0.169	0.107	0.083	0.069	0.044	0.413	0.142	0.277	0.180
		Coverag	5.20	97.50	**95.90**	**94.60**	90.20	0.00	97.30	96.50	91.20	78.80
	$\text{pPIF}_{X_1X_2}$	BIAS	-23.19	-3.37	-18.77	1.98	1.94	-18.41	-13.10	-15.07	-12.12	3.94
		MCSE	0.009	0.067	0.051	0.069	0.043	0.023	0.133	0.060	0.199	0.077
		ASE	0.009	0.201	0.074	0.071	0.038	0.021	0.633	0.093	0.260	0.054
		RMSE	0.096	0.070	0.089	0.069	0.044	0.107	0.160	0.094	0.202	0.083
		Coverage	0.00	97.20	76.90	**93.80**	88.30	56.50	98.10	86.10	93.10	81.30
250	pPIF_{X_1}	BIAS	-27.60	-0.84	-15.38	0.96	0.75	-26.27	-2.96	-12.94	0.50	0.65
		MCSE	0.008	0.060	0.046	0.037	0.028	0.020	0.077	0.061	0.091	0.050
		ASE	0.008	0.082	0.047	0.039	0.029	0.018	0.146	0.069	0.096	0.048
		RMSE	0.081	0.062	0.063	0.037	0.028	0.097	0.079	0.076	0.092	0.052
		Coverage	0.00	92.90	83.50	**95.70**	**95.00**	0.00	**93.60**	91.30	93.40	92.00
	pPIF_{X_2}	BIAS	-20.09	-11.98	-20.88	1.31	-0.46	-19.89	-21.14	-19.16	-7.39	-0.48
		MCSE	0.009	0.099	0.072	0.048	0.041	0.023	0.160	0.104	0.123	0.079
		ASE	0.009	0.122	0.075	0.049	0.039	0.021	0.232	0.114	0.130	0.077
		RMSE	0.033	0.104	0.079	0.049	0.042	0.045	0.174	0.111	0.129	0.082
		Coverage	6.50	93.30	93.50	93.30	92.70	49.50	**94.10**	**95.50**	**93.80**	92.50
	$\text{pPIF}_{X_1X_2}$	BIAS	-23.50	-1.45	-14.77	1.26	0.62	-21.72	-2.11	-10.90	-0.05	1.35
		MCSE	0.009	0.041	0.030	0.039	0.026	0.020	0.052	0.040	0.089	0.040
		ASE	0.009	0.057	0.034	0.040	0.025	0.019	0.095	0.047	0.094	0.038
		RMSE	0.098	0.042	0.064	0.039	0.026	0.109	0.054	0.066	0.091	0.041
		Coverage	0.00	**94.90**	55.50	**94.60**	92.80	0.00	**94.50**	77.80	93.40	92.30

Table 3.3: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 500, 1000$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 3.4, *Bayes* = proposed. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease) and $\beta_0 = -4$ (rare disease), multivariate normal true reclassification model, and the moderate measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators									
			$\beta_0 = -2$					$\beta_0 = -4$				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*	*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
500	pPIF_{X_1}	BIAS	-27.63	-1.51	-14.58	-0.02	0.72	-26.17	-2.52	-11.91	0.63	0.26
		MCSE	0.008	0.038	0.031	0.026	0.020	0.018	0.056	0.048	0.058	0.038
		ASE	0.008	0.040	0.033	0.027	0.020	0.018	0.058	0.050	0.058	0.036
		RMSE	0.082	0.018	0.051	0.026	0.022	0.096	0.058	0.064	0.059	0.039
		Coverage	0.00	**94.50**	77.00	**95.40**	**93.80**	0.00	**95.10**	87.00	93.30	93.00
	pPIF_{X_2}	BIAS	-19.32	-2.70	-18.27	-0.32	-0.83	-19.10	-5.19	-16.53	-2.24	-0.30
		MCSE	0.008	0.059	0.050	0.033	0.028	0.021	0.094	0.081	0.078	0.057
		ASE	0.009	0.063	0.053	0.034	0.028	0.021	0.102	0.084	0.078	0.057
		RMSE	0.032	0.060	0.057	0.033	0.028	0.044	0.097	0.088	0.079	0.060
		Coverage	5.00	**94.60**	92.30	**94.00**	**94.40**	53.50	**94.40**	**93.80**	**94.00**	93.00
	$\text{pPIF}_{X_1X_2}$	BIAS	-23.28	-1.40	-13.98	0.04	0.36	-21.42	-1.48	-10.35	0.59	0.66
		MCSE	0.009	0.027	0.022	0.027	0.018	0.020	0.036	0.030	0.054	0.030
		ASE	0.009	0.029	0.023	0.027	0.019	0.019	0.040	0.032	0.056	0.029
		RMSE	0.097	0.028	0.058	0.026	0.020	0.108	0.038	0.057	0.054	0.032
		Coverage	0.00	**95.10**	29.20	**94.80**	**93.70**	0.00	**95.20**	65.20	**94.10**	**94.20**
1000	pPIF_{X_1}	BIAS	-27.71	-2.51	-14.60	-0.19	0.37	-26.39	-2.27	-11.47	0.42	0.09
		MCSE	0.008	0.029	0.024	0.018	0.015	0.018	0.047	0.038	0.040	0.027
		ASE	0.008	0.030	0.025	0.019	0.015	0.018	0.046	0.041	0.039	0.028
		RMSE	0.082	0.030	0.047	0.017	0.014	0.097	0.048	0.055	0.040	0.028
		Coverage	0.00	**93.60**	65.20	**95.40**	**94.30**	0.00	**94.10**	85.90	**94.00**	**93.80**
	pPIF_{X_2}	BIAS	-18.72	-1.81	-16.33	0.46	-0.57	-19.09	-5.86	-14.06	-0.70	0.30
		MCSE	0.010	0.047	0.040	0.024	0.020	0.021	0.081	0.065	0.054	0.044
		ASE	0.009	0.047	0.040	0.024	0.021	0.021	0.080	0.070	0.053	0.043
		RMSE	0.030	0.047	0.047	0.024	0.020	0.044	0.082	0.071	0.055	0.045
		Coverage	5.50	**94.40**	90.50	**94.40**	**95.20**	49.50	**94.50**	**94.80**	**94.50**	92.70
	$\text{pPIF}_{X_1X_2}$	BIAS	-23.15	-2.08	-13.61	0.09	0.14	-21.54	-2.12	-9.81	0.46	0.46
		MCSE	0.009	0.021	0.017	0.018	0.013	0.019	0.031	0.024	0.037	0.023
		ASE	0.009	0.021	0.017	0.019	0.014	0.018	0.031	0.027	0.038	0.023
		RMSE	0.096	0.023	0.055	0.017	0.014	0.108	0.033	0.052	0.037	0.024
		Coverage	0.00	92.90	10.70	**94.20**	**94.80**	0.00	92.60	54.80	**94.40**	**93.80**

Tables 3.2 and 3.3 present the simulation results when disease is common ($\beta_0 = -2$) and rare ($\beta_0 = -4$), and the amount of measurement error is moderate, under varying validation study sizes. It is evident that the uncorrected approach of ignoring the measurement error in the modifiable risk factors leads to substantial negative bias in estimating all three pPIFs. The two RC estimators frequently reduced the bias in estimating the pPIF compared to the uncorrected estimator, and demonstrate complementary behaviours with different levels of IVS sample size.

Interestingly, when the IVS was small ($N_v = 100$ and 250), $RC_{M/I}$ generally had smaller bias compared to RC_I, and demonstrates smaller bias with a larger IVS sample size. RC_I, however, shows significant bias as the size of IVS increases, which is caused by the bias in estimating conditional disease probability model coefficients. This can be inferred by comparing results from the IVS only estimator and RC_I, because the IVS only estimator gives unbiased estimates of both the conditional disease probability model and the pPIF, and the only difference between the two estimator lies in the way they estimate the conditional disease probability model coefficients. As suggested by Carroll et al. (2006), RC estimators give biased estimates of logistic disease model coefficients when the effect of the exposure subject to measurement error is moderate to high, especially as the disease rate increases and/or measurement error increases, which is the case in our simulations (Rosner et al., 1989). We investigate this issue further in section 3.7.3 of the Appendix, and observe improved performance for RC estimators as the effect of $\widetilde{\boldsymbol{X}}_i$ lessens, although we still see an advantage to the proposed Bayesian approach in the presence of substantial measurement error, as often occurs in epidemiology.

Finally, the proposed Bayesian approach demonstrates the smallest and negligible bias and RMSE regardless of the size of the IVS and the amount of measurement error. In addition, the Bayesian estimator has the highest efficiency in almost all scenarios. Throughout, the ASE of the Bayesian estimator is close to its MCSE, suggesting that the Bayesian posterior variance estimator accurately quantifies the uncertainty in the pPIF.

When the disease is common, the frequentist coverage percentage of the Bayesian credible interval estimator is close to 95% when $N_v \geq 250$. Under the rare disease scenario, the required level of N_v needs to be 500 to obtain nominal coverage percentage of the Bayesian credible interval estimator (shown in Table 3.3).

The comparisons between the proposed approach and other estimators under low and high measurement error scenarios are qualitatively similar and presented in Tables 3.8 to 3.11 in section 3.7.2 of the Appendix. In particular, the bias of the uncorrected approach became positive in the low measurement error scenario, echoing the observations of Wong et al. (2018), that the bias of the pPIF with multiple risk factors could be either positive or negative. In the low measurement error scenario, the coverage percentages of the proposed Bayesian interval estimator also improved compared to that in moderate or high measurement error scenarios, as it consistently stayed at least 90% even with limited validation data. In contrast, the coverage percentage of the Bayesian interval estimator further declined with high measurement error, small validation study and rare disease (Table 3.11). In section 3.7.4 of the Appendix, we also include an additional simulation study where the IVS is obtained as a biased sample from the MS, and find the proposed Bayesian estimator still maintains satisfactory performance.

Because our Bayesian estimator assumed a multivariate normal reclassification model, we carried out additional simulations to investigate the sensitivity of the results when the normality assumption is violated. We maintain the above data generating process except that the true exposures were simulated from a bivariate Gamma distribution with the same mean and covariance. We update the simulations parameters so that: $\widetilde{\boldsymbol{X}_i} = \widetilde{\boldsymbol{\Gamma}}(1, \boldsymbol{Z}_i')' + \mathcal{G}_2(\boldsymbol{\Sigma}_{\boldsymbol{x}})$, where $\mathcal{G}_2(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ is a mean-centered bivariate Gamma density with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. We further choose the parameters so that the two marginal distributions implied from $\mathcal{G}_2(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ are univariate $\mathcal{G}(2, 2)$ with mild skewness $\sqrt{2}$. For brevity, we present selected simulation results in Tables 3.12 to 3.15 in section 3.7.2 of the Appendix. Our finding is that, when the distribution of the true exposures exhibits mild skewness, the proposed Bayesian approach

could still yield accurate point estimate for the pPIF with slightly reduced precision, as well as reasonable frequentist coverage percentage of the interval estimator.

3.5 Application to the HPFS Data

We applied the proposed Bayesian approach to correct the measurement errors for estimating the pPIF of CRC in the Health Professionals Follow-up Study (Platz et al., 2000). For illustration, our analysis is based on complete data records (regarding intakes of red meat, alcohol, and folate, as well as age and body mass index) only, leading to 48691 participants in the main study and 126 participants in the validation study. A total of 1913 (3.9%) and 6 (4.8%) CRC cases occurred over follow-up in the main study and validation study, respectively. Our objective was to estimate the pPIF for CRC when several continuous exposures are modified. We focus on three modifiable exposures that are error-prone: red meat intake (servings per day), alcohol intake (servings per day), and folate intake (grams per day). A descriptive summary of self-reported values of the three modifiable exposures is given in Table 3.4.

Table 3.4: A descriptive statistics at baseline of self-reported (surrogate) exposures and factors for HPFS participants. Red Meat: red meat intake. Alcohol: alcohol intake. Folate: folate intake. $N_m = 44849$, $N_v = 126$.

Variable (Unit)	Mean (Std. Dev.)	(Min, Max)
Red Meat (servings/day)	1.270 (0.617)	(0.000, 3.000)
Alcohol (servings/day)	0.729 (0.627)	(0.000, 2.571)
Folate (g/day)	0.434 (0.190)	(0.067, 1.003)
Age (year)	54.49 (9.935)	(32.00, 79.00)
BMI (kg/m^2)	25.53 (3.365)	(12.91, 91.67)

We also consider two error-free risk factors: age (years) and body mass index (BMI, kg/m^2). In the main study, records containing surrogate exposures (any in red meat, alcohol, or folate intakes) with values below the first quartile minus $1.5\times$IQR (inter quarter range)

or above the third quartile plus $1.5\times$IQR were considered outliers and were removed to prevent their undue influence on the analysis. The final main study dataset includes 44975 participants with 1748 (3.9%) CRC cases.

We estimated a set of pPIFs by first modifying one exposure (red meat intake, alcohol intake, or folate intake) only. Then we modified these exposures in pairs (i.e., red meat and alcohol, red meat and folate, folate and alcohol). Finally, we simultaneously modified all three exposures and computed the pPIF. Previous studies have shown that increasing red meat intake and alcohol intake leads to increased risk of CRC occurrence, while increasing folate intake protected against CRC risk (Giovannucci et al., 1994, 1995). Therefore, we studied the impact of reductions in red meat intake by 0.5 servings per day, in alcohol intake by 0.5 servings per day, and increase in folate intake by 0.5 grams per day.

Table 3.5: Correlations between true and surrogate exposures computed from validation study data, $N_v = 126$. (X_1, Z_1): true and surrogate exposures of red meat intake (servings/day). (X_2, Z_2): true and surrogate exposures of alcohol intake (servings/day). (X_3, Z_3): true and surrogate exposures of folate intake (grams/day).

	Z_1	Z_2	Z_3
X_1	0.685	0.206	-0.252
X_2	0.263	0.852	-0.175
X_3	-0.150	-0.276	0.596

In notation, let X_1 and Z_1 represent the true and surrogate values for red meat intake, X_2 and Z_2 represent the true and surrogate values for alcohol intake, while X_3 and Z_3 represent the true and surrogate values for folate intake. Correlations between true and surrogate exposures computed from the validation study data are presented in Table 3.5. The non-modifiable risk factors, age and BMI (correctly-measured and denoted by X_4 and X_5), were transformed to quintiles to allow for nonlinear associations. We assume that reclassification transportablility holds, and for the conditional disease probability model, we used the logistic regression model (3.5). We assumed a multivariate normal reclassification process, as in our simulation study. In addition, an uncorrected approach that ignored

measurement error was used to estimate the same set of pPIFs.

For each estimator, we specified a non-informative $\mathcal{N}(0, 10^3)$ prior for each element of $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\Gamma}}$, and a $\mathcal{IW}(4, I)$ prior for $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. We ran a chain of 15000 iterations (with the first 5000 discarded as burn-in), retaining every fifth iteration. Posterior convergence was monitored by trace plots and the Geweke's z-scores. As a further comparison, we also included estimates from the two RC approaches investigated in Section 3.4. The disease model coefficients, pPIF estimates and associated 95% credible intervals (confidence intervals for RC estimators) are presented in Tables 3.6 to 3.7. Estimation results for other model parameters are presented in Tables 3.16 to 3.17 in section 3.7.2 of the Appendix, with empirical density plots of surrogate and estimated true exposures provided in Figure 3.2 in the same section of the Appendix.

Table 3.6: Association estimates ($\widehat{\beta}$) in the conditional disease probability model from the uncorrected, $RC_{M/I}$, RC_I, and the proposed Bayesian methods, the HPFS (1986-2010), $N_m = 44,849$, $N_v = 126$. The 95% posterior credible intervals are provided in the parentheses.

Method	Red Meat	Alcohol	Folate	Age	BMI
Uncorrected	0.069 *(-0.004, 0.148)*	0.140 *(0.068, 0.215)*	-0.223 *(-0.481, 0.032)*	0.371 *(0.333, 0.409)*	0.050 *(0.020, 0.085)*
$RC_{M/I}$	0.008 *(-0.156, 0.139)*	0.164 *(0.075, 0.253)*	-0.287 *(-0.995, 0.420)*	0.372 *(0.334, 0.411)*	0.055 *(0.017, 0.092)*
RC_I	0.008 *(-0.153, 0.137)*	0.158 *(0.061, 0.256)*	-0.287 *(-0.974, 0.401)*	0.372 *(0.332, 0.412)*	0.055 *(0.017, 0.092)*
Bayesian	0.001 *(-0.116, 0.115)*	0.163 *(0.070, 0.266)*	-0.338 *(-1.072, 0.288)*	0.375 *(0.333, 0.420)*	0.055 *(0.016, 0.093)*

Table 3.6 indicates that association of alcohol intake and folate intake with disease becomes greater after measurement error correction, which leads to larger estimates for the pPIF associated with reductions in alcohol intake (pPIF=0.045 before correction versus pPIF=0.064 using the proposed Bayesian estimator) and with increases in folate intake (pPIF=0.105 before correction versus pPIF=0.135 using the proposed Bayesian estimator) by given increments. This indicates that measurement errors in the two exposures could

have resulted in an underestimate of the impact of potential interventions targeting alcohol and folate intake.

Table 3.7: The pPIF estimates from the uncorrected and the proposed method. Red meat intake and alcohol intake were decreased by 0.5 servings per day for all participants. (The intake level was set to zero for if the original value was below 0.5.) Folate intake was increased by 0.5 grams per day for all participants. '✓' indicates that exposure was modified when estimating the pPIF. The 95% posterior credible intervals are given in the parentheses.

Modified Exposures			Methods			
Red Meat	Alcohol	Folate	Uncorrected	$RC_{M/I}$	RC_I	Bayesian
✓			0.030 *(-0.007, 0.064)*	0.004 *(-0.075, 0.067)*	0.004 *(-0.071, 0.064)*	0.006 *(-0.066, 0.075)*
	✓		0.045 *(0.020, 0.068)*	0.069 *(0.029, 0.110)*	0.062 *(0.024, 0.100)*	0.064 *(0.028, 0.100)*
		✓	0.105 *(-0.011, 0.209)*	0.128 *(-0.163, 0.419)*	0.127 *(-0.153, 0.407)*	0.135 *(-0.130, 0.361)*
✓	✓		0.074 *(0.036, 0.113)*	0.073 *(-0.006, 0.138)*	0.066 *(-0.011, 0.127)*	0.071 *(-0.004, 0.126)*
✓		✓	0.133 *(0.021, 0.235)*	0.132 *(-0.129, 0.377)*	0.131 *(-0.126, 0.373)*	0.140 *(-0.109, 0.343)*
	✓	✓	0.146 *(0.029, 0.247)*	0.189 *(-0.078, 0.456)*	0.181 *(-0.076, 0.438)*	0.191 *(-0.047, 0.400)*
✓	✓	✓	0.172 *(0.065, 0.272)*	0.193 *(-0.043, 0.414)*	0.185 *(-0.047, 0.404)*	0.198 *(-0.025, 0.378)*

In contrast, the conditional disease probability model coefficient and the pPIF corresponding to red meat intake are larger without measurement error correction (pPIF$= 0.030$ before correction versus pPIF$= 0.006$ using the proposed Bayesian estimator), suggesting that measurement error may have led to an overestimation of the impact of potential interventions targeting red meat intake alone. This could be resulted by accounting for folate intake in the conditional disease probability model. Nevertheless, correcting for bias due to measurement error still led to a larger pPIF estimate for an intervention modifying all three exposures (pPIF=0.172 before correction versus pPIF=0.198 using the proposed Bayesian estimator).

The pPIF values estimated by the two RC approaches were often similar, and were larger than those from the proposed Bayesian approach for the folate intake increase intervention.

However, the RC results may not be as accurate as the proposed approach due to the caveat discussed in our simulations. In summary, the CRC disease burden in the HPFS and similar populations could be reduced by 19.1% when simultaneously reducing alcohol intake while increasing folate intake, which is similar to simultaneously conducting interventions on all three exposures. The estimated additional reduction in CRC occurrence results from a decrease in red meat intake, however, appeared to be relatively minimal.

3.6 Discussion

The pPIF extends the concept of the pPAR, and provides a useful means for assessing the potential impact of interventions targeting continuous exposures in a population. Motivated by the HPFS, we developed a computationally tractable Bayesian approach for estimating the pPIF when continuous exposures are measured with error. This estimator is based upon the reclassification model, and specifies conditional distributions of the true exposures given the observed surrogate exposures. In the MS/IVS design considered here, we assumed the *reclassification transportability* condition, requiring that the reclassification process to be exchangeable between the MS and the IVS, and clarified the relationship between this assumption and the transportability conditions proposed previously. Finally, this Bayesian procedure allows us to easily obtain both point and interval estimates for a range of pPIF values, without resorting to complex numerical integration.

Across a range of sample sizes, disease prevalences, and degrees of measurement error, our simulations suggested that, in the presence of exposure measurement error, the proposed Bayesian approach can reduce the bias in estimating the pPIF and accurately quantify the uncertainty. In particular, the estimation bias can be dramatically reduced with an IVS as small as 100, while nominal frequentist coverage may require a larger IVS, at least 250 in our simulation scenarios, in the presence of moderate to high amount of measurement error. Ignoring exposure measurement error leads to substantial bias and under coverage of the

credible interval, and the bias could be towards either direction.

There are several potential limitations of our approach, both of which will be pursued in our future work. First, only continuous surrogate and true exposures can enter the model, which implies a possible improvement of including both discrete/categorical and continuous exposures into the model. Second, we have assumed that there are no missing surrogate exposures in both the MS and the IVS. In the HPFS, missing exposures could arise from non-response of dietary questionnaires, and have been excluded from our analysis. Assuming exposures missing at random (MAR), a potential modification within our Bayesian framework could regard the missing surrogate exposures as model parameters and include additional Gibbs updates for these missing values, alongside estimating the remaining model parameters. This procedure is akin to the joint modeling approach used in multiple imputation (Molenberghs et al., 2014). It remains to be explored whether such an approach would perform well in simulations and substantially change our empirical pPIF estimates for the HPFS. Third, we have assumed the availability of an IVS, as motivated by the HPFS data. Often, the validation study occurs in an external population, where the disease outcome is not available. Given the importance of this design in practice, it would be worthwhile to extend the current approach to the main study/external validation study design in the presence of exposure measurement error.

3.7 Appendix

3.7.1 Proofs for Claimed Relationships between ST, RT, and DT

We assume all marginal, conditional, and joint densities exist and are positive on their respective supports. Following notations in the main article, let $\boldsymbol{X}$ and $\boldsymbol{Z}$ denote the true and surrogate exposures, let $f(\boldsymbol{x})$ and $f(\boldsymbol{z})$ denote marginal distributions of true and surrogate exposures, let $f(\boldsymbol{z}|\boldsymbol{x})$ and $f(\boldsymbol{x}|\boldsymbol{z})$ denote corresponding conditional distributions,

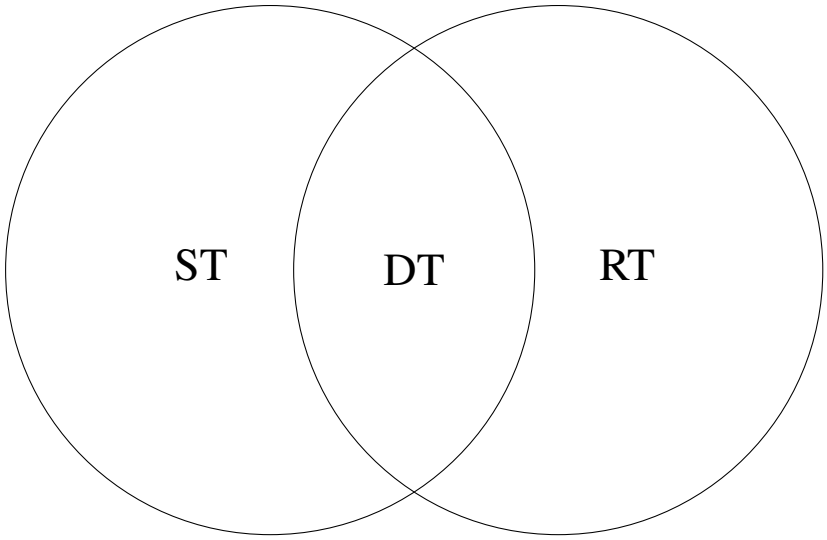


Figure 3.1: A Venn Diagram for Claimed Relationships between ST, RT, and DT.

and let $f(\boldsymbol{x}, \boldsymbol{z})$ denote the joint distribution of the true and surrogate exposures. Here we reiterate the definitions of ST, RT, and DT:

- ST: *Single Transportability*. $f(\boldsymbol{z}|\boldsymbol{x})$ is transportable between the validation study (VS) sample and the main study (MS) sample, but no transportability assumption is placed on $f(\boldsymbol{x})$.
- RT:*Reclassification Transportability*. $f(\boldsymbol{x}|\boldsymbol{z})$ is transportable between VS and MS, but no transportability assumption is placed on $f(\boldsymbol{z})$.
- DT: *Double Transportability*. $f(\boldsymbol{x}, \boldsymbol{z})$ is transportable between VS and MS.

We show relationships between ST, RT, and DT are as claimed in the Venn diagrams above.

1. $DT \subseteq ST$

Proof. Self-evident from the definition. □

2. $DT \subseteq RT$

Proof. Self-evident from the definition. □

3. $DT = ST \cap RT$

Proof. Since $DT \subseteq RT$ and $DT \subseteq ST$, we obtain $RT \cap ST \neq \emptyset$ and $DT \subseteq RT \cap ST$. Next, to show the converse, we will show that if ST and RT both hold, which means both $f(\boldsymbol{z}|\boldsymbol{x})$ and $f(\boldsymbol{x}|\boldsymbol{z})$ are transportable between VS and MS, then $f(\boldsymbol{x})$ is also transportable. As a general matter, $f(\boldsymbol{x})f(\boldsymbol{z}|\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{z})f(\boldsymbol{x}|\boldsymbol{z})$, which gives

$$\frac{f(\boldsymbol{z}|\boldsymbol{x})}{f(\boldsymbol{x}|\boldsymbol{z})} = \frac{f(\boldsymbol{z})}{f(\boldsymbol{x})},$$

and

$$\frac{1}{f(\boldsymbol{x})} = \frac{\int f(\boldsymbol{z})\mathrm{d}\boldsymbol{z}}{f(\boldsymbol{x})} = \int \frac{f(\boldsymbol{z})}{f(\boldsymbol{x})}\mathrm{d}\boldsymbol{z} = \int \frac{f(\boldsymbol{z}|\boldsymbol{x})}{f(\boldsymbol{x}|\boldsymbol{z})}\mathrm{d}\boldsymbol{z},$$

so that the marginal $f(\boldsymbol{x})$ is determined by the two conditional distributions $f(\boldsymbol{z}|\boldsymbol{x})$ and $f(\boldsymbol{x}|\boldsymbol{z})$. Thus, the transportability of $f(\boldsymbol{z}|\boldsymbol{x})$ and $f(\boldsymbol{x}|\boldsymbol{z})$ implies that of $f(\boldsymbol{x})$, and therefore implies double transportability. □

4. $RT - ST \neq \emptyset$

Proof. Let $f_m(\boldsymbol{x}, \boldsymbol{z})$ and $f_v(\boldsymbol{x}, \boldsymbol{z})$ denote the joint distributions in the main and validation studies. We want to show the existence of an example where RT holds but ST does not. For simplicity assume both joint densities are positive. Assume $f_m(\boldsymbol{x}|\boldsymbol{z}) = f_v(\boldsymbol{x}|\boldsymbol{z})$ so that RT holds, and denote the common conditional distribution by $f(\boldsymbol{x}|\boldsymbol{z})$. Let $r(\boldsymbol{z}) = f_v(\boldsymbol{z})/f_m(\boldsymbol{z})$ denote the ratio of the marginal densities of $\boldsymbol{z}$ in the validation and main studies; for example, this situation could be realized if the validation study participants were sampled from the main study participants with probabilities proportional to $r(\boldsymbol{z})$. Unless $r(\boldsymbol{z}) = 1$ for all $\boldsymbol{z}$, in which case double transportability holds, the function r must take at least two different values (at least one larger than 1 and one smaller than 1). Let $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ be such that $r(\boldsymbol{z}_1) > r(\boldsymbol{z}_2)$, so that $\frac{f_v(\boldsymbol{z}_1)}{f_m(\boldsymbol{z}_1)} > \frac{f_v(\boldsymbol{z}_2)}{f_m(\boldsymbol{z}_2)}$, which is equivalent to $\frac{f_v(\boldsymbol{z}_1)}{f_v(\boldsymbol{z}_2)} > \frac{f_m(\boldsymbol{z}_1)}{f_m(\boldsymbol{z}_2)}$. Then for

arbitrary $\boldsymbol{x}$ we have

$$\frac{f_v(\boldsymbol{z}_1|\boldsymbol{x})}{f_v(\boldsymbol{z}_2|\boldsymbol{x})} = \frac{f_v(\boldsymbol{z}_1)f(\boldsymbol{x}|\boldsymbol{z}_1)}{f_v(\boldsymbol{z}_2)f(\boldsymbol{x}|\boldsymbol{z}_2)} > \frac{f_m(\boldsymbol{z}_1)f(\boldsymbol{x}|\boldsymbol{z}_1)}{f_m(\boldsymbol{z}_2)f(\boldsymbol{x}|\boldsymbol{z}_2)} = \frac{f_m(\boldsymbol{z}_1|\boldsymbol{x})}{f_m(\boldsymbol{z}_2|\boldsymbol{x})},$$

so that the conditional distribution of $\boldsymbol{Z}|\boldsymbol{X}$ in the validation study is not the same as that in the main study, and ST does not hold. □

5. $ST - RT \neq \emptyset$

Proof. This follows from the proof of $RT - ST \neq \emptyset$ with the roles of $\boldsymbol{X}$ and $\boldsymbol{Z}$ interchanged. □

3.7.2 Tables and Figures

Table 3.8: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 100, 250$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 4, *Bayes* = proposed. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease) and $\beta_0 = -4$ (rare disease), multivariate normal true reclassification model, and the low measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators									
			$\beta_0 = -2$					$\beta_0 = -4$				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*	*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
100	pPIF_{X_1}	BIAS	82.60	1.05	-5.07	1.91	1.31	75.98	0.75	-6.04	5.27	1.46
		MCSE	0.005	0.030	0.027	0.035	0.022	0.006	0.038	0.039	0.063	0.033
		ASE	0.004	0.033	0.030	0.038	0.020	0.006	0.045	0.039	0.076	0.029
		RMSE	0.128	0.030	0.029	0.036	0.022	0.158	0.039	0.041	0.066	0.035
		Coverage	0.00	**94.90**	**93.90**	**96.00**	90.50	0.00	**94.20**	92.80	96.60	91.20
	pPIF_{X_2}	BIAS	170.84	-3.79	-10.48	3.23	-0.77	159.13	-6.53	-10.90	6.53	-2.79
		MCSE	0.005	0.037	0.033	0.039	0.026	0.007	0.050	0.048	0.075	0.043
		ASE	0.004	0.040	0.036	0.042	0.024	0.007	0.058	0.048	0.087	0.038
		RMSE	0.135	0.038	0.035	0.040	0.026	0.173	0.052	0.050	0.077	0.044
		Coverage	0.00	**94.30**	**93.60**	**95.10**	91.40	0.00	**95.20**	92.50	**94.10**	91.00
	$\text{pPIF}_{X_1X_2}$	BIAS	101.64	-0.02	-6.30	2.46	0.81	87.65	-0.37	-6.43	6.19	0.65
		MCSE	0.005	0.012	0.019	0.030	0.011	0.006	0.015	0.026	0.052	0.015
		ASE	0.004	0.014	0.020	0.034	0.011	0.005	0.018	0.027	0.067	0.014
		RMSE	0.232	0.012	0.024	0.032	0.013	0.263	0.016	0.032	0.057	0.017
		Coverage	0.00	**95.00**	88.10	**96.30**	92.20	0.00	**94.10**	89.30	97.40	92.90
250	pPIF_{X_1}	BIAS	82.44	-0.12	-7.18	1.06	0.91	75.76	0.16	-7.17	1.83	0.90
		MCSE	0.004	0.018	0.017	0.022	0.013	0.007	0.023	0.022	0.035	0.019
		ASE	0.004	0.018	0.018	0.022	0.013	0.006	0.024	0.024	0.036	0.019
		RMSE	0.128	0.018	0.020	0.022	0.014	0.158	0.023	0.027	0.035	0.020
		Coverage	0.00	**95.60**	89.50	93.40	**94.20**	0.00	**95.70**	91.00	**95.00**	**94.30**
	pPIF_{X_2}	BIAS	170.43	-0.97	-6.58	0.61	-0.71	161.11	-1.98	-9.04	-0.20	-1.49
		MCSE	0.005	0.022	0.020	0.025	0.016	0.008	0.030	0.027	0.042	0.024
		ASE	0.005	0.022	0.021	0.025	0.016	0.007	0.031	0.029	0.042	0.024
		RMSE	0.135	0.022	0.021	0.024	0.017	0.175	0.030	0.029	0.042	0.024
		Coverage	0.00	**95.20**	**95.20**	**94.00**	**93.60**	0.00	**94.50**	**95.50**	**93.60**	**94.00**
	$\text{pPIF}_{X_1X_2}$	BIAS	101.43	-0.28	-6.64	0.97	0.45	88.07	-0.26	-7.06	1.53	0.33
		MCSE	0.005	0.008	0.011	0.017	0.007	0.006	0.010	0.016	0.028	0.009
		ASE	0.004	0.008	0.012	0.019	0.007	0.005	0.010	0.016	0.029	0.009
		RMSE	0.231	0.008	0.019	0.017	0.010	0.264	0.010	0.027	0.028	0.012
		Coverage	0.00	**95.00**	76.80	**95.60**	93.00	0.00	**94.00**	76.10	**94.30**	**94.90**

Table 3.9: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 500, 1000$). ***UN*** = uncorrected, $RC_{M/I}$, RC_I, and ***IVS*** are defined in Section 4, ***Bayes*** = proposed. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease) and $\beta_0 = -4$ (rare disease), multivariate normal true reclassification model, and the low measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators									
			$\beta_0 = -2$					$\beta_0 = -4$				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*	*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
500	pPIF_{X_1}	BIAS	82.56	-0.13	-6.86	0.03	0.29	75.64	-0.44	-7.51	0.22	0.07
		MCSE	0.005	0.013	0.013	0.015	0.010	0.007	0.018	0.018	0.024	0.013
		ASE	0.004	0.013	0.013	0.015	0.010	0.006	0.018	0.018	0.024	0.014
		RMSE	0.128	0.013	0.017	0.015	0.009	0.158	0.018	0.024	0.024	0.014
		Coverage	0.00	**95.10**	87.10	**95.30**	93.00	0.00	**94.70**	84.40	**94.90**	**96.00**
	pPIF_{X_2}	BIAS	170.84	-0.93	-7.65	0.96	-0.15	160.66	-1.36	-8.27	-0.83	-0.01
		MCSE	0.005	0.016	0.015	0.016	0.011	0.007	0.023	0.023	0.027	0.017
		ASE	0.005	0.016	0.016	0.017	0.012	0.007	0.023	0.022	0.028	0.018
		RMSE	0.134	0.016	0.017	0.017	0.011	0.174	0.023	0.025	0.028	0.017
		Coverage	0.00	**95.00**	91.90	**94.40**	**94.70**	0.00	**94.30**	91.40	**94.10**	**93.90**
	$\text{pPIF}_{X_1X_2}$	BIAS	101.63	-0.34	-6.80	0.41	0.19	87.83	-0.59	-7.11	0.11	0.15
		MCSE	0.005	0.006	0.008	0.012	0.005	0.006	0.007	0.012	0.018	0.007
		ASE	0.004	0.006	0.008	0.013	0.005	0.005	0.008	0.012	0.019	0.007
		RMSE	0.231	0.006	0.017	0.014	0.010	0.264	0.008	0.024	0.017	0.008
		Coverage	0.00	**95.20**	57.70	**94.90**	**95.80**	0.00	**95.00**	58.00	**95.70**	**95.80**
1000	pPIF_{X_1}	BIAS	82.93	-0.66	-6.93	0.65	0.22	76.04	-1.11	-8.11	0.35	0.26
		MCSE	0.004	0.010	0.010	0.010	0.007	0.006	0.015	0.014	0.017	0.011
		ASE	0.004	0.010	0.010	0.011	0.007	0.006	0.015	0.014	0.017	0.011
		RMSE	0.128	0.010	0.015	0.010	0.006	0.159	0.015	0.022	0.017	0.011
		Coverage	0.00	92.70	82.20	**94.60**	**96.30**	0.00	**94.00**	79.00	**94.30**	**94.80**
	pPIF_{X_2}	BIAS	170.03	-0.73	-7.52	-0.25	0.05	159.45	-0.62	-8.02	0.32	-0.51
		MCSE	0.005	0.012	0.012	0.012	0.008	0.007	0.019	0.017	0.020	0.014
		ASE	0.004	0.013	0.012	0.012	0.009	0.007	0.019	0.018	0.020	0.014
		RMSE	0.134	0.013	0.013	0.006	0.004	0.173	0.019	0.020	0.020	0.014
		Coverage	0.00	**94.70**	92.70	**95.00**	**95.30**	0.00	**94.30**	92.20	**95.70**	**95.10**
	$\text{pPIF}_{X_1X_2}$	BIAS	101.64	-0.65	-6.84	0.37	0.18	87.77	-0.85	-7.48	0.42	0.07
		MCSE	0.004	0.004	0.006	0.008	0.004	0.005	0.006	0.008	0.013	0.005
		ASE	0.004	0.005	0.006	0.009	0.004	0.005	0.006	0.009	0.013	0.005
		RMSE	0.231	0.005	0.017	0.008	0.007	0.263	0.006	0.024	0.014	0.010
		Coverage	0.00	**94.50**	30.50	**96.30**	**94.50**	0.00	**93.60**	26.00	**95.60**	**95.60**

Table 3.10: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 100, 250$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 4, *Bayes* = proposed. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease) and $\beta_0 = -4$ (rare disease), multivariate normal true reclassification model, and the high measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators									
			$\beta_0 = -2$					$\beta_0 = -4$				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*	*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
100	pPIF_{X_1}	BIAS	-51.94	-29.52	-26.11	0.89	1.80	-50.07	-95.95	-26.14	-13.98	5.70
		MCSE	0.010	0.147	0.068	0.075	0.058	0.0254	0.405	0.093	0.228	0.138
		ASE	0.010	0.507	0.128	0.078	0.050	0.0231	1.963	0.166	0.339	0.080
		RMSE	0.164	0.173	0.102	0.076	0.060	0.189	0.542	0.133	0.246	0.150
		Coverage	0.00	98.30	**95.90**	93.00	88.10	0.00	99.20	**96.00**	90.40	74.90
	pPIF_{X_2}	BIAS	-52.46	-100.65	-43.93	-0.52	-2.50	-52.55	-326.10	-49.28	-57.32	-15.13
		MCSE	0.011	0.218	0.104	0.094	0.081	0.028	0.686	0.152	0.344	0.235
		ASE	0.011	0.682	0.189	0.096	0.072	0.026	2.634	0.263	0.508	0.131
		RMSE	0.089	0.273	0.126	0.096	0.082	0.113	0.949	0.181	0.392	0.252
		Coverage	0.00	98.60	98.50	93.30	89.40	0.00	98.80	92.00	99.00	73.80
	$\text{pPIF}_{X_1X_2}$	BIAS	-49.53	-22.19	-27.30	1.10	1.21	-47.16	-67.45	-25.40	-13.49	3.96
		MCSE	0.011	0.133	0.066	0.081	0.055	0.027	0.386	0.090	0.259	0.128
		ASE	0.012	0.495	0.132	0.084	0.051	0.027	1.900	0.172	0.347	0.075
		RMSE	0.219	0.171	0.130	0.082	0.057	0.240	0.524	0.154	0.294	0.139
		Coverage	0.00	98.10	83.90	**93.70**	90.00	0.00	98.60	92.40	93.10	74.80
250	pPIF_{X_1}	BIAS	-51.28	-9.12	-20.56	-0.08	0.27	-50.62	-24.06	-19.01	-1.32	2.57
		MCSE	0.010	0.098	0.059	0.044	0.036	0.024	0.194	0.085	0.108	0.069
		ASE	0.010	0.206	0.076	0.044	0.034	0.023	0.652	0.104	0.132	0.062
		RMSE	0.162	0.104	0.085	0.045	0.036	0.191	0.229	0.110	0.111	0.072
		Coverage	0.00	**95.30**	88.30	**94.90**	92.20	0.00	**96.30**	92.00	92.90	89.60
	pPIF_{X_2}	BIAS	-53.67	-28.45	-23.97	-1.11	0.95	-51.89	-83.25	-23.80	-1.79	1.02
		MCSE	0.011	0.163	0.097	0.056	0.049	0.025	0.334	0.141	0.151	0.111
		ASE	0.011	0.311	0.122	0.055	0.047	0.026	0.983	0.173	0.171	0.098
		RMSE	0.092	0.174	0.106	0.057	0.051	0.110	0.384	0.150	0.158	0.115
		Coverage	0.00	**95.20**	**95.10**	**94.00**	92.40	0.00	**96.40**	**95.50**	92.30	90.10
	$\text{pPIF}_{X_1X_2}$	BIAS	-49.50	-4.58	-18.04	-0.13	0.78	-47.33	-13.51	-14.33	0.75	3.34
		MCSE	0.011	0.080	0.049	0.047	0.035	0.027	0.159	0.065	0.113	0.065
		ASE	0.012	0.188	0.069	0.047	0.034	0.026	0.599	0.091	0.129	0.056
		RMSE	0.218	0.086	0.089	0.048	0.036	0.241	0.192	0.097	0.115	0.069
		Coverage	0.00	97.00	74.50	**94.20**	92.10	0.00	97.10	89.00	**93.80**	88.20

Table 3.11: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 500, 1000$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 4, *Bayes* = proposed. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease) and $\beta_0 = -4$ (rare disease), multivariate normal true reclassification model, and the high measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators									
			$\beta_0 = -2$					$\beta_0 = -4$				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*	*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
500	pPIF_{X_1}	BIAS	-51.56	-3.64	-17.47	0.16	-0.24	-49.45	-5.39	-13.49	0.02	0.69
		MCSE	0.009	0.059	0.045	0.029	0.024	0.023	0.086	0.068	0.068	0.048
		ASE	0.010	0.080	0.049	0.031	0.024	0.023	0.150	0.075	0.071	0.047
		RMSE	0.163	0.062	0.068	0.030	0.024	0.187	0.092	0.084	0.069	0.049
		Coverage	0.00	**95.30**	82.50	**95.90**	**95.00**	0.00	**95.80**	93.00	**93.80**	**93.70**
	pPIF_{X_2}	BIAS	-53.46	-4.58	-20.50	-0.53	-0.43	-52.21	-15.48	-21.17	-1.50	0.01
		MCSE	0.010	0.104	0.078	0.037	0.033	0.024	0.167	0.122	0.089	0.073
		ASE	0.010	0.130	0.083	0.038	0.034	0.025	0.263	0.134	0.092	0.073
		RMSE	0.091	0.108	0.085	0.037	0.035	0.111	0.178	0.133	0.090	0.075
		Coverage	0.00	**95.20**	**93.60**	**95.00**	**94.50**	0.01	**95.90**	**95.00**	**93.80**	93.10
	$\text{pPIF}_{X_1X_2}$	BIAS	-49.62	-1.56	-15.72	0.08	-0.12	-46.65	-2.14	-11.19	0.41	1.12
		MCSE	0.011	0.052	0.038	0.032	0.024	0.026	0.068	0.053	0.068	0.046
		ASE	0.011	0.068	0.041	0.032	0.025	0.025	0.128	0.061	0.071	0.044
		RMSE	0.219	0.054	0.075	0.032	0.024	0.238	0.071	0.077	0.070	0.047
		Coverage	0.00	**95.40**	58.60	**94.60**	**95.10**	0.00	**95.40**	87.40	93.50	92.10
1000	pPIF_{X_1}	BIAS	-51.87	-2.89	-15.67	0.43	0.31	-49.34	-3.98	-12.04	-0.33	0.39
		MCSE	0.010	0.042	0.034	0.021	0.018	0.023	0.067	0.056	0.049	0.036
		ASE	0.010	0.043	0.036	0.021	0.018	0.022	0.070	0.061	0.047	0.036
		RMSE	0.164	0.043	0.057	0.022	0.017	0.186	0.070	0.071	0.049	0.037
		Coverage	0.00	93.00	74.80	**93.80**	**95.60**	0.00	**94.20**	90.80	93.50	**93.80**
	pPIF_{X_2}	BIAS	-53.09	-5.48	-18.51	-1.06	-1.06	-53.17	-9.50	-18.27	-1.74	0.42
		MCSE	0.011	0.074	0.060	0.027	0.024	0.027	0.127	0.103	0.061	0.053
		ASE	0.010	0.076	0.062	0.027	0.025	0.025	0.131	0.111	0.062	0.054
		RMSE	0.091	0.075	0.066	0.026	0.0244	0.114	0.131	0.110	0.062	0.055
		Coverage	0.00	**94.70**	92.50	**94.60**	**94.40**	0.01	**94.20**	**95.10**	**94.60**	**95.20**
	$\text{pPIF}_{X_1X_2}$	BIAS	-49.70	-2.98	-14.42	0.04	-0.04	-46.89	-2.66	-10.23	-0.25	0.68
		MCSE	0.012	0.036	0.027	0.022	0.018	0.026	0.053	0.044	0.047	0.033
		ASE	0.011	0.039	0.029	0.022	0.018	0.025	0.057	0.048	0.047	0.034
		RMSE	0.219	0.038	0.065	0.022	0.017	0.239	0.056	0.067	0.047	0.033
		Coverage	0.00	**94.30**	40.90	**95.20**	**94.70**	0.00	**94.80**	84.10	93.40	**94.10**

Table 3.12: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of the uncorrected and proposed Bayesian estimators for estimating the three pPIFs, with different validation study size. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease), multivariate gamma true reclassification model, and the low measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

N_v	Estimand	Method	Bias	MCSE	ASE	RMSE	Coverage
100	PIF_{X_1}	Uncorrected	82.32	0.005	0.004	0.134	0.00
		Bayesian	0.06	0.018	0.018	0.020	**94.00**
	PIF_{X_2}	Uncorrected	169.27	0.005	0.005	0.138	0.00
		Bayesian	-0.05	0.022	0.022	0.028	**94.70**
	PIF_{X_1,X_2}	Uncorrected	100.54	0.004	0.004	0.239	0.00
		Bayesian	0.22	0.009	0.009	0.010	**94.00**
500	PIF_{X_1}	Uncorrected	82.57	0.005	0.004	0.134	0.00
		Bayesian	0.36	0.009	0.009	0.010	**95.00**
	PIF_{X_2}	Uncorrected	169.77	0.005	0.004	0.138	0.00
		Bayesian	-0.65	0.011	0.011	0.010	**94.10**
	PIF_{X_1,X_2}	Uncorrected	100.84	0.005	0.003	0.239	0.00
		Bayesian	0.09	0.004	0.004	0.010	**94.80**

Table 3.13: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of the uncorrected and proposed Bayesian estimators for estimating the three pPIFs, with different validation study size. The simulation results are based on 1000 data replications, under $\beta_0 = -4$ (rare disease), multivariate gamma true reclassification model, and the low measurement error scenario. Coverage percentage between 93.6% and 96.4% bold font are within the margin of error for 1000 replications.

N_v	Estimand	Method	Bias	MCSE	ASE	RMSE	Coverage
100	PIF_{X_1}	Uncorrected	75.46	0.007	0.006	0.164	0.00
		Bayesian	-0.23	0.025	0.025	0.026	93.40
	PIF_{X_2}	Uncorrected	159.60	0.007	0.007	0.182	0.00
		Bayesian	0.33	0.032	0.032	0.033	**94.20**
	PIF_{X_1,X_2}	Uncorrected	86.83	0.007	0.004	0.270	0.00
		Bayesian	0.35	0.011	0.011	0.010	**95.20**
500	PIF_{X_1}	Uncorrected	75.24	0.007	0.006	0.164	0.00
		Bayesian	0.50	0.013	0.013	0.014	**94.30**
	PIF_{X_2}	Uncorrected	159.15	0.008	0.007	0.179	0.00
		Bayesian	-1.09	0.016	0.017	0.017	**94.70**
	PIF_{X_1,X_2}	Uncorrected	86.57	0.006	0.004	0.270	0.00
		Bayesian	0.12	0.006	0.006	0.010	**93.60**

Table 3.14: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of the uncorrected and proposed Bayesian estimators for estimating the three pPIFs, with different validation study size. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease), multivariate gamma true reclassification model, and the high measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

N_v	Estimand	Method	Bias	MCSE	ASE	RMSE	Coverage
100	PIF_{X_1}	Uncorrected	-51.39	0.011	0.010	0.158	0.00
		Bayesian	4.34	0.063	0.058	0.063	91.30
	PIF_{X_2}	Uncorrected	-53.56	0.012	0.011	0.089	0.00
		Bayesian	-6.11	0.105	0.093	0.110	89.90
	PIF_{X_1,X_2}	Uncorrected	-49.56	0.013	0.012	0.212	0.00
		Bayesian	2.36	0.063	0.058	0.063	91.80
500	PIF_{X_1}	Uncorrected	-51.44	0.010	0.010	0.158	0.00
		Bayesian	1.75	0.028	0.028	0.032	93.10
	PIF_{X_2}	Uncorrected	-53.54	0.010	0.011	0.089	0.00
		Bayesian	1.76	0.043	0.042	0.045	93.10
	PIF_{X_1,X_2}	Uncorrected	-49.59	0.012	0.012	0.212	0.00
		Bayesian	1.92	0.028	0.028	0.032	93.20

Table 3.15: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of the uncorrected and proposed Bayesian estimators for estimating the three pPIFs, with different validation study size. The simulation results are based on 1000 data replications, under $\beta_0 = -4$ (rare disease), multivariate gamma true reclassification model, and the high measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

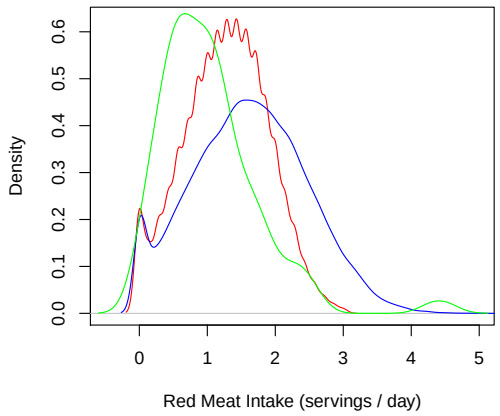
N_v	Estimand	Method	Bias	MCSE	ASE	RMSE	Coverage
100	PIF_{X_1}	Uncorrected	-50.61	0.023	0.022	0.184	0.00
		Bayesian	8.20	0.150	0.088	0.179	79.80
	PIF_{X_2}	Uncorrected	-51.60	0.026	0.025	0.105	0.00
		Bayesian	-63.86	0.392	0.180	0.473	78.70
	PIF_{X_1,X_2}	Uncorrected	-47.44	0.025	0.026	0.235	0.00
		Bayesian	-4.22	0.185	0.091	0.217	80.80
500	PIF_{X_1}	Uncorrected	-50.40	0.022	0.022	0.184	0.00
		Bayesian	0.90	0.047	0.049	0.055	**94.90**
	PIF_{X_2}	Uncorrected	-53.31	0.027	0.025	0.110	0.00
		Bayesian	-4.89	0.084	0.085	0.089	**95.00**
	PIF_{X_1,X_2}	Uncorrected	-47.87	0.025	0.026	0.237	0.00
		Bayesian	0.02	0.047	0.047	0.045	**93.70**

Table 3.16: Bayesian estimation results from reclassification model parameters ($\mathbf{\Gamma}^*$), illustrative example.

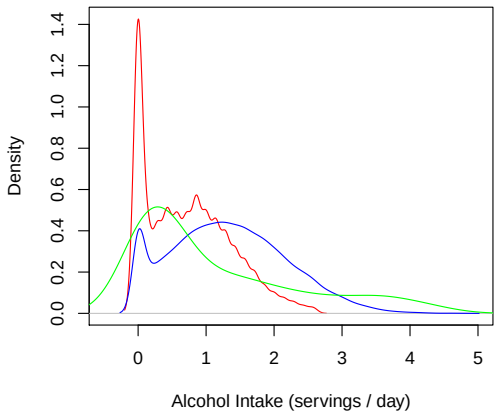
	α	Z_1	Z_2	Z_3	X_4	X_5
X_1	0.696	0.991	-0.056	-0.714	0.022	-0.008
X_2	0.425	0.297	0.830	-0.411	-0.011	-0.043
X_3	0.251	-0.061	-0.030	0.503	0.003	-0.004

Table 3.17: Bayesian estimation results from reclassification model parameters ($\mathbf{\Sigma}_{\boldsymbol{x}}$), the HPFS.

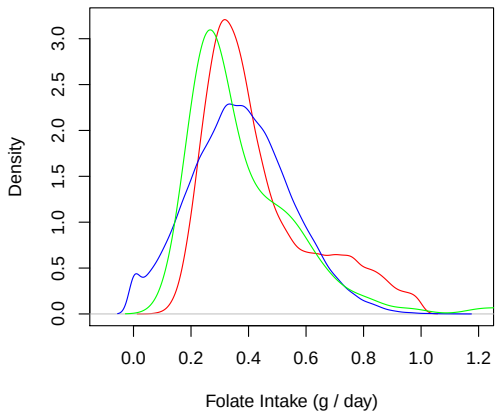
	X_1	X_2	X_3
X_1	0.284	0.039	-0.004
X_2	0.039	0.358	0.007
X_3	-0.004	0.007	0.020



(a) Red Meat Intake



(b) Alcohol Intake



(c) Folate Intake

Figure 3.2: An illustration of empirical densities for red meat intake, alcohol intake, and folate intake obtained from the proposed Bayesian approach. Red lines represent surrogate exposures. Blue lines represent imputed true exposures. Green lines represent true exposures in the internal validation study.

3.7.3 A Simulation Study Investigating Performances of RC Estimators

We conducted a simulation study to further investigate the performances of RC estimators. As suggested in Carroll et al. (2006), RC methods give suboptimal estimates of logistic conditional disease probability model coefficients when the under one or more of the following conditions:

1. The rare disease condition is violated;
2. Large measurement error;
3. Prognostic effects of the exposure are moderate.

In our main simulation setting, where the conditional disease probability model coefficients, $\boldsymbol{\beta} = (\beta_0, 1, 0.5, 0.5)'$, where $\beta_0 = -2$ (base disease prevalence 12 %) or -4 (base disease prevalence 2 %). This coefficient setting is unfavorable towards RC methods, because the relative risk with respect to X_1 and X_2 are 1.859 and 1.324 when $\beta_0 = -2$ (2.513 and 1.571 when $\beta_0 = -4$), which are considerably substantial. We thus consider a more favorable simulation setting for RC estimators, where $\boldsymbol{\beta} = (-4, 0.25, 0.125, 0.125)'$. This reduces the relative risks to 1.276 and 1.129. We follow the same settings for the generating distributions of surrogate exposures and the reclassification process coefficients. We considered two different validation study sizes, 250 and 1000, and three different amount of measurement errors. Results are given in Tables 3.18 to 3.20.

under more favorable conditions, RC estimators exhibit improved performances in terms of bias, the RMSE, and coverage probability. RC methods show comparable estimation accuracy and efficiency compared to the Bayesian estimator under the low measurement error setting. As measurement error amount increases, Bayesian estimator starts to outperform RC estimators, as it shows less bias and smaller RMSE. Hence, compared to RC estimators,

the proposed Bayesian estimator shows better stability under various scenarios. However, RC estimators are easier to implement since they do not require a long running time.

Table 3.18: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 250, 1000$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 4, *Bayes* = proposed. The simulation results are based on 1000 data replications, multivariate normal true reclassification model, and the low measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
250	pPIF_{X_1}	BIAS	82.46	-1.82	-3.11	-13.73	-0.36
		MCSE	0.024	0.041	0.038	0.151	0.017
		ASE	0.024	0.041	0.040	0.191	0.017
		RMSE	0.096	0.042	0.039	0.159	0.019
		Coverage	4.00	**95.10**	**94.80**	91.70	**95.00**
	pPIF_{X_2}	BIAS	162.27	-2.98	-2.82	-72.55	0.88
		MCSE	0.025	0.051	0.049	0.206	0.039
		ASE	0.026	0.051	0.050	0.261	0.038
		RMSE	0.098	0.052	0.50	0.228	0.039
		Coverage	4.00	**94.10**	**95.40**	90.80	**95.60**
	$\text{pPIF}_{X_1X_2}$	BIAS	99.24	0.019	-0.75	-3.40	-9.46
		MCSE	0.025	0.017	0.017	0.123	0.049
		ASE	0.026	0.018	0.017	0.142	0.048
		RMSE	0.166	0.017	0.017	0.133	0.050
		Coverage	0.00	**94.50**	**95.60**	91.60	**94.40**
1000	pPIF_{X_1}	BIAS	80.21	-1.72	-0.47	-2.17	0.31
		MCSE	0.024	0.039	0.037	0.061	0.016
		ASE	0.022	0.038	0.038	0.061	0.015
		RMSE	0.094	0.040	0.037	0.062	0.016
		Coverage	5.00	92.70	**94.80**	**93.70**	**94.60**
	pPIF_{X_2}	BIAS	167.57	-3.23	-1.28	-5.44	-2.05
		MCSE	0.027	0.048	0.045	0.074	0.032
		ASE	0.024	0.047	0.047	0.073	0.032
		RMSE	0.101	0.049	0.046	0.075	0.032
		Coverage	4.00	92.70	**95.30**	92.90	**95.80**
	$\text{pPIF}_{X_1X_2}$	BIAS	99.39	-0.11	-1.34	0.89	1.02
		MCSE	0.028	0.016	0.016	0.045	0.039
		ASE	0.024	0.017	0.016	0.045	0.039
		RMSE	0.166	0.016	0.016	0.045	0.040
		Coverage	0.00	**95.90**	**95.10**	93.50	**94.80**

Table 3.19: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 250, 1000$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 4, *Bayes* = proposed. The simulation results are based on 1000 data replications, multivariate normal true reclassification model, and the moderate measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
250	pPIF_{X_1}	BIAS	-31.09	-24.30	-3.72	-50.81	-0.67
		MCSE	0.032	0.123	0.084	0.270	0.081
		ASE	0.032	0.284	0.094	0.328	0.066
		RMSE	0.048	0.145	0.086	0.298	0.107
		Coverage	77.00	**95.20**	**95.10**	88.80	**94.60**
	pPIF_{X_2}	BIAS	-25.02	-90.12	-40.70	-131.28	-4.73
		MCSE	0.035	0.195	0.131	0.299	0.091
		ASE	0.033	0.395	0.143	0.373	0.079
		RMSE	0.038	0.228	0.136	0.327	0.103
		Coverage	88.00	**95.40**	**95.00**	87.70	93.20
	$\text{pPIF}_{X_1X_2}$	BIAS	-27.96	-8.95	-3.74	-40.82	-20.63
		MCSE	0.038	0.083	0.067	0.293	0.127
		ASE	0.038	0.191	0.074	0.344	0.115
		RMSE	0.061	0.087	0.068	0.332	0.137
		Coverage	79.00	**96.00**	**95.40**	89.50	93.20
1000	pPIF_{X_1}	BIAS	-24.39	-4.86	-2.54	-7.86	-3.71
		MCSE	0.031	0.081	0.079	0.094	0.048
		ASE	0.031	0.082	0.079	0.092	0.051
		RMSE	0.041	0.083	0.080	0.096	0.049
		Coverage	78.00	**94.80**	**94.90**	92.70	**96.40**
	pPIF_{X_2}	BIAS	-32.92	-27.72	-23.37	-15.42	-0.88
		MCSE	0.035	0.120	0.117	0.108	0.054
		ASE	0.032	0.124	0.118	0.105	0.055
		RMSE	0.040	0.123	0.119	0.111	0.054
		Coverage	86.00	**94.80**	**94.40**	91.60	**96.00**
	$\text{pPIF}_{X_1X_2}$	BIAS	-26.13	-2.54	-3.74	-5.97	-20.76
		MCSE	0.036	0.065	0.062	0.108	0.078
		ASE	0.038	0.065	0.064	0.104	0.076
		RMSE	0.057	0.066	0.064	0.109	0.080
		Coverage	76.00	**94.90**	**95.10**	92.90	**94.40**

Table 3.20: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of different estimators for estimating the three pPIFs, with different validation study size ($N_v = 250, 1000$). *UN* = uncorrected, $RC_{M/I}$, RC_I, and *IVS* are defined in Section 4, *Bayes* = proposed. The simulation results are based on 1000 data replications, multivariate normal true reclassification model, and the high measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

			Estimators				
N_v	pPIF		*UN*	$RC_{M/I}$	RC_I	*IVS*	*Bayes*
250	pPIF_{X_1}	BIAS	-52.53	-92.98	-18.45	-63.81	-17.76
		MCSE	0.034	0.250	0.112	0.302	0.142
		ASE	0.033	0.904	0.136	0.347	0.112
		RMSE	0.069	0.301	0.116	0.333	0.180
		Coverage	45.00	96.50	97.60	87.50	90.60
	pPIF_{X_2}	BIAS	-57.76	-364.98	-55.04	-145.53	-6.96
		MCSE	0.037	0.433	0.174	0.331	0.116
		ASE	0.034	1.324	0.217	0.408	0.102
		RMSE	0.050	0.518	0.179	0.357	0.123
		Coverage	81.00	**96.30**	96.50	87.70	91.60
	$\text{pPIF}_{X_1X_2}$	BIAS	-53.04	-50.73	-11.32	-56.12	-88.29
		MCSE	0.044	0.206	0.112	0.338	0.208
		ASE	0.041	0.821	0.136	0.403	0.155
		RMSE	0.099	0.235	0.115	0.372	0.249
		Coverage	38.00	96.50	96.70	86.40	91.60
1000	pPIF_{X_1}	BIAS	-53.17	-13.57	-13.95	-6.10	-2.76
		MCSE	0.033	0.112	0.108	0.099	0.071
		ASE	0.032	0.118	0.111	0.101	0.075
		RMSE	0.070	0.115	0.112	0.101	0.074
		Coverage	51.00	**95.00**	**94.80**	**94.40**	**95.20**
	pPIF_{X_2}	BIAS	-60.52	-63.28	-37.85	-28.64	-5.23
		MCSE	0.038	0.192	0.172	0.114	0.063
		ASE	0.033	0.196	0.173	0.113	0.067
		RMSE	0.052	0.200	0.178	0.117	0.066
		Coverage	77.00	**95.40**	**94.60**	93.50	**95.80**
	$\text{pPIF}_{X_1X_2}$	BIAS	-54.28	-10.75	-9.57	-9.50	-8.77
		MCSE	0.040	0.118	0.108	0.119	0.088
		ASE	0.040	0.118	0.110	0.121	0.090
		RMSE	0.099	0.122	0.111	0.122	0.091
		Coverage	37.00	**93.70**	**95.80**	**93.80**	**94.00**

3.7.4 A Simulation Study when the Internal Validation Study is a Biased Sample from the Main Study

We included an additional simulation study to examine the operating characteristics of the Bayesian estimator when internal validation study (IVS) is a biased sample from the entire population. Even though the IVS is not uniformly sampled from the main study, the reclassification transportability still holds (while neither the single nor double transportability conditions may not hold), and our estimator in theory should still be effective in accurately estimating the pPIF in the presence of exposure measurement errors. This simulation study aims to confirm this analytical conjecture, and the results may endorse the application of our approach in general MS/IVS designs, where the IVS is not a random sample from the MS.

We conducted the simulation by first generating surrogate exposures $(Z_{i1}, Z_{i2}) \sim \mathcal{N}((-0.2, 0.8)^T, \boldsymbol{\Sigma_z})$, where

$$\boldsymbol{\Sigma_z} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}.$$

Each observation $i = 1, ..., N$ is independently selected into the IVS according to a Bernoulli variable T_i with

$$\mathbb{P}(T_i = 1 | Z_{i1}, Z_{i2}) = \frac{\exp(\phi_0 + \phi_1 Z_{i1} + \phi_2 Z_{i2})}{1 + \exp(\phi_0 + \phi_1 Z_{i1} + \phi_2 Z_{i2})}$$

and $(\phi_1, \phi_2) = (1, -0.5)$. We vary size of the IVS via changing the value of ϕ_0. In specific, we set the total number of observations in this simulation study $N = 10000$, and vary $\phi_0 \in \{-4.5, -3.5, -2.8, -2\}$, which corresponds to sizes of IVS at around 100, 250, 500, and 1000. Similar to previous simulation studies, we generate the two true exposures from a bivariate normal measurement error model, $\widetilde{\boldsymbol{X}}_i \sim \mathcal{N}(\boldsymbol{\Gamma}^* \widetilde{\boldsymbol{Z}}_i, \boldsymbol{\Sigma_x})$, where

$\widetilde{\boldsymbol{Z}}_i = (1, Z_{i1}, Z_{i2}, X_{i3})'$ and $\boldsymbol{\Sigma_x} = \boldsymbol{\Sigma_z}$. The correctly measured exposure, X_{i3}, is generated from the standard normal distribution. The parameter matrix for the reclassification model is set as

$$\boldsymbol{\Gamma}^* = \begin{pmatrix} -0.2 & 0.6 & 0.2 & -0.7 \\ 0.1 & 0.2 & 0.4 & -0.5 \end{pmatrix},$$

which leads to moderate degree of measurement errors. The exposure-outcome relationship model is assumed logistic with

$$\mu_i = \mathbb{E}(Y_i|\boldsymbol{X}_i) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3})},$$

and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3) = (1, 0.5, 0.5)$. The relative bias, MCSE, ASE, RMSE and coverage probability are reported in Table 11 and 12, where the proposed sampler performs similarly on the IVS sampled with bias.

Table 3.21: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of the uncorrected and proposed Bayesian estimators for estimating the three pPIFs, with different validation study size. The simulation results are based on 1000 data replications, under $\beta_0 = -2$ (common disease), multivariate normal true reclassification model, biased validation study sample, and the moderate measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

N_v^*	Estimand	Method	Bias	MCSE	ASE	RMSE	Coverage
100	PIF_{X_1}	Uncorrected	-28.61	0.007	0.007	0.071	0.00
		Bayesian	1.41	0.042	0.036	0.045	89.10
	PIF_{X_2}	Uncorrected	-19.27	0.008	0.008	0.032	12.00
		Bayesian	0.92	0.057	0.051	0.055	90.30
	PIF_{X_1,X_2}	Uncorrected	-24.42	0.008	0.008	0.089	0.00
		Bayesian	1.66	0.037	0.034	0.045	91.00
250	PIF_{X_1}	Uncorrected	-28.23	0.007	0.007	0.071	0.00
		Bayesian	0.76	0.025	0.024	0.032	92.60
	PIF_{X_2}	Uncorrected	-19.54	0.008	0.008	0.032	10.50
		Bayesian	-0.42	0.033	0.033	0.032	93.30
	PIF_{X_1,X_2}	Uncorrected	-24.26	0.008	0.008	0.089	0.00
		Bayesian	0.53	0.023	0.022	0.032	**93.80**
500	PIF_{X_1}	Uncorrected	-28.66	0.007	0.007	0.071	0.00
		Bayesian	0.45	0.017	0.018	0.017	**93.90**
	PIF_{X_2}	Uncorrected	-19.16	0.008	0.008	0.032	10.00
		Bayesian	0.49	0.023	0.024	0.032	**95.40**
	PIF_{X_1,X_2}	Uncorrected	-24.41	0.008	0.008	0.089	0.00
		Bayesian	0.52	0.016	0.017	0.014	**94.50**
1000	PIF_{X_1}	Uncorrected	-28.33	0.007	0.007	0.071	0.00
		Bayesian	-0.07	0.013	0.013	0.020	**93.60**
	PIF_{X_2}	Uncorrected	-19.37	0.008	0.008	0.032	9.50
		Bayesian	0.24	0.017	0.018	0.018	**94.70**
	PIF_{X_1,X_2}	Uncorrected	-24.27	0.008	0.008	0.087	0.00
		Bayesian	0.07	0.012	0.012	0.017	**95.50**

*Average N_v based on $N = 10000$ and values of (ϕ_0, ϕ_1, ϕ_2).

Table 3.22: Simulation results for the relative bias (% Bias), Monte Carlo standard error (MCSE), average (posterior) standard error (ASE), root mean squared error (RMSE) and empirical coverage percentage of 95% CIs (Coverage %) of the uncorrected and proposed Bayesian estimators for estimating the three pPIFs, with different validation study size. The simulation results are based on 1000 data replications, under $\beta_0 = -4$ (rare disease), multivariate normal true reclassification model, biased validation study sample, and the moderate measurement error scenario. Coverage percentage between 93.6% and 96.4% in bold font are within the margin of error for 1000 replications.

N_v^*	Estimand	Method	Bias	MCSE	ASE	RMSE	Coverage
100	PIF_{X_1}	Uncorrected	-26.90	0.015	0.014	0.089	0.00
		Bayesian	3.87	0.074	0.058	0.077	86.20
	PIF_{X_2}	Uncorrected	-18.60	0.016	0.016	0.033	43.50
		Bayesian	-3.90	0.111	0.093	0.114	87.40
	PIF_{X_1,X_2}	Uncorrected	-22.26	0.016	0.015	0.105	0.00
		Bayesian	3.15	0.058	0.049	0.063	88.60
250	PIF_{X_1}	Uncorrected	-27.27	0.016	0.014	0.089	0.00
		Bayesian	0.83	0.041	0.039	0.044	92.70
	PIF_{X_2}	Uncorrected	-18.98	0.016	0.016	0.032	39.00
		Bayesian	-1.75	0.063	0.060	0.062	93.30
	PIF_{X_1,X_2}	Uncorrected	-22.60	0.016	0.015	0.107	0.00
		Bayesian	0.71	0.033	0.033	0.031	93.50
500	PIF_{X_1}	Uncorrected	-26.99	0.014	0.014	0.089	0.00
		Bayesian	0.17	0.031	0.030	0.031	**94.00**
	PIF_{X_2}	Uncorrected	-18.82	0.016	0.016	0.032	41.00
		Bayesian	1.68	0.047	0.045	0.046	93.30
	PIF_{X_1,X_2}	Uncorrected	-22.33	0.016	0.015	0.105	0.00
		Bayesian	0.93	0.025	0.025	0.032	93.40
1000	PIF_{X_1}	Uncorrected	-27.06	0.017	0.014	0.091	0.00
		Bayesian	0.33	0.023	0.023	0.031	**93.60**
	PIF_{X_2}	Uncorrected	-18.82	0.018	0.016	0.046	46.00
		Bayesian	-0.47	0.0350	0.0339	0.036	93.30
	PIF_{X_1,X_2}	Uncorrected	-22.41	0.017	0.015	0.109	0.00
		Bayesian	0.27	0.020	0.020	0.012	93.40

*Average N_v based on $N = 10000$ and values of (ϕ_0, ϕ_1, ϕ_2).

Chapter 4

When customer dynamics is more than relationship: A coupled hidden Markov model framework[1]

Abstract

Despite the growing interest in using hidden Markov model (HMM) to study customer dynamics and implement customer relationship management (CRM), little is known about whether a single Markov process can adequately capture dynamics of customers. In this research we propose a coupled non-homogeneous hidden Markov model (CNHMM) framework that simultaneously considers two distinct yet (potentially) correlated Markov processes, respectively representing the latent relational and monetary value of customers. Leveraging data from a major telecommunication carrier in China, our findings indicate that the proposed method is able to uncover the multi-dimensional latent states of customers (dynamic customer values) and possible effects of covariates of interests (including marketing mixes) on the evolutions of the latent states. Consumers' choice of products (and services) are jointly influenced by their relational and monetary value over time, and the evolution of customers' relational states is significantly dependent on their monetary states

[1] Co-authored with Yiwei Li and Xiangnan Feng

(but not vice versa), suggesting customer heterogeneity in monetary value is a potential antecedent of customer-firm relationship. Furthermore, we show how scenario analysis using the proposed model can help firms formulate effective multidimensional dynamic segmentation strategies for customer relationship management.

4.1 Introduction

Currently, firms widely use the recency, frequency, and monetary value (RFM) framework to measure customer value (Petersen and Kumar, 2015; Lewis, 2006; Zhang et al., 2015), conduct customer segmentation (Fader et al., 2005; Feinberg et al., 2016; Haenlein et al., 2006), and allocate marketing resources to customers (Avery et al., 2012; Venkatesan and Farris, 2012; Wübben and v. Wangenheim, 2008). This customer-centric approach, coupled with the increasing availability of customer-generated big data, has led to an interest in both the notion and the calculation of RFM-based metrics and scores (Ansari et al., 2008; Ertekin et al., 2019; Venkatesan et al., 2007). Conceptually, while it seems straightforward that customer value can be calculated independently on each of the three dimensions (i.e. the recency, frequency, and monetary value), firms in practice assign different weights to them according to their perceived importance. For an example, Reinartz and Kumar (2000) suggest firms assign maximum importance to recency then to monetary value and the lowest importance to frequency.

In understanding the differential importance of customer value in various dimensions, a common practice is to associate each of the dimension (e.g. RFM metric) with firms' sales performance, typically in regression analysis (Ansari et al., 2008; Ertekin et al., 2019). This approach however implicitly assumes orthogonal relationships between the dimensions, by including them in the regression as independent drivers of product sales. In particular, we argue customer value measured in different dimensions are potentially interdependent. Without loss of generality, we consider a two-dimensional case where a

customer is measured in relational (e.g. recency and frequency in service-encounters or product purchases) and monetary value (e.g. average spending per transaction). For instance, loyal customers (i.e. high in relational value) may tend to spend more with the firm (i.e. also high in monetary value), and the interdependence argument still holds true if we frame the relationship reversely: customers who spends more can (i.e. high in monetary value) easily become loyal customers. Therefore, we aim to develop a methodological framework that helps firms identify potential interdependence between distinctive dimensions of customer value, and also obtain rigorous empirical evidence revealing such interdependent relationships.

For customer-centric firms, another challenge in understanding customer value is that customer preferences and behaviors are fundamentally dynamic (Ansari et al., 2008; Rhee and McIntyre, 2008; Zhang and Chang, 2021). For instance, customers who received a direct mail piece last week may be less receptive to a direct mail piece this week (Neslin et al., 2013), and some customers may stop interacting with a firm by starting to ignore the communications coming from it (Ascarza et al., 2018). Leveraging real-time data instead of static list-based audiences of yesteryear, firms start to manage the timing of marketing efforts in order to obtain optimal results. Consequently, to contact the right customer (whom) at the right time (when) has now become the central theme of customer relationship management (CRM) and dynamic segmentation (Ma et al., 2015a; Netzer et al., 2008; Zhang et al., 2014).

Insofar as the complexity lies in both the interactive nature of the dimensions of customer value and the embedded dynamic process, we propose a flexible approach that builds on the coupled nonhomogeneous hidden Markov model (CNHMM) framework (Sherlock et al., 2013; Touloupou et al., 2020a,b). Hidden Markov models (HMMs) have been widely used in marketing to study the dynamics in customer behavior, delineating how customers move back and forth between different unobserved (or latent) states. These latent states, which govern the observed customer behavior, include, for example the relationship between

the customer and the firm (Ansari et al., 2012; Holtrop et al., 2017; Zhang et al., 2017), purchasing propensity (Liechty et al., 2003; Schwartz et al., 2014), and price sensitivity (Zhang et al., 2014).

Unlike the conventional HMM which includes a single Markov process (e.g. customer-firm relationship states), the proposed CNHMM employs multiple underlying Markov processes that evolve dynamically (e.g. two simultaneously evolving latent processes representing customers' relational and monetary states). Importantly, the CNHMM relaxes the complete independence assumption among underlying Markov processes by explicitly modeling the interdependence among latent states across underlying processes (e.g. whether customer relational states drive monetary states or vice versa), and thus help firms to evaluate the differential importance of each underlying process. In addition, we allow customer latent states to be influenced by time-varying covariates (such as service encounters and customer spending), leading to time-nonhomogeneous transition matrices, while accounting for unobserved customer heterogeneity through individual-specific parameters both in the latent state evolution and observed customer behavior (given state membership) stages.

We apply the proposed CNHMM model to a large-scale longitudinal mobile service data from a major Chinese telecommunication carrier, comprising 217,726 subscribers from a major city in southwest China for two years from October 2013 to September 2015. We identify two distinctive latent processes, which correspond to the evolution of customer relational and monetary states respectively. The results suggest an interesting interactive pattern of the two underlying processes: the evolution of customers' relational states is dependent upon their monetary states but not vice versa. Specifically, customers in higher monetary states are likely to transition to higher relational states, implying that the level of customer engagement may be partially determined by how much money customers spend with the firm. While the extant literature on HMM has predominantly focused on single-Markov-process customer-firm relationships (Ascarza et al., 2018; Ma et al., 2015a; Zhang et al., 2017), our results indicate dynamic monetary value serves as a potential source

of customer heterogeneity in CRM. By considering the important heterogeneity, our model also performs better than competing conventional frameworks at identifying the latent states and predicting future customer behavior (e.g. service plan renewal, and phone purchase).

The proposed framework generates useful guides for firms to actively manage customer base and effectively allocate marketing resources. Through identifying different factors that can influence either the evolution of customer latent states or the observed behaviors, we demonstrate that firms can achieve better performance via meaningful service interventions. In particular, insights from (the evolution of) customer latent states can help firms formulate effective promotion strategies. This is exemplified through our scenario analyses in which we suggest, by focusing on customers in some specific segments, firms can increase the expected revenue from service plan renewal and phone purchase of customers. In addition, firms can dynamically (e.g. for every month) recover the latent monetary-relational state (segment) of their existing customers and effectively allocate marketing resources according to the perceived importance of each segment.

The remainder of this paper is organized as follows. Section 4.2 continues by discussing the literature pertaining to customer value, dynamic segmentation, and HMM-based models. Section 4.3 develops a flexible CNHMM model for firms to dynamically segment customers based on their monetary and relational value. In Section 4.4 we apply our model to a large-scale longitudinal mobile service data, and show how two Markov processes simultaneously and interactively reflect the evolution of the customers' monetary and relational states, which govern customers' phone purchase and service plan renewal behavior. Employing scenario analyses, we also demonstrate how the telecom carrier could increase sales through actively targeting the right segments. Finally, Section 4.5 concludes the paper with a discussion.

4.2 Literature Review

In this section we discuss two streams of relevant literature. From a substantive point of view, our work relates to customer value, more specifically, to how it serves as a basis for CRM and marketing resource allocation. From a methodological point of view, our work relates to a growing number of HMM applications in marketing.

4.2.1 Customer Value-based CRM

The contention that loyal customers are always more profitable is a gross oversimplication.

— Dowling and Uncles

Using a customer-value-based approach to conduct CRM yields several benefits (Kumar and Reinartz, 2012), such as decreased cost (Lewis, 2004), reactivation of dormant customers (Neslin et al., 2013), acquisition and retention of profitable customers (Lewis, 2006), and increased profits from marketing investments (Venkatesan et al., 2007). The importance of customer value has led to keen and growing interests on customer lifetime value (CLV) in a wide sense, for firms that wish to continuously and dynamically align their resources with drivers of customer value (Kumar et al., 2008; Venkatesan et al., 2007; Zhang et al., 2015). Customer value, or CLV, is a dynamic concept, not only because its magnitude is likely to change over time, but also because the determinants of customer value may alter significantly (Haenlein et al., 2006; Parasuraman, 1997).

With a dynamic perspective, firms are now finding their ways to optimize the current as well as the future value of customers (Fader et al., 2005; Kumar and Reinartz, 2012). Naturally customer-firm relationship has become the central focus and a key indicator of future profitability (Ascarza et al., 2018; Reinartz and Kumar, 2003). A substantial body of literature has evaluated the effectiveness of relationship-oriented interventions, such as the

influence of loyalty programs (Lewis, 2004; Liu, 2007; Wang et al., 2016). While great efforts have been directed to create and maintain strong relational bond between firms and customers, doubts have been cast on the basic assumption that a loyal customer (high in relational value) is always more profitable. For instance, Rust and Verhoef (2005) suggest that action-oriented CRM interventions such as sales promotions, and direct mailing with coupons are less effective among loyal customers. As loyal customers might have reached their potential relational value in view of the number of products purchased, they might be less likely to purchase additional ones, despite a received direct mailing with a call for action (Dwyer et al., 1987; Grant and Schlesinger, 1995).

Our paper provides a new perspective for Dowling and Uncles (1997) which claims that it is a gross oversimplication to equate loyal customers with higher profits. Specifically, we consider the monetary value of customers as potential antecedents of their relational value, and therefore a loyal customer is not necessarily more profitable, unless he reaches a "minimum spend" in history. While several studies have included monetary value in predicting CLV, they simply discount the future cashflows associated with a customer to yield a net present value (NPV) (Reinartz and Kumar, 2000, 2003), or summarize the relational and monetary value as independent predictors of CLV, such as through RFM framework (Haenlein et al., 2006; Schmittlein and Peterson, 1994). Our approach used to assess the customer value provides several benefits relative to standard CLV calculations. First, by explicitly modeling the interdependence between the relational and monetary value, we allow customers' relational value to be endogenously and dynamically affected by their monetary value (and vice versa); Second, the proposed framework allows firms to validate the widely-held view that relational value (e.g. recency) is usually a more powerful discriminator than monetary value (Fader et al., 2005; Hughes, 2000), and therefore enhances the flexibility for firms directly to link CLV to marketing decisions, such as to identify profitable yet short-lived group and then to stop chasing these customers (Reinartz and Kumar, 2003), or to simply abandon unprofitable but strongly bonded customers (Haenlein

et al., 2006). Third, customer value assessed under our proposed alternative approach can serve as a basis for segment classification, enabling firms to allocate scarce marketing resources and make individual customer purchase-level forecast in a real time fashion.

The results based on our mobile service data suggest customers' relational value can be influenced by their monetary value, i.e., a customer in higher spending level tends to remain more loyal to the firm (but not the vice versa). We emphasize that the purpose of this empirical study is not to straightly conclude monetary value as the determinant of customer relational value, since the relationship between these two important dimensions are likely to change across industries and perhaps over time. That being said, we would join the sentiment with the work of Dowling and Uncles that loyalty does not always go first. Managers would be wise if they can flexibly conduct CRM based on both the relational and monetary value of customers, assessing potential interdependences in between simultaneously.

4.2.2 HMM-based CRM

Customers' relational and monetary value are not only individual-specific, but also time-varying. For example, a customer can move from a loyal to a disloyal segment over time. To capture such dynamic evolution of customer value, HMMs are a representative setup wherein customers migrate among a set of latent "states" (akin to latent segments in dynamic customer segmentation) over time.

HMMs have made significant inroads into marketing over the past several decades (Kappe et al., 2018; Montgomery et al., 2004; Zhang and Chang, 2021). Increasing attempts are being made to use HMMs to model the dynamic change in customer-firm relationship. Netzer et al. (2008) used an HMM in the context of university alumni donation and classified alumni into dynamic relationship states based on their changing propensities for donation. Montoya et al. (2010) employed an HMM to explore how pharmaceutical

marketing managers optimize targeting activities to individual customers (physicians), and found that detailing is more effective for acquisition whereas sampling is more effective for retention. Ascarza et al. (2018) applied HMMs to two contexts, a daily deal website and a performing arts organization, and separated two types of customer churns, the observed and unobserved customer attrition (i.e. "overt" churn and "silent" churn).

Leveraging HMMs in the area of CRM has unique advantages (Du and Kamakura, 2006; Moon et al., 2007; Zhang and Chang, 2021). First, in today's information-rich environments, customer-firms encounters take places across multiple channels and in various forms, and customer relationship encompassing multiple encounters measures various facets of a richer relational construct; Second, HMMs classify customers into a set of "latent states" (e.g. low and high loyalty states) based on their observed (buying) behaviors, and such empirically determined states naturally shed light on the segmentation strategies for CRM; Third, HMMs estimate the movements of customers in and out of the segments and identify drivers of the segment transitions, and therefore allow firms to perform dynamic targeting, through tailoring marketing actions to nudge customers towards desirable segments; Forth, marketing mix tailored for each segment can be created to directly influence customer (state-dependent) behaviors, representing another useful feature of dynamic targeting.

Our CNHMM approach pushes forward the marketing literature of the HMM and its applications in CRM from several aspects. First, the CNHMM provides a framework to include multiple potentially correlated underlying Markov processes, by relaxing the assumptions that only one or independent underlying process(es) should be there (e.g. for CRM). We also find superior explanation power and predictive validity of the proposed CNHMM relative to the traditional HMM and the factorial HMM (FHMM, e.g. an HMM with two independent hidden Markov chains). Second, by adding monetary value as another distinct yet potentially more fundamental segmentation criterion, the CNHMM allows firms to implement multidimensional segmentation strategies. For example, managers will be able to identify high-loyalty but low-monetary, and low-loyalty yet high-monetary

customers; Third, based on the results generated by the CNHMM, firms can have richer inputs (considering both latent customer relational and monetary states) to dynamically tailor targeting actions for customers in a just-in-time fashion, as well as allocate marketing resources to maximize the long-run profitability.

4.3 Model Development

We extend the existing methods for modeling customer relational value and monetary value by proposing a joint modeling framework based on an CNHMM that can incorporate both dimensions as two distinct yet interacting latent processes.

There is a collection of Marketing literature on using Markov models to characterize customer dynamics (Zhang and Chang, 2021), within which, a sizable section of research focuses on using HMMs to model dynamic changes in customer-firm relationships (Netzer et al., 2008; Montoya et al., 2010; Ascarza et al., 2018). These studies predominantly consider a single aspect, i.e., relational value, in the RFM framework for CRM. We note that some studies modeled multidimensional behavior of customers (e.g., Ascarza et al. (2018) modeled customers' behaviors of opening promotion emails, and clicking on deals or the unsubscribe button included in those emails) without considering the existence or effects of multiple latent customer values. Therefore, in this study, we propose to analyze dynamics of customer behaviors by simultaneously modeling two latent customer values using a CNHMM.

Different from the traditional HMM and the factorial HMM, the CNHMM allows for dependence between the two latent Markov chains representing the two evolving customer values, monetary and relational, respectively. The approach enables firms to implement finer multidimensional segmentation strategies. For instance, managers will be able to segment customers into various groups based on their combinations of loyalty and spent levels, and devise more specific and effective promotion strategies targeting each different

group.

4.3.1 Preliminaries

We model customers' multidimensional observed behaviors regarding telecom services. At each period, we observe two customer behaviors for every individual: (i). whether or not the customer has changed her mobile phone tier, and (ii). whether or not the customer has changed her telecom service plan. For the customer behavior of changing mobile phones, the customer first makes the decision on whether or not to change the tier of her mobile phone; and if she decides to change phone tiers, she then will decide on choosing an upgrade or a downgrade. The tiers of mobile phones are largely characterized by phone prices. For the other customer behavior, we can observe in the data whether or not the customer decides to change her telecom service plan at each period. Here, we solely focus on the behavior of changing telecom service plans because such it can largely reflect the level of engagement of a customer with the firm.

More formally, we observe a three-dimensional binary random vector $\boldsymbol{Y}_{i,t_i} = [Y^c_{i,t_i}, Y^p_{i,t_i}, Y^{pl}_{i,t_i}]$ with the realization $\boldsymbol{y}_{i,t_i} = [y^c_{i,t_i}, y^p_{i,t_i}, y^{pl}_{i,t_i}]$, where $y^c_{i,t_i} = 1$ if customer i changes the tier of her mobile phone at time t_i (0 otherwise), $y^p_{i,t_i} = 1$ if customer i upgrades her mobile phone at time t_i conditional on a phone tier change happens in the same period (0 for downgrade), and $y^{pl}_{i,t_i} = 1$ if customer i changes her telecom service plan at time t_i (0 otherwise).

We assume the existence of a pair of latent variables that respectively reflect the customer's monetary and relational values to the firm. Each latent variable takes values from a set of latent states, and we model the likelihood, $\mathbb{P}(\boldsymbol{Y}_{i,t_i} = \boldsymbol{y}_{i,t_i})$, as a function of the latent state pairing occupied by customer i at time t_i. For instance, we expect that a customer who is highly interested in buying a newly released mobile phone that is more expensive and switching to a new service plan will be captured by a latent state pairing

that exhibits high probabilities of upgrading mobile phones and changing service plans. In contrast, a customer who might be tightening her financial budget on wearable tech gadgets and showing little interest in viewing different service plan options will be captured by a latent state pairing with a high probability of downgrading mobile phones and perhaps a low probability of switching service plans.

We assume that customers transition among the latent state pairings following a first-order coupled Markov process, where for a given customer, the two latent states in period t are independent conditional on latent states in period $t-1$. This conditional independence structure in the CNHMM bridges the gap between the overly simplistic assumption of complete independence between the two latent states by the factorial HMM, and a potentially over-parameterized HMM that assumes general dependence between the two latent states. Furthermore, to better understand the evolutions and impacts of the two latent customer values, we allow managerially relevant covariates such as average revenue per user (ARPU) of different services and customers' frequencies of service calls with the firm to affect customers' transitions among latent state pairings as well as their behaviors given membership of a particular latent state pairing.

4.3.2 Model Specification

The model consists of two main components, both describing dynamics at the individual level: (i). the latent state pairing evolution, and (ii). the customer's state-dependent behaviors (e.g., the probability of changing phone tiers, upgrading phones, and switching service plans). We account for individual heterogeneity across customers by including individual-specific parameters in both the states evolution and customer behaviors (given state membership) processes.

Latent States Evolution

We assume K^m latent monetary value states, which differ with respect to the customer's probability of changing and upgrading/downgrading her mobile phone, and K^r latent relational value states, which differ with respect to the probability of the customer switching to other service plans. To capture the evolution of customers' behaviors, we allow customers to transition among latent state pairings over time. Let Z^m_{i,t_i} and Z^r_{i,t_i} denote the latent state pairing (monetary and relational, respectively) occupied by customer i at time t_i. The evolution of Z^m_{i,t_i} and Z^r_{i,t_i} each follows a first-order hidden Markov process.

Conceptually, however, it will be overly simplistic to assume evolution processes of a customer's monetary and relational values to the firm are completely independent. For instance, a customer is more likely to select products from a certain brand/firm even if they are more expensive than alternatives with equal value of functionality when she occupies a state of higher monetary values, and such phenomenon of brand/firm loyalty rarely happens when the customer falls into a state of low monetary value. Considering such possible scenarios, we assume that there could exist potential dependence between the two latent variables when they are evolving over time. This assumption of dependent latent processes leads to us adopting a CNHMM to describe the underlying transition dynamics. In specific, with the CNHMM, we assume the following transition probability for the customer's latent state pairings:

$$\mathbb{P}(Z^m_{i,t_i+1}, Z^r_{i,t_i+1} | Z^m_{i,t_i}, Z^r_{i,t_i}) = \mathbb{P}(Z^m_{i,t_i+1} | Z^m_{i,t_i}, Z^r_{i,t_i}) \times \mathbb{P}(Z^r_{i,t_i+1} | Z^m_{i,t_i}, Z^r_{i,t_i}). \tag{4.1}$$

The two latent variables are interactively evolving in a way that the value of one latent variable at time $t_i + 1$ is dependent upon values of both latent variables at time t_i. This is equivalent to assuming the independence between Z^m_{i,t_i+1} and Z^r_{i,t_i+1} conditional on Z^m_{i,t_i} and Z^r_{i,t_i}. Such conditional independence structure in the CNHMM serves as an intermediary

between the overly simplistic assumption of complete independence between the two latent states by the FHMM, and a potentially over-parameterized HMM that assumes general dependence between the two latent states.

We model the customer's probabilities of moving from one latent state to another using multinomial logit models:

$$\begin{aligned}&\mathbb{P}(Z^m_{i,t_i} = k^m_2 | Z^m_{i,t_i-1} = k^m_1, Z^r_{i,t_i-1} = k^r) \\ &\qquad = \frac{\exp\left(\zeta^m_{i,k^m_1,k^m_2} + \eta^m_{k^m_1,k^m_2} + \boldsymbol{W}^m_{i,t_i-1}(k^r)'\boldsymbol{\rho}^m_{k^m_2}\right)}{\sum_{\kappa^m=1}^{K^m}\exp\left(\zeta^m_{i,k^m_1,\kappa^m} + \eta^m_{k^m_1,\kappa^m} + \boldsymbol{W}^m_{i,t_i-1}(k^r)'\boldsymbol{\rho}^m_{\kappa^m}\right)}\end{aligned} \tag{4.2}$$

for $k^m_1, k^m_2 \in \{1, \ldots, K^m\}$;

$$\begin{aligned}&\mathbb{P}(Z^r_{i,t_i} = k^r_2 | Z^r_{i,t_i-1} = k^r_1, Z^m_{i,t_i-1} = k^m) \\ &\qquad = \frac{\exp\left(\zeta^r_{i,k^r_1,k^r_2} + \eta^r_{k^r_1,k^r_2} + \boldsymbol{W}^r_{i,t_i-1}(k^m)'\boldsymbol{\rho}^r_{k^r_2}\right)}{\sum_{\kappa^r=1}^{K^r}\exp\left(\zeta^r_{i,k^r_1,\kappa^r} + \eta^r_{k^r_1,\kappa^r} + \boldsymbol{W}^r_{i,t_i-1}(k^m)'\boldsymbol{\rho}^r_{\kappa^r}\right)},\end{aligned} \tag{4.3}$$

for $k^r_1, k^r_2 \in \{1, \ldots, K^r\}$. Here $\boldsymbol{W}^m_{i,t_i-1}(k^r)$ and $\boldsymbol{W}^r_{i,t_i-1}(k^m)$ are vectors of transition covariates which also characterized the coupling mechanism between two Markov chains, and we have

$$\boldsymbol{W}^m_{i,t_i-1}(k^r) = \left[\boldsymbol{W}^m_{i,t_i-1}, \mathbb{I}\{k^r = 2\}, \ldots, \mathbb{I}\{k^r = K^r\}\right],$$

and

$$\boldsymbol{W}^r_{i,t_i-1}(k^m) = \left[\boldsymbol{W}^r_{i,t_i-1}, \mathbb{I}\{k^m = 2\}, \ldots, \mathbb{I}\{k^m = K^m\}\right].$$

Parameter vectors $\boldsymbol{\rho}^m_{k^m}$ and $\boldsymbol{\rho}^r_{k^r}$ capture the effects of those covariates. Parameters $\eta^m_{k^m_1,k^m_2}$ and $\eta^r_{k^r_1,k^r_2}$ determine propensities to transition from state k^m_1 to k^m_2 and k^r_1 to k^r_2, respectively.

Customers are assumed to differ in their propensities to transition among latent states. These differences in transition probabilities reflect the hypothesis that customers may exhibit different lifetimes or shorter (versus longer) runs of frequent activity. This heterogeneity across customers is captured by individual-level propensities ζ^m_{i,k^m_1,k^m_2} and ζ^r_{i,k^r_1,k^r_2}. For identification purposes, we set $\zeta^m_{i,k^m_1,1}$, $\eta^m_{k^m_1,1}$, $\boldsymbol{\rho}^m_1$, $\zeta^r_{i,k^r_1,1}$, $\eta^r_{k^r_1,1}$, and $\boldsymbol{\rho}^r_1$ as zero.

We assume the following initial condition to determine the latent state memberships for customers in period 1. In specific, we assume that the probabilities that a customer belongs to latent state pairing k^m and k^r at time $t_i = 1$ are determined by parameters $\boldsymbol{\pi}^m = (\pi^m_2, \ldots, \pi^m_{K^m})$ and $\boldsymbol{\pi}^r = (\pi^r_2, \ldots, \pi^r_{K^r})$, where

$$\mathbb{P}(Z^m_{i,1} = k^m) = \begin{cases} \dfrac{1}{1 + \sum_{\kappa^m=2}^{K^m} \exp(\pi^m_{\kappa^m})}, & \text{for } k^m = 1 \\ \dfrac{\exp(\pi^m_{k^m})}{1 + \sum_{\kappa^m=2}^{K^m} \exp(\pi^m_{\kappa^m})}, & \text{for } k^m = 2, \ldots, K^m \end{cases}, \tag{4.4}$$

and

$$\mathbb{P}(Z^r_{i,1} = k^r) = \begin{cases} \dfrac{1}{1 + \sum_{\kappa^r=2}^{K^r} \exp(\pi^r_{\kappa^r})}, & \text{for } k^r = 1 \\ \dfrac{\exp(\pi^r_{k^r})}{1 + \sum_{\kappa^r=2}^{K^r} \exp(\pi^r_{\kappa^r})}, & \text{for } k^r = 2, \ldots, K^r \end{cases}. \tag{4.5}$$

Observed Behaviors

In each period, we observe whether the customer changes her model phone tier or service plan, represented by the random vector $\boldsymbol{Y}_{i,t_i} = [Y^c_{i,t_i}, Y^p_{i,t_i}, Y^{pl}_{i,t_i}]$. In the model, conditional on the latent state membership, we allow the customer's behavior to be affected by covariates (e.g., time to the nearest iPhone release date, number of available phone choices) that might influence behavior without altering the latent states she occupies at the time. For example, customers might be more likely to upgrade their mobile phones when a new influential hi-tech smartphone is released to the market.

Since the customer needs to make the decision on whether or not to upgrade/downgrade her mobile phone conditional on her decision to change her mobile phone tiers, we model the customer's phone changing behavior via a two-stage approach, where the customer first makes a decision on whether or not to change the tier of her mobile phone, and conditional on the scenario where the customer decides to change phone tiers, she makes a second decision on whether to upgrade or downgrade. Feng et al. (2020) suggest that (the price of) mobile phone is an indicator of the social capital of the user, and thus we consider the phone changing behavior is primarily associated with customers' latent monetary value state. This can also be reflected by our model-free evidence in section 4.4. The probability of customer i deciding to change phone tier at period t_i given the latent monetary state k^m is

$$\mathbb{P}(Y^c_{i,t_i} = 1 | \boldsymbol{X}^c_{i,t_i}, Z^m_{i,t_i} = k^m) = \frac{\exp(\xi^c_{i,k^m} + \mu^c_{k^m} + \boldsymbol{X}^c_{i,t_i}{}' \boldsymbol{\beta}^c_{k^m})}{1 + \exp(\xi^c_{i,k^m} + \mu^c_{k^m} + \boldsymbol{X}^c_{i,t_i}{}' \boldsymbol{\beta}^c_{k^m})}. \tag{4.6}$$

This probability is modeled as a function of underlying propensities of changing phone tiers that varies across latent monetary value states, $\mu^c_{k^m}$, across customers, ξ^c_{i,k^m}, and customer-level time-varying covariates that might affect the phone tier changing behavior conditional on a given state, $\boldsymbol{X}^c_{i,t_i}$. Since both the propensity of changing phone tiers, $\mu^c_{k^m}$, and effects of covariates, $\boldsymbol{\beta}^c_{k^m}$, are state-specific, customers in different monetary value states can have different underlying propensities of changing phone tiers and different sensitivity to stimuli from the time-varying covariates.

In a similar fashion, we can model the second stage of customers' phone changing behavior, where the probability of customer i in latent monetary value state k^m upgrading her mobile phone at time t_i, conditional on she has made the decision of changing phone

tiers, is

$$\mathbb{P}(Y^p_{i,t_i} = 1|\boldsymbol{X}^p_{i,t_i}, Z^m_{i,t_i} = k^m) = \frac{\exp(\xi^p_{i,k^m} + \mu^p_{k^m} + \boldsymbol{X}^p_{i,t_i}{}'\boldsymbol{\beta}^p_{k^m})}{1 + \exp(\xi^p_{i,k^m} + \mu^p_{k^m} + \boldsymbol{X}^p_{i,t_i}{}'\boldsymbol{\beta}^p_{k^m})}. \tag{4.7}$$

$\mu^p_{k^m}$ is the propensity of upgrading mobile phones across latent states. ξ^p_{i,k^m} represents the customer-level heterogeneity in phone upgrading propensity, and $\boldsymbol{X}^p_{i,t_i}$ is the customer-level time-varying covariates that might affect phone upgrading.

Finally, we model the remaining observed behavior of customers changing telecom service plans. In the literature, the behavior of customers switching service plans often reflects their latent relational value states to the firm. In specific, customers with higher relational value exhibit higher levels of engagement with the firm (Lee et al., 2018), and therefore show higher levels of interest in learning and switching to new service plans provided by the same firm. We thus model the probability that customer i will switch service plans in period t_i given membership of relational value state k^r as

$$\mathbb{P}(Y^{pl}_{i,t_i} = 1|\boldsymbol{X}^{pl}_{i,t_i}, Z^r_{i,t_i} = k^r) = \frac{\exp(\xi^{pl}_{i,k^r} + \mu^{pl}_{k^r} + \boldsymbol{X}^{pl}_{i,t_i}{}'\boldsymbol{\beta}^{pl}_{k^r})}{1 + \exp(\xi^{pl}_{i,k^r} + \mu^{pl}_{k^r} + \boldsymbol{X}^{pl}_{i,t_i}{}'\boldsymbol{\beta}^{pl}_{k^r})}, \tag{4.8}$$

where $\mu^{pl}_{k^r}$ and ξ^{pl}_{i,k^r} are population- and customer-level propensities of customer changing service plans, $\boldsymbol{X}^{pl}_{i,t_i}$ is the customer-level time-varying covariates that might affect the service plan switching behavior when in a given state, and $\boldsymbol{\beta}^{pl}_{k^r}$ is the corresponding effects vector.

The specification of our model assumes that the two observed behaviors are interconnected via the hidden states, but conditionally independent given the latent state membership. Time-specific random effects could potentially be added to the probabilities of observed behaviors given the latent state membership provided there is more interest in measuring further correlations among these behaviors. We perform the model estimation using the method of stochastic variational Bayes (SVB) due to the large size of our data.

Details of the estimation approach is available in section 4.6.1 of the Appendix.

4.4 Results

We apply our CNHMM framework to a large-scale longitudinal mobile service data obtained from a major Chinese telecommunication carrier. The telecom carrier operates by providing customers with different options of mobile phones and service plans. Customers can choose the type of products in terms of service plans and/or mobile phones based on their needs. The more frequently appeared purchasing paradigms are: (i). a new customer joining the carrier's network by purchasing a service plan only, (ii). a new customer purchasing a new mobile phone and a service plan from the carrier, (iii). an existing customer switching to another service plan without purchasing new mobile phones from the carrier, and (iv). an existing customer purchasing a new mobile phone and switching to another service plan both provided by the carrier. A customer becomes a subscriber to the carrier once she purchases a service plan. Subscribers can switch to other available service plans or terminate the current service plan (unsubscribe from the carrier) at any time they desire without additional costs.

Service plans offered by the carrier primarily differ in prices and volumes of three main telecom services, namely, phone calls, SMSs, and mobile Internet data. Customers who have higher technology affinity may tend to select service plans providing lower prices on mobile Internet data. The telecom carrier also offers a large variety of mobile phones, ranging from phones that do not support 3G network to those equipped with the most cutting-edge technology. Service plans could also be sold together with a newly purchased mobile phone, but such bundling is not compulsory. From the perspective of the carrier, one important operating aspect is to retain subscribers through identifying their characteristics and making timely targeting offers that improves profitability in the long-run.

4.4.1 Data Description and Patterns

We collected data from a branch of the carrier located in a city of southwest China. An overwhelming majority of the customers included in the data live in the same area, which eliminates the factor of different development levels of telecommunication infrastructure to a large extent. We focus on a cohort of customers that were (any point in time) subscribers to the carrier's service network within a time window from October 2013 to September 2015, a total of 24 months. This observation window includes 24 periods, as we record customers' behaviors and other related factors at the end of each month. The data comprise 217,726 subscribers. Since the data became uncollectable once subscribes terminate their relationship with the carrier and no conclusive longitudinal results could be obtained with data from a single observation period, we observe customers for a minimum of two and a maximum of 24 periods.

In each period, two behaviors are observed for every customer: their decision on whether or not to change phone tiers (including whether it is an upgrade or a downgrade), and their decision on whether or not to change service plans. Phone tiers are formed by categorizing mobile phones into different tiers mainly based on their prices, which showed little fluctuation during the observation window. In addition, we observe some general telecom service usage of customers, including total phone call time (made or received) in minutes, total number of SMSs (sent or received), and the total mobile Internet data usage in KB. Other time-varying factors were also recorded in the data, i.e., the tenure length of the customer with the carrier, late fee frequencies of the customer, and the number of available service plans/mobile phones etc.

In the data, the majority (169,648 subscribers, 77.9%) of customers ended up unsubscribing from the carrier within the observation window. Customers unsubscribing from the carrier during the observation window on average spends 9.41 months (with a standard deviation of 5.56) with the carrier, which differs significantly from customers who

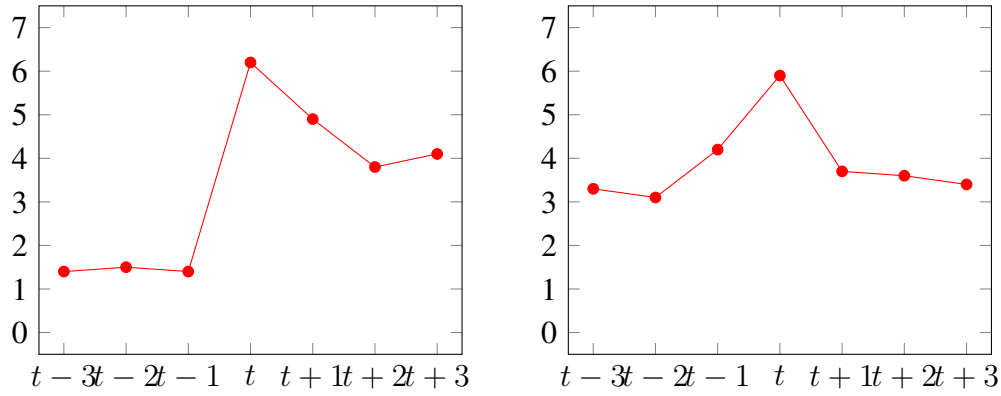
remain with the carrier at the end of the observation window. For customers who did not unsubscribe, the average time with the carrier is 23.94 months with a standard deviation of 0.25. The two groups of customers in terms of whether or not unsubscribe during the observation window also differ considerably in the frequency of changing service plans, where customers did not unsubscribe on average change service plan 0.88 times during their time with the carrier, compared to those who unsubscribed at 0.28. This observation seemingly suggests that there exists an association between customers' loyalty to the carrier and their frequency of changing service plans.

During our observation window, a total of 2,746,398 instances (person$\times$month) were rerecorded in the data, among which, the change of phone tiers and service plans occur in 3.1% and 3.3% of these instances, respectively. We categorize mobile phones into five different tiers, from low to high, according to their prices. It is observed in the data that probabilities of customers purchasing new phones while using mobile phones in these five tiers are 2.6%, 2.6%, 3.3%, 4.7%, and 5.6%, respectively. This observation roughly implies that there might exist an association between customers' financial capability and their behavior of phone change.

The behaviors of changing mobile phone tiers and service plans also appear to exhibit associations when observed longitudinally. In specific, we can compute the empirical probability of one behavior occurring conditional on the other behavior is observed in the same period (t). In a similar fashion, the empirical probability of a behavior occurring can also be computed for times when the other behavior occurs after the next or before the previous certain number of periods (in period $t-n$ or $t+n$). We computed empirical probabilities for both behaviors and plotted our findings in Figure 4.1.

In the left panel of Figure 4.1, we see that, for a customer who changes phone tiers at time t, there is a considerable increase in the empirical probability of her switching service plans at the same time (the empirical probability at time t is over 6%). This could be partially due to the fact that a large number of service plan sales are bundled

together with mobile phone sales. However, it is also noted that this increase in service plan changing probability persists into times several observation periods after the phone change has occurred (empirical probabilities of changing service plans are less than 2% before the phone change, but grow to around 4% after the phone change). In contrast, the right panel of Figure 4.1 shows the empirical probabilities of phone change conditional on a service plan change occurs at time t, where the phone changing probability spikes in the same period of service plan change large due to the bundling sale strategy (the empirical probability at time t is over 6%). The persistent impact of service plan changes on the behavior of phone change, however, does not seem to exist in observation periods after the plan change occurs (empirical probabilities of changing mobile phones are around 4% before and after the service plan change). From this observation, we postulate that there could be longitudinal dependence between the corresponding latent attributes behind the two observed behaviors, which we later capture and analyze using our proposed CNHMM framework.



(a) Plan change probability (%) in neighboring time periods conditional on phone change happens at time t

(b) Phone change probability (%) in neighboring time periods conditional on plan change happens at time t

Figure 4.1: Empirical Distributions of Plan Change Probability (%) and Phone Change Probability (%) Conditional on the Occurrence of the Other Behavior.

There are certain limitations with the model-free analysis presented: (i). we are making

inference about individual-level customer dynamics from aggregate data patterns without controlling for individual heterogeneity, and (ii). the patterns analyzed are largely qualitative, which leads to suggestive results rather than more substantiated explanatory mechanism. Therefore, in the ensuing sections, we utilize our proposed modeling framework to address these problems by incorporating individual heterogeneity and by providing more quantitative and detailed explanations for the findings.

4.4.2 Covariates

In this section, we introduce covariates included in each part of the model. The model comprises two main components: (i). the two observed behaviors (mobile phone change and service plan change), and (ii). the latent states evolution (affecting the two behaviors) over time. While some covariates are more likely to impact customers' behavior instantaneously, other variables might affect their decision-making long-term in the future. For example, if a customer has been continually having more phone calls or SMSs with people who are not in the carrier's service network, she becomes less likely to stay with the carrier and search for other service plans in the future. We account for both types of effects by including variables that are expected to have a short-term impact via the observation part of the model and variables that are likely to have a longer-term impact in the future via part of the model that describes latent state transition dynamics.

Observed Behaviors

We consider several variables that might have effects on customers' decision on changing mobile phones. One set of variables included in the model are the customer's total usage volumes of three primary telecom services, namely, phone calls in minutes ($Call\ Vol_{i,t_i}$), SMSs ($SMS\ Vol_{i,t_i}$), and mobile Internet data in KB ($Net\ Vol_{i,t_i}$). Customers' decision to upgrade or downgrade could partially depend on whether their current mobile phones

matches well with the functions (e.g., mobile Internet) they desire to use. For example, customers who exhibit high usage of mobile Internet data are more likely to upgrade if their current mobile phones do not support high-speed mobile Internet networks.

The using time of current mobile phone ($Phone\ Using\ Time_{i,t_i}$) is another variable that may factor into customers' decision process of changing phones, where a sense of attachment or convenience is often established if the customer has been using her current mobile phone for a longer duration, leading to a reduced change of her changing phones. Customers' affinity for technology may also be an influential factor in their decision to change mobile phones. To capture this effect, we included two related variables, where one dummy variable ($Smartphone_{i,t_i}$) indicates whether or not the current mobile is a smartphone, and another variable ($iPhone\ Release_{i,t_i}$) measures the time between the current observation period and the nearest iPhone release date. Here we use iPhone as a surrogate for an influential mobile phone that represents high-end technology with a large base of followers, and its release might sway customers' decision of changing phones substantially. Besides aforementioned variables, we included another variable ($Available\ Phone\ Choices_{i,t_i}$), which represents the number of different available mobile phone types offered by the carrier, as a control covariate. We assume that both stages of the phone changing behavior might be influenced by the same set of variables. Therefore, with reference to (4.6) and (4.7), we have

$$\boldsymbol{X}^c_{i,t_i} = \boldsymbol{X}^p_{i,t_i} = [Call\ Vol_{i,t_i}, SMS\ Vol_{i,t_i}, Net\ Vol_{i,t_i}, Phone\ Using\ Time_{i,t_i}, \\ Smartphone_{i,t_i}, iPhone\ Release_{i,t_i}, Available\ Phone\ Choices_{i,t_i}]$$

For the customer behavior of changing service plans, we expect it to also be affected by customer's total usage volumes of three primary telecom services, phone calls in minutes ($Call\ Vol_{i,t_i}$), SMSs ($SMS\ Vol_{i,t_i}$), and mobile Internet data in KB ($Net\ Vol_{i,t_i}$). We also include the variable, $Available\ Phone\ Choices_{i,t_i}$, representing the number of different

available telecom service plans offered by the carrier, as a control covariate. With reference to (4.8), we have

$$\boldsymbol{X}_{i,t_i}^{pl} = [Call\ Vol_{i,t_i}, SMS\ Vol_{i,t_i}, Net\ Vol_{i,t_i}, Available\ Phone\ Choices_{i,t_i}]\,.$$

Summary statistics of covariates included in the observation sub-model are given in Table 4.1.

Table 4.1: Summary statistics of covariates included in the observation sub-model ($N = 2,746,398$).

Covariate	Minimum	Maximum	Mean	Std. Dev.	Median
Call Volume (in mins)	0.00	9721.90	197.61	283.47	104.93
SMS Volume	0.00	21788.00	92.44	154.32	42.00
Net Volume (in KB)	0.00	107802.66	192.73	713.72	16.72
Phone Using Time (in days)	0.00	2717.00	410.90	354.38	313.00
Smartphone (dummy)	0.00	1.00	0.83	0.38	1.00
iPhone Release (in days)	0.00	181.00	92.69	54.07	92.00
Available Phone Choices	255.00	289.00	277.20	11.22	281.00
Available Plan Choices	189.00	243.00	216.20	19.03	212.00

Latent State Transitions

As discussed previously, variables that are more likely to affect customers' behaviors long-term in the future are included in the latent state transition part of the model. For the latent process representing the regime-switching of customers' monetary value, we include each customer's average revenue per user (ARPU) for the three primary telecom services, phone call ($Call\ ARPU_{i,t_i}$), SMS ($SMS\ ARPU_{i,t_i}$), and mobile Internet data ($Net\ ARPU_{i,t_i}$) as

part of the variables that may affect changes of the customer's latent monetary value. ARPU values in each separate category contains information of customers revenue generating value to the carrier, which can be integrated into a unified customer monetary value by being incorporated into the transition model. We add a dummy variable ($Roaming_{i,t_i}$) indicating whether or not the customer spends significant time in a roaming status[2], which might impact customer monetary value. Since our modeling framework allows dependence between the two latent processes, dummy variables ($\mathbb{I}\{k^r = 2\}, \ldots, \mathbb{I}\{k^r = K^r\}$) indicating the value of the other latent variable, Z^r_{i,t_i-1}, is also included in the transition covariates. Therefore, with reference to (4.2), we have

$$\begin{aligned}\boldsymbol{W}^m_{i,t_i}(k^r) = &[Call\ ARPU_{i,t_i}, SMS\ ARPU_{i,t_i}, Net\ ARPU_{i,t_i},\\ &\mathbb{I}\{k^r = 2\}, \ldots, \mathbb{I}\{k^r = K^r\}]\end{aligned}$$

For the latent process representing the evolution of customers' latent relational value to the carrier, we include variables that are expected to affect the customers' relationship with the firm. One variable we considered is a customer's tenure length with the carrier ($CI\ Tenure_{i,t_i}$), where longer tenure with carrier creates stronger sense of affinity or convenience, leading to higher relational value. The other variable ($Call-In-Call-Out_{i,t_i}$) included is a variable indicating the customer's communication patterns with others who are within or outside of the carrier's service network. $Call-In-Call-Out_{i,t_i} = 1$ if customer i has more within-network phone call time in period t_i; $Call-In-Call-Out_{i,t_i} = -1$ if the customer has more out-of-network phone call time in period t_i; $Call-In-Call-Out_{i,t_i} = 0$ if the customer has equal phone call time in both categories in period t_i. In addition, we included carrier's service phone call frequency with the customer ($Service\ Call\ Freq_{i,t_i}$) and the customer's late fee frequency ($Late\ Fee\ Freq_{i,t_i}$)

[2]Every telecom service account has a registration area. When the customer is using telecom service outside of the registration area, a roaming fee is incurred. The roaming fee exists for all Chinese telecom carriers.

to capture the effect of qualities of past interactions between the customer and the carrier on the customer's latent relational value. Similar to the monetary value latent process, we also include dummy variables ($\mathbb{I}\{k^m = 2\}, \ldots, \mathbb{I}\{k^m = K^m\}$) indicating the value of the other latent variable, Z^m_{i,t_i-1}, in the transition covariates. With reference to (4.3), we thus have

$$\boldsymbol{W}^r_{i,t_i}(k^m) = [CI\ Tenure_{i,t_i}, Call - In - Call - Out_{i,t_i}, Service\ Call\ Freq_{i,t_i},$$
$$Late\ Fee\ Freq_{i,t_i}, \mathbb{I}\{k^m = 2\}, \ldots, \mathbb{I}\{k^m = K^m\}] .$$

Table 4.2 reports the summary statistics for covariates included in the transition sub-model.

Table 4.2: Summary statistics of covariates included in the transition sub-model ($N = 2,746,398$).

Covariate	Minimum	Maximum	Mean	Std. Dev.	Median
Call ARPU	0.00	18020.05	61.76	59.08	51.00
SMS ARPU	0.00	704.40	0.85	4.13	0.00
Net ARPU	0.00	2239.36	10.02	28.06	0.00
Roaming (dummy)	0.00	1.00	0.13	0.34	0.00
CI Tenure	0.00	333.00	43.45	61.82	23.00
Call-In-Call -Out	-1.00	1.00	-0.56	0.75	-1.00
Service Call Freq	0.00	706.00	0.95	2.93	0.00
Late Fee Freq	0.00	654.00	9.05	7.43	7.00

4.4.3 Model Selection

We split the data into a calibration period (from October 2013 to March 2015) and a validation period (from April 2015 to September 2015). We conduct the model estimation varying the number of states from one to four in both latent Markov chains and compute the deviance information criterion (DIC) (Spiegelhalter et al., 2002), in-sample mean squared error (ISMSE), and in-sample prediction accuracy (ISPA), as well as two out-of-sample

metrics: out-of-sample mean squared error (OSMSE) and out-of-sample prediction accuracy (OSPA) to select a suitable model. We use these metrics to select number(s) of states for the CNHMM, the factorial HMM (FHMM), and the nonhomoenegous HMM (NHMM) with one latent Markov chain.

With reference to Table 4.12 in section 4.6.4 of the Appendix, the CNHMM with the best DIC and ISMSE is the model with three latent states in both Markov chains, whereas models with other numbers of states occasionally show better performances in ISPA. As for out-of-sample metrics, the CNHMM with three latent states in both Markov chains displays the best performance in both OSPA and OSMSE. We therefore use the specification with three latent states in both Markov chains as the model for data analysis. The model selection results for the FHMM is shown in Table 4.13 in section 4.6.4 of the Appendix, where the FHMM with four monetary value states and two relational value states are chosen because it gives the best performance in OSMSE and OSPA. While the FHMM with four states in both Markov chains performs slightly better in DIC, ISMSE, and ISPA, we chose the chose the more parsimonious specification with fewer number of states. Table 4.14 in section 4.6.4 of the Appendix shows the model selection results of the NHMM, where a three state model performs the best in multiple metrics, including DIC, ISMSE, and OSMSE, and among the best in other metrics. In addition to the three latent state models, we also estimate a logistic regression model without the latent state dynamics. The model comparison results for these four competing models are shown in Table 4.3 below. The proposed CNHMM compared favorably to other model specifications in both in-sample and out-of-sample metrics, where it shows lower DIC, ISMSE, and OSMSE, and higher ISPA and OSPA.

4.4.4 Model Estimation Results

In this section, we present our estimation results by first summarizing the state-specific behaviors for each latent state in order to characterize them. This is then followed by a

Table 4.3: Model comparison results.

Model	DIC	ISMSE	ISPA	OSMSE	OSPA
CNHMM	54127.10	0.059	0.978	0.169	0.933
FHMM	56982.58	0.061	0.979	0.177	0.930
HMM	55618.99	0.060	0.977	0.179	0.923
Logistic Regression	98093.10	0.165	0.848	0.357	0.641

discussion of the effects of covariates on customer behavior in each of the latent states and latent state transition dynamics. Values of covariates are standardized before entering into the model to prevent disproportional impacts of certain variables.

4.4.5 Characterizing Latent States

Table 4.4 presents estimated averages of customers' probabilities of changing mobile phone tiers, upgrading mobile phones, and changing service plans for each of the hidden states. Following previous modeling assumption, customers' phone changing behavior

Table 4.4: Estimated averages of state-specific behaviors.

Monetary Value State	Prob (Change Phone Tier)	Prob (Upgrade)	Relational Value State	Prob (Change Service Plan)
M1	0.054	0.577	R1	0.049
M2	0.035	0.490	R2	0.055
M3	0.047	0.334	R3	0.084

is mainly associated with their latent monetary value states, and customers' service plan changing behavior is primarily affected by their latent relational value states. Customers in monetary value state M1 show the highest probability of changing mobile phone tiers ($\mathbb{P}(\text{Change Phone Tier}) = 0.054$) as well as highest probability of upgrading their mobile phones ($\mathbb{P}(\text{Upgrade}) = 0.577$). These two probabilities both decrease on average when customers transition from state M1 to either state M2 or state M3. Customers in monetary state M2 are characterized by having the lowest probability of changing

phone tiers ($\mathbb{P}(\text{Change Phone Tier}) = 0.035$) and the median probability of upgrading ($\mathbb{P}(\text{Upgrade}) = 0.490$). In comparison, customers in monetary state M3 have the median probability of changing phone tiers ($\mathbb{P}(\text{Change Phone Tier}) = 0.047$) and the lowest probability of upgrading ($\mathbb{P}(\text{Upgrade}) = 0.334$). Since customers in states M1, M2, and M3 respectively show the lowest to the highest average probability of upgrading their mobile phones, we correspondingly label states M1, M2, and M3 as latent states with high, moderate, and low monetary values. As for service plan changing behaviors, customers in relational value state R1 shows the lowest average probability of changing service plans ($\mathbb{P}(\text{Change Service Plan}) = 0.049$). This probability increases as customers transition from state R1 to states R2 ($\mathbb{P}(\text{Change Service Plan}) = 0.055$) and R3 ($\mathbb{P}(\text{Change Service Plan}) = 0.084$). We therefore label states R1, R2, and R3 as latent states with low, moderate, and high relational values, respectively.

Covariate Effects on Customer Behaviors

We now discuss how certain factors affect customers' state-specific behaviors of changing mobile phones and service plans. Table 4.5 presents the estimates and standard deviations for the effects of covariates on the behavior of changing phone tiers. In model specification, we allow the effects of covariates to be state-specific. Hence, effects of the same covariate could vary across columns, depending on which latent state customers occupy.

From Table 4.5, we can see that for customers in the highest and lowest monetary value states (M1 and M3), if the current phone is a smartphone, they are less likely to switch for a mobile phone in a different tier. For customers in the lowest monetary state (M3) who are already using a smartphone, there is a higher chance that their need for technology is readily satisfied by their current equipment; and a lack of monetary budget could further prevent them from switching to phones in other tiers. For customers in the highest monetary state (M1) who are already using smartphones, their current equipment is likely to be more technologically advanced mobile compared to other customers. Hence there could be a

Table 4.5: Covariate effects on changing mobile phone tiers.

Covariate	Monetary Value States		
	M1 ("highest" monetary value)	M2	M3
Call Volume	−0.151	0.113	0.288
	(0.481)	(0.761)	(0.532)
SMS Volume	−0.134	−0.050	−0.093
	(0.505)	(0.751)	(0.522)
Net Volume	−0.110	0.113	0.258
	(0.447)	(0.705)	(0.484)
Phone Using Time	−1.240**	−0.903	−1.226**
	(0.529)	(0.734)	(0.584)
Smartphone	−1.874**	−1.891**	−1.907**
	(0.559)	(0.720)	(0.588)
iPhone Release	−0.039	0.089	−0.282
	(0.447)	(0.730)	(0.512)
Available Phone Choices	0.248	0.958	0.343
	(0.439)	(0.653)	(0.530)

1. ** indicates 95% CIs not including 0.
2. Standard deviations are given in parentheses.

lack of motivation for them to switch to phones that belong to other tiers. Decisions of customers in the moderate monetary states (M2) are less likely to be affected by the type of their current mobile phones. The effects of phone using time are negative across all three latent states, which substantiates our assumption that longer using time builds up familiarity and reduces customers' chances of switching to other phones in different tiers.

Results of covariate effects on the behavior of upgrading mobile phones are presented in Table 4.6. Customers in all three latent monetary value states are more likely to upgrade their mobile phones with increased usage in primary telecom services, phone calls, SMSs, and mobile Internet data. Increased usages appear to have largest impacts on customers in the moderate monetary value state (M2).

For customers who have made the decision to change phone tiers, longer using time of their current mobile phones seem to lower their probabilities of upgrading. This could

Table 4.6: Covariate effects on upgrading mobile phones.

Covariate	Monetary Value States		
	M1 ("highest" monetary value)	M2	M3
Call Volume	0.336^{**}	0.950^{**}	0.747^{**}
	(0.171)	(0.303)	(0.106)
SMS Volume	0.536^{**}	0.939^{**}	0.805^{**}
	(0.075)	(0.213)	(0.094)
Net Volume	0.328^{**}	0.920^{**}	0.776^{**}
	(0.069)	(0.135)	(0.142)
Phone Using Time	-0.920^{**}	-1.209^{**}	-1.102^{**}
	(0.228)	(0.282)	(0.144)
Smartphone	0.144	0.244	0.112
	(0.955)	(0.257)	(0.144)
iPhone Release	-0.763^{**}	-0.908^{**}	-0.575^{**}
	(0.230)	(0.169)	(0.208)
Available Phone Choices	0.865^{**}	0.769^{**}	0.869^{**}
	(0.272)	(0.109)	(0.275)

1. ** indicates 95% CIs not including 0.
2. Standard deviations are given in parentheses.

be caused by their familiarity established with their current equipment, which curbs their enthusiasm, to certain extent, on pursuing expensive phones that offer additional functions which they might not use. On the other hand, however, the impact of newly released hi-tech mobile phones might not be overlooked, as the closer it is to the iPhone release data, the more likely for customers to upgrade their mobile phones. The number of available mobile phone types offered by the carrier also has positive effects on customers upgrading mobile phones, since it is more likely for customers to search for mobile phones that offer significantly more functions with slightly higher prices when given more choices.

Table 4.7 summarizes the covariate effects on the customer behavior of changing service plans. Customers in the highest relational value state (R3) are likely to be affected by their usages of SMSs and mobile Internet data. When customers are enjoying good relationships with the carrier, they tend to search for more suitable service plans when they

are experiencing increased telecom service usage. For customers in all three relational value states, an increased number of service plans could also increase their likelihoods in switching to a more suitable plan according to their own situations, which corroborates our assumption.

Table 4.7: Covariate effects on changing service plans.

Covariate	Monetary Value States		
	R1 ("lowest" relational value)	R2	R3
Call Volume	-0.069	0.044	0.298
	(0.147)	(0.145)	(0.090)
SMS Volume	-0.005	-0.104	0.629^{**}
	(0.182)	(0.135)	(0.189)
Net Volume	0.059	0.036	0.128^{**}
	(0.056)	(0.059)	(0.066)
Available Phone Choices	0.710^{**}	0.769^{**}	1.707^{**}
	(0.013)	(0.015)	(0.058)

1. ** indicates 95% CIs not including 0.
2. Standard deviations are given in parentheses.

Covariate Effects on Latent State Transition

Covariate effects on latent monetary state transition dynamics are given in Table 4.8. Recall that coefficients corresponding to the first latent state in (4.2) and (4.3) are set to zero for identification purposes. Therefore, covariate effects, $\boldsymbol{\rho}_{k^m}^m$, can be interpreted as increasing the value of a covariate renders it more or less likely for customers to transit to state k^m than to state 1. In Table 4.8, we can see that ARPU values for primary telecom services have negative effects on customers transitioning to latent states with lower monetary values (states M2 and M3). The effects of phone call ARPU are significant for transitions to both moderate monetary value state and low monetary value state, whereas SMS ARPU and mobile Internet data ARPU are significant for transition to the low monetary value state. Effects of dummy variables representing membership of the relational value state are not

statistically significant, which implies that the transition dynamics of customers' latent monetary value states do not depend on their latent relational value.

Table 4.8: Covariate effects on monetary value latent state transition.

Covariate	Monetary Value States	
	M2	M3
Call ARPU	-1.107^{**}	-2.143^{**}
	(0.509)	(0.607)
SMS ARPU	-1.053	-1.584^{**}
	(0.813)	(0.778)
Net ARPU	-1.121	-1.389^{**}
	(0.619)	(0.535)
Roaming	-1.493	-1.294
	(1.391)	(1.333)
R-State-1	-1.024	0.192
	(1.237)	(1.120)
R-State-2	-0.277	0.172
	(1.142)	(1.203)

1. ** indicates 95% CIs not including 0.
2. Standard deviations are given in parentheses.

Table 4.9 shows the covariate effects on relational value latent state transition. Customers are more likely to transition to higher relational value states if they have longer tenure with the carrier, which corroborates our initial assumption. It is also observed from the results that customers are more likely to transition to latent states with higher relational values if customers have higher degree of dependence of the carrier's because majorities of their contacts are within the carrier's network. Service quality of the carrier could play a significant role in increasing customers' likelihoods of transitioning to higher relational value states. Higher service call frequencies and lower frequencies of late fees can both shift customers to having higher relational values to the carrier.

Different from the latent monetary value states transition, dummy variables representing membership of the monetary value state in fact show significant effects on the transition dynamics of customers' latent relational value states. In specific, the customer is less likely

to transition to states with higher relational values in the future if she is currently in a lower monetary value state. Such effects are significant for transitions to both moderate relational value state (R2) and high relational value state (R3). This result implies that the evolution of customers' latent relational value, to certain extent, depends on their latent monetary value. However, the evolution of customers' latent monetary value shows no significant dependence on their relational value. This finding can lead to various managerial implications for the carrier to implement with the goal of achieving more effective CRM and increasing profits in the long-run.

Table 4.9: Covariate effects on relational value latent state transition.

Covariate	Relational Value States	
	R2	R3
CI Tenure	0.920^{**}	0.787^{**}
	(0.087)	(0.357)
Call-in-Call-out	1.212^{**}	2.201^{**}
	(0.163)	(0.128)
Service Call Freq	2.459^{**}	0.585^{**}
	(0.023)	(0.162)
Late Fee Freq	-0.826^{**}	-1.304^{**}
	(0.071)	(0.071)
M-State-1	-1.125^{**}	-1.861^{**}
	(0.086)	(0.109)
M-State-2	-1.513^{**}	-1.422^{**}
	(0.243)	(0.352)

1. ** indicates 95% CIs not including 0.
2. Standard deviations are given in parentheses.

4.4.6 Scenario Analyses

In this section, we present some examples of possible managerial implications that can be gained for the carrier using our proposed CNHMM framework. We start by recovering latent state memberships using model estimation results. Managerial implications are then given via scenario analyses where we compare effectiveness of different targeting strategies

based on the model.

Recovering Latent State Memberships

We use parameter estimates to compute the probability of each customer belonging to each state at any time period. The forward-backward algorithm is adopted to calculate the posterior probability that customer i is in monetary value state k^m and relational value state k^r at time t_i. Details of the forward-backward algorithm is available in section 4.6.3 of the Appendix. We then averaging these probabilities across customers to obtain estimates of the proportion of the customer base in each state pairing at any time period. The evolution of these proportions over time during the calibration period is given in Table 4.10.

Table 4.10: Latent state membership proportions.

Time	Latent State Pairings								
	M1R1	M1R2	M1R3	M2R1	M2R2	M2R3	M3R1	M3R2	M3R3
2013/11	0.282	0.089	0.015	0.025	0.005	0.035	0.416	0.084	0.050
2013/12	0.332	0.025	0.017	0.019	0.001	0.039	0.502	0.021	0.043
2014/01	0.279	0.079	0.015	0.017	0.004	0.056	0.449	0.063	0.040
2014/02	0.289	0.094	0.014	0.023	0.006	0.062	0.413	0.065	0.035
2014/03	0.260	0.083	0.013	0.018	0.004	0.038	0.471	0.074	0.038
2014/04	0.268	0.094	0.014	0.025	0.006	0.026	0.450	0.078	0.039
2014/05	0.264	0.094	0.015	0.022	0.005	0.025	0.459	0.074	0.042
2014/06	0.267	0.093	0.016	0.022	0.005	0.026	0.455	0.073	0.043
2014/07	0.261	0.096	0.017	0.024	0.005	0.037	0.441	0.076	0.043
2014/08	0.265	0.101	0.015	0.034	0.006	0.029	0.428	0.078	0.043
2014/09	0.258	0.098	0.015	0.056	0.005	0.019	0.429	0.077	0.044
2014/10	0.249	0.088	0.014	0.070	0.006	0.005	0.449	0.076	0.043
2014/11	0.252	0.081	0.012	0.066	0.006	0.003	0.464	0.073	0.042
2014/12	0.239	0.072	0.013	0.070	0.006	0.003	0.477	0.074	0.047
2015/01	0.239	0.073	0.013	0.066	0.006	0.004	0.478	0.073	0.050
2015/02	0.241	0.074	0.013	0.068	0.007	0.004	0.469	0.073	0.050
2015/03	0.257	0.076	0.013	0.030	0.008	0.001	0.495	0.074	0.046

Understanding the evolution of customers' latent state membership could help the carrier gain a better understanding of the customer base. For example, we find that proportions of

customers within the calibration period remain relatively stable. In specific, high monetary, moderate, and low monetary value customers respectively take up around 37%, 6%, and 57% of the customer base with certain fluctuations from time to time. There exists a gradual decline in the proportion of customers in the high monetary value state, accompanied by a gradual increase in the proportion of customers in the low monetary value state, with the proportion of customers in the moderate monetary value state staying relatively stable. For relational value states, customers occupy low, moderate, and high relational states at proportions of roughly 75%, 15%, and 10% in each time period. There appears to be a gradual increase in the proportion of customers in the low relational value state, accompanied by a gradual decline in the proportion of customers in the high relational value state, with the proportion of customers in the moderate relational value state staying relatively stable. In the next section, we explore the strategies in terms of marketing actions that the carrier could potentially implement based on customers' latent state memberships to achieve better CRM and improved profitability.

Scenario Analyses

We use scenario analyses to demonstrate how the model can help the carrier increase customers' activity levels in changing and upgrading mobile phones as well as changing service plans. We compare results from strategies based on the model with other approaches using the validation data (from April to September 2015, six months in total). Specifically, we first estimate the latent state probability at the end of the calibration period for every customer in the validation data, and assign her to a latent state pairing if the probability of belonging to that paring is the largest. Using estimated parameters and transition covariates, the latent state probability at the beginning of the validation period for each included customer can also be estimated and a latent state pairing can be assigned to her in a similar fashion. Sequentially, latent state membership at each period in the validation data for all included customers can be then estimated.

Since we have obtained estimates of latent state membership at each period in the validation data, we can then estimate the behavior probabilities of phone mobile change, mobile phone upgrade, and service plan change for each customer. We then compute average behavior probabilities across all customers for each time period in the validation data. These average behavior probability estimates can be obtained for all competing strategies. If certain average behavior probabilities from the strategy based on our model are greater than those from competing strategies, then it could be said that our model is able to provide benefits for the carrier to make more effective managerial decisions.

Stimulation with Promotion Packages

A commonly adopted marketing action is to give out free promotion packages with low cost in order to stimulate customers into purchasing products and thus achieves higher profits. For the telecom carrier, such low-cost promotion packages that can be given out freely often comprises of mobile usage quotas. In this analysis, we simulate a scenario where the carrier is interested in increasing the probabilities of customers purchasing more expensive mobile phones and switching service plans via giving out free mobile usage quotas, including phone call minutes, SMSs, and mobile Internet data. We here make the assumption that customers will use up all the free quotas given to them and their original telecom service volumes observed in the validation data.

We consider three competing strategies, i.e., status quo (SQ), universal targeting (UT), and segmentation (SG). Status quo refers to the strategy where the carrier does not give out free quotas, and customers only use up their original telecom service volumes. Universal targeting refers to the carrier gives out the same free quotas to all customer, and segmentation is the strategy based on the model, where only certain segments of the customers are selected to be given free quotas.

In this analysis, we assume the carrier gives out an extra 1/3 of the standard deviation of phone call minutes, SMSs, and mobile Internet data to every customer in the universal

targeting strategy. From table 4.10, we found that, on average, around 10% of the customers belong to the moderate monetary value state (M2) or the high relational value state (R3) at each time period. Because mobile usages have significant effects on customers in state M2 or R3, they are chosen as the targeted group. Hence for the segmentation strategy, we assume the carrier gives out all the free quotas, which would have been given out to all customers in the universal targeting strategy, only to customers in state M2 or R3, and does not give out free quotas to other customers. We measure the effectiveness of the strategies using differences in behavior probabilities, which are obtained by subtracting corresponding behavior probabilities under the status quo strategy from the other two. A more effective strategy is expected to yield more positive probability differences. Results are given in Table 4.11.

Table 4.11: Scenario analysis results for free promotion packages (free quotas only).

Time	Δ Phone Change		Δ Phone Upgrade		Δ Plan Change	
	UT	SG	UT	SG	UT	SG
2015/04	0.000	0.000	0.086	0.107	0.003	0.043
2015/05	0.000	0.000	0.084	0.103	0.005	0.047
2015/06	0.000	0.000	0.080	0.096	0.005	0.047
2015/07	0.000	0.000	0.075	0.090	0.005	0.050
2015/08	0.000	0.000	0.071	0.087	0.005	0.050
2015/09	0.000	0.000	0.066	0.077	0.006	0.052

UT: Universal Targeting. SG: Segmentation.

Because covariate effects of mobile usages on the behavior of changing mobile phone tiers are not significant, the marketing action of giving out free quotas does not increase customers' likelihoods of changing mobile phone tiers. However, from Table 4.11, we can see that this marketing action could lead to increased probabilities in customers upgrading mobile phones and changing service plans. With the same level of cost, the segmentation strategy, which is more effective, results in a larger increase in both probabilities compared to the UT strategy. The free promotion strategies do not lead to increased probabilities of

customers changing phones because the corresponding covariate effects are not significant. With the model, the carrier can identify a small segment of the customer base and possibly achieve the same or higher level of profitability.

4.5 Discussion

In this paper, we developed a CNHMM framework for a multidimensional and dynamic understanding of customer values. Such CNHMM framework, illustrated through an application in the large-scale telecommunication data, offers several insights into both the drivers of customer value and dynamic CRM.

The first contribution of this research is to suggest a behaviorally grounded model that help marketers to identify multiple underlying sources of customer value. Leveraging a large-scale longitudinal mobile service data, we identify two distinctive latent drivers of customer value, i.e., the relational and monetary values. Indeed, as customer value can be measured beyond the relational and monetary dimensions, our proposed model presents a general framework for marketers to capture multiple sources of heterogeneity in customer value. Future research can also construct and assess alternative or additional dimensions that offer additional managerial insights.

Second, our proposed CNHMM model not only uncovers latent drivers of customer value, but also explicitly identifies the interdependence among different dimensions of customer value. Interestingly, our empirical results suggest that customer monetary value drives relational value but not vice versa, indicating that firms can improve sales performance (e.g. in eliciting more phone purchases) by strategically considering the weights or relative importance of customers' monetary and relational value when allocating scare marketing resources. While the relationship between monetary and relational value of customers may vary across industries, our empirical evidence suggests that the sole reliance on customers' relational dimension in implementing CRMs may lead to incomplete understanding of

customer values.

Third, the dynamic nature of the proposed model allows firms to understand the evolution of customer value, and therefore to implement dynamic segmentation strategies. In our sample, firms can recover the latent monetary and relational states (segments) of their customers (e.g. for every month), and then dynamically trace those customers that have "silently" transferred to another segment each month, and eventually adjust the marketing mix offered to those customers accordingly. With this dynamic approach, marketing resources can be allocated more efficiently over time, with respect to the perceived importance of each segment.

Besides the methodological developments, our paper offers fruitful managerial insights for customer-value based CRM. Our simulation studies suggest the proposed model can help firms formulate customer acquisition strategies. For example, by acquiring customers high in both monetary and relational value, firms can substantially increase the expected revenue from customers' phone purchases and service plan renewals. In addition, firms can tailor make their (targeted) promotions based on the recovered segments of customers, which not only enhances the effectiveness of promotion by improving sales performance (e.g. eliciting more phone purchase), but also leads to higher promotion efficiency as firms may choose to focus on particular (e.g. customers in the moderate-monetary-value segment) instead of universal segments, better allocating their scarce marketing resources.

To summarize, we believe our study provides a first step to explore the multidimensional customer value using HMM-based models. We have also provided CRM practitioners with an implementable model for evaluating the monetary and relational values of customers, as well as their interdependence over time. Such models are necessary today as consumer-generated big data plays an increasingly important role for firms. We encourage future research to continue investigating the multidimensional characterization of customer value and derive managerial-relevant insights in creating a better customer-centric CRM system.

4.6 Appendix

4.6.1 Model Estimation

In this section, we describe the stochastic variational Bayes (SVB) method we used in estimating the coupled non-homogeneous hidden Markov model (CNHMM).

Variational Bayesian Inference

Variational Bayesian (VB) inference is a Bayesian estimation approach that uses density functions from simple distribution families to approximate intractable posteriors (Jordan et al., 1999; Blei et al., 2017).

Given a generic model $p(\boldsymbol{y}|\boldsymbol{\theta})$ with $\boldsymbol{y}$ denoting the observed data and $\boldsymbol{\theta}$ as unknown parameters, the aim of VB is to approximate the intractable posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ through a variational posterior distribution $\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ from a tractable variational distribution family $\widetilde{\mathcal{P}}$, where $\boldsymbol{\phi}$ is the set of variational parameters that govern the variational distribution. The distance between $\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is often measured by the Kullback-Leibler (KL) divergence

$$\begin{aligned}
\mathrm{KL}\left[\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\boldsymbol{y})\right] &= \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}}\left[\log \widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{y})\right] \\
&= \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}}\left[\log \widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{y}, \boldsymbol{\theta})\right] + \log p(\boldsymbol{y}) \\
&= -\mathcal{L}(\boldsymbol{\phi}) + \log p(\boldsymbol{y}).
\end{aligned}$$

Since $\mathrm{KL}\left[\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\boldsymbol{y})\right]$ is nonnegative, equations above imply that $\mathcal{L}(\boldsymbol{\phi}) \leq \log p(\boldsymbol{y})$, $\forall \widetilde{p}_{\boldsymbol{\phi}}$. So $\mathcal{L}(\boldsymbol{\phi})$, being a lower bound for the log marginal likelihood, is called the evidence lower bound (ELBO) function.

The optimal variational posterior can be obtained by maximizing $\mathcal{L}(\boldsymbol{\phi})$ over $\boldsymbol{\phi}$, which is performed via stochastic gradient ascent (SGA) (Robbins and Monro, 1951). Letting

$\nabla\mathcal{L}(\boldsymbol{\phi})$ denote the gradient of $\mathcal{L}(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$, after selecting an initial value $\boldsymbol{\phi}^{(0)}$, $\boldsymbol{\phi}$ is updated via

$$\boldsymbol{\phi}^{(\tau+1)} = \boldsymbol{\phi}^{(\tau)} + \boldsymbol{\psi}_\tau \circ \nabla\mathcal{L}(\boldsymbol{\phi}^{(\tau)}),$$

where superscript (τ) denotes the τ-th iteration, operator $\circ$ denotes the Hadamard (element-wise) product, and $\{\boldsymbol{\psi}_\tau\}_{\tau\geq 0}$ contains a sequence of learning rates satisfying the Robbins-Monro conditions (Robbins and Monro, 1951). Each updating step involves determining values of the learning rates and the gradient. In this study, we apply the well-received *Adam* optimizer for adaptive learning rates; *Adam* is efficient with limited tuning required, capable of handling noisy and/or spare gradients and non-stationary objectives, and particularly suitable for dealing with non-convex objective functions (Kingma and Ba, 2015). Section 4.6.2 gives a more detailed specification of the Adam optimizer.

The complexity and accuracy of the above VB method is significantly influenced by the choice of variational posterior family $\widetilde{\mathcal{P}}$ (Blei et al., 2017). In practice, a trade-off between approximation accuracy and computational complexity is sought. The mean-field family, which assumes complete independence for model parameters, is a common choice, but while it reduces computational complexity, the over-simplified form may lead to suboptimal approximations. Here, we use a structured mean-field family that adds dependence among certain variational posteriors to better approximate the structures of true posteriors. The dependence is introduced based on the factor covariance structure developed by (Ong et al., 2018). Numerical studies in latter sections show that the utilized posterior family performs satisfactorily.

The Stochastic Variational Bayes Estimation of The CNHMM

Following the model setting in the main article, consider the CNHMM with observed response sequence $\{\boldsymbol{Y}_{i,\mathcal{T}_i}\}$ where $\boldsymbol{Y}_{i,t_i} = (Y^c_{i,t_i}, Y^p_{i,t_i}, Y^{pl}_{i,t_i})$, emission covariates sequence

$\{\boldsymbol{X}_{i,\mathcal{T}_i}\}$ with $\boldsymbol{X}_{i,t_i} = (\boldsymbol{X}^c_{i,t_i}, \boldsymbol{X}^p_{i,t_i}, \boldsymbol{X}^{pl}_{i,t_i})$, and transition covariate sequence $\{\boldsymbol{W}_{i,\mathcal{T}_i}\}$ with $\boldsymbol{W}_{i,t_i} = (\boldsymbol{W}^m_{i,t_i}, \boldsymbol{W}^r_{i,t_i})$ for $t_i = 1, \ldots, T_i$ and $i = 1, \ldots, N$ from N independent individuals.

Let $\{\boldsymbol{Z}_{i,\mathcal{T}_i}\}$ denote the hidden states sequences such that $\boldsymbol{Z}_{i,t_i} = (Z^m_{i,t_i}, Z^r_{i,t_i})$. The complete-data likelihood for the CNHMM is

$$
\begin{aligned}
&p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\pi} | \boldsymbol{X}, \boldsymbol{W}) \\
=&\prod_{i=1}^{N} p(\boldsymbol{Z}_{i,1}, \boldsymbol{\pi}) \prod_{t_i=1}^{T_i} p(\boldsymbol{Y}_{i,t_i} | \boldsymbol{Z}_{i,t_i}, \boldsymbol{X}_{i,t_i}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta}) \prod_{t_i=2}^{T_i} p(\boldsymbol{Z}_{i,t_i} | \boldsymbol{Z}_{i,t_i-1}, \boldsymbol{W}_{i,t_i}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho}) \\
&\times p(\boldsymbol{\mu}) p(\boldsymbol{\xi}) p(\boldsymbol{\beta}) p(\boldsymbol{\zeta}) p(\boldsymbol{\eta}) p(\boldsymbol{\rho}) p(\boldsymbol{\pi}),
\end{aligned}
$$

where $p(\boldsymbol{\mu})$, $p(\boldsymbol{\xi})$, $p(\boldsymbol{\beta})$, $p(\boldsymbol{\zeta})$, $p(\boldsymbol{\eta})$, $p(\boldsymbol{\rho})$ and $p(\boldsymbol{\pi})$ are priors, $p(\boldsymbol{Y}_{i,t_i} | \boldsymbol{Z}_{i,t_i}, \boldsymbol{X}_{i,t_i}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\xi})$ is the emission distribution, $p(\boldsymbol{Z}_{i,t_i} | \boldsymbol{Z}_{i,t_i-1}, \boldsymbol{W}_{i,t_i}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho})$ is the transition distribution, and $p(\boldsymbol{Z}_{i,1}, \boldsymbol{\pi})$ is the initial state distribution.

In specific, for model parameters, we have

- $\boldsymbol{\mu} = (\boldsymbol{\mu}^c, \boldsymbol{\mu}^p, \boldsymbol{\mu}^{pl})$, where $\boldsymbol{\mu}^c = \{\mu^c_{k^m} | k^m \in [K^m]\}$, $\boldsymbol{\mu}^p = \{\mu^p_{k^m} | k^m \in [K^m]\}$, and $\boldsymbol{\mu}^{pl} = \{\mu^{pl}_{k^r} | k^r \in [K^r]\}$;

- $\boldsymbol{\xi} = \{\boldsymbol{\xi}^c_i, \boldsymbol{\xi}^p_i, \boldsymbol{\xi}^{pl}_i\}_{i=1,\ldots,N}$, where $\boldsymbol{\xi}^c_i = \{\xi^c_{i,k^m} | k^m \in [K^m]\}$, $\boldsymbol{\xi}^p_i = \{\xi^p_{i,k^m} | k^m \in [K^m]\}$, and $\boldsymbol{\xi}^{pl}_i = \{\xi^{pl}_{i,k^r} | k^r \in [K^r]\}$;

- $\boldsymbol{\beta} = (\boldsymbol{\beta}^c, \boldsymbol{\beta}^p, \boldsymbol{\beta}^{pl})$, where $\boldsymbol{\beta}^c = \{\boldsymbol{\beta}^c_{k^m} | k^m \in [K^m]\}$, $\boldsymbol{\beta}^p = \{\boldsymbol{\beta}^p_{k^m} | k^m \in [K^m]\}$, and $\boldsymbol{\beta}^{pl} = \{\boldsymbol{\beta}^{pl}_{k^r} | k^r \in [K^r]\}$;

- $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}^m_i, \boldsymbol{\zeta}^r_i\}_{i=1,\ldots,N}$, where $\boldsymbol{\zeta}^m_i = \{\zeta^m_{i,k^m_1,k^m_2} | k^m_1 \in [K^m], k^m_2 \in [K^m - 1]\}$ and $\boldsymbol{\zeta}^r_i = \{\zeta^r_{i,k^r_1,k^r_2} | k^r_1 \in [K^r], k^r_2 \in [K^r - 1]\}$;

- $\boldsymbol{\eta} = (\boldsymbol{\eta}^m, \boldsymbol{\eta}^r)$, where $\boldsymbol{\eta}^m = \{\eta^m_{k^m_1,k^m_2} | k^m_1 \in [K^m], k^m_2 \in [K^m - 1]\}$ and $\boldsymbol{\eta}^r = \{\eta^r_{k^r_1,k^r_2} | k^r_1 \in [K^r], k^r_2 \in [K^r - 1]\}$;

- $\boldsymbol{\rho} = (\boldsymbol{\rho}^m, \boldsymbol{\rho}^r)$, where $\boldsymbol{\rho}^m = \{\boldsymbol{\rho}^m_{k^m} | k^m \in [K^m]\}$ and $\boldsymbol{\rho}^r = \{\boldsymbol{\rho}^r_{k^r} | k^r \in [K^r]\}$;

- $\boldsymbol{\pi} = (\boldsymbol{\pi}^m, \boldsymbol{\pi}^r)$, where $\boldsymbol{\pi}^m = \{\pi^m_{k^m} | k^m \in [K^m - 1]\}$ and $\boldsymbol{\pi}^r = \{\pi^r_{k^r} | k^r \in [K^r - 1]\}$.

VB computes the posteriors of $\boldsymbol{\mu}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\zeta}$, $\boldsymbol{\eta}$, $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$, as well as hidden states sequences $\{\boldsymbol{Z}_{i,\mathcal{T}_i}\}$. We first specify the variational posteriors with the factorized form

$$\widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{Z}) = \widetilde{p}_{\boldsymbol{\phi}_\mu}(\boldsymbol{\mu})\widetilde{p}_{\boldsymbol{\phi}_\xi}(\boldsymbol{\xi})\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})\widetilde{p}_{\boldsymbol{\phi}_\zeta}(\boldsymbol{\zeta})\widetilde{p}_{\boldsymbol{\phi}_\eta}(\boldsymbol{\eta})\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})\widetilde{p}_{\boldsymbol{\phi}_\pi}(\boldsymbol{\pi})\widetilde{p}(\boldsymbol{Z}),$$

where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_\mu, \boldsymbol{\phi}_\xi, \boldsymbol{\phi}_\beta, \boldsymbol{\phi}_\zeta, \boldsymbol{\phi}_\eta, \boldsymbol{\phi}_\rho, \boldsymbol{\phi}_\pi\}$ are the variational parameters. We allow dependence within each of $\boldsymbol{\mu}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\zeta}$, $\boldsymbol{\eta}$, $\boldsymbol{\rho}$, $\boldsymbol{\pi}$, and $\boldsymbol{Z}$ to achieve better approximation, and retain independence between $\boldsymbol{\mu}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\zeta}$, $\boldsymbol{\eta}$, $\boldsymbol{\rho}$, $\boldsymbol{\pi}$, and $\boldsymbol{Z}$ for computational tractability. The ELBO objective,

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}}\Big\{ & \mathbb{E}_{\widetilde{p}(\boldsymbol{Z})}[\log p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\pi} | \boldsymbol{X}, \boldsymbol{W}) \\ & - \log \widetilde{p}_{\boldsymbol{\phi}_\mu}(\boldsymbol{\mu})\widetilde{p}_{\boldsymbol{\phi}_\xi}(\boldsymbol{\xi})\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})\widetilde{p}_{\boldsymbol{\phi}_\zeta}(\boldsymbol{\zeta})\widetilde{p}_{\boldsymbol{\phi}_\eta}(\boldsymbol{\eta})\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})\widetilde{p}_{\boldsymbol{\phi}_\pi}(\boldsymbol{\pi})\widetilde{p}(\boldsymbol{Z})]\Big\},\end{aligned}$$

is maximized by updating $\widetilde{p}_{\boldsymbol{\phi}_\mu}(\boldsymbol{\mu})$, $\widetilde{p}_{\boldsymbol{\phi}_\xi}(\boldsymbol{\xi})$, $\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})$, $\widetilde{p}_{\boldsymbol{\phi}_\zeta}(\boldsymbol{\zeta})$, $\widetilde{p}_{\boldsymbol{\phi}_\eta}(\boldsymbol{\eta})$, $\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})$, $\widetilde{p}_{\boldsymbol{\phi}_\pi}(\boldsymbol{\pi})$, and $\widetilde{p}(\boldsymbol{Z})$ iteratively.

The updating of $\widetilde{p}(\boldsymbol{Z})$ involves calculating posteriors of $\boldsymbol{Z}_{i,t_i}$ and $(\boldsymbol{Z}_{i,t_i-1}, \boldsymbol{Z}_{i,t_i})$ with current values of variational parameters. We jointly update the posterior of Z^m_{i,t_i} and Z^r_{i,t_i} due to their dependence on each other. The joint posterior of $\{\boldsymbol{Z}_{i,\mathcal{T}_i}\}$ is proportional to

$$\begin{aligned}\prod_{i=1}^{N} \exp\Bigg\{ & \sum_{t_i=1}^{T_i} \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\mu}(\boldsymbol{\mu})\widetilde{p}_{\boldsymbol{\phi}_\xi}(\boldsymbol{\xi})\widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})} [\log p(\boldsymbol{Y}_{i,t_i} | \boldsymbol{Z}_{i,t_i}, \boldsymbol{X}_{i,t_i}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta})] \\ & + \sum_{t_i=2}^{T_i} \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\zeta}(\boldsymbol{\zeta})\widetilde{p}_{\boldsymbol{\phi}_\eta}(\boldsymbol{\eta})\widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})} [\log p(\boldsymbol{Z}_{i,t_i} | \boldsymbol{Z}_{i,t_i-1}, \boldsymbol{W}_{i,t_i}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho})] \Bigg\}.\end{aligned}$$

Based on the posterior, we compute the marginal posteriors $\widetilde{p}(\boldsymbol{Z}_{i,t_i})$ and $\widetilde{p}(\boldsymbol{Z}_{i,t_i-1}, \boldsymbol{Z}_{i,t_i})$ using the forward and backward probabilities of the Baum-Welch procedure (Baum et al., 1970). Details of the forward-backward algorithm are given in section 4.6.3.

The updating of variational parameters $\boldsymbol{\phi}_{\mu}$, $\boldsymbol{\phi}_{\xi}$, $\boldsymbol{\phi}_{\beta}$, $\boldsymbol{\phi}_{\zeta}$, $\boldsymbol{\phi}_{\eta}$, $\boldsymbol{\phi}_{\rho}$, and $\boldsymbol{\phi}_{\pi}$ is conducted through SGA. Here we use $\boldsymbol{\phi}_{\rho}$ and $\boldsymbol{\phi}_{\beta}$ as examples to describe the procedure of updating variational parameters, since other parameters can be updated in a similar fashion. For the updating of $\boldsymbol{\phi}_{\rho}$, we assign the following factorized variational posterior for $\boldsymbol{\rho}$:

$$\prod_{k^m=1}^{K^m}\prod_{k^r=1}^{K^r}\widetilde{p}_{\boldsymbol{\phi}_{\rho^m,k^m}}(\boldsymbol{\rho}_{k^m}^m)\widetilde{p}_{\boldsymbol{\phi}_{\rho^r,k^r}}(\boldsymbol{\rho}_{k^r}^r),$$

where $\widetilde{p}_{\boldsymbol{\phi}_{\rho^m,k^m}}(\boldsymbol{\rho}_{k^m}^m)$ and $\widetilde{p}_{\boldsymbol{\phi}_{\rho^r,k^r}}(\boldsymbol{\rho}_{k^r}^r)$ are R_w^m- and R_w^r-dimensional multivariate normal distributions with factor covariance structures as follows (Ong et al., 2018):

$$\widetilde{p}_{\boldsymbol{\phi}_{\rho^m,k^m}}(\boldsymbol{\rho}_{k^m}^m)=\mathcal{N}\left(\boldsymbol{\rho}_{k^m}^m;\boldsymbol{m}_{\rho^m,k^m},\boldsymbol{G}_{\rho^m,k^m}\boldsymbol{G}'_{\rho^m,k^m}+\boldsymbol{H}^2_{\rho^m,k^m}\right),$$

and

$$\widetilde{p}_{\boldsymbol{\phi}_{\rho^r,k^r}}(\boldsymbol{\rho}_{k^r}^r)=\mathcal{N}\left(\boldsymbol{\rho}_{k^r}^r;\boldsymbol{m}_{\rho^r,k^r},\boldsymbol{G}_{\rho^r,k^r}\boldsymbol{G}'_{\rho^r,k^r}+\boldsymbol{H}^2_{\rho^r,k^r}\right).$$

Thus, by definition, $\boldsymbol{\phi}_{\rho^m,k^m}=\{\boldsymbol{m}_{\rho^m,k^m},\boldsymbol{G}_{\rho^m,k^m},\boldsymbol{H}_{\rho^m,k^m}\}$ and $\boldsymbol{\phi}_{\rho^r,k^r}=\{\boldsymbol{m}_{\rho^r,k^r},\boldsymbol{G}_{\rho^r,k^r},\boldsymbol{H}_{\rho^r,k^r}\}$ are variational parameters to be estimated, where $\boldsymbol{m}_{\rho^m,k^m}$ and $\boldsymbol{m}_{\rho^r,k^r}$ are variational mean vectors for $\boldsymbol{\rho}_{k^m}^m$ and $\boldsymbol{\rho}_{k^r}^r$, $\boldsymbol{G}_{\rho^m,k^m}$ and $\boldsymbol{G}_{\rho^r,k^r}$ are $R_w^m\times r_{\rho^m,k^m}$ and $R_w^r\times r_{\rho^r,k^r}$ matrices with $r_{\rho^m,k^m}\leq R_w^m$ and $r_{\rho^r,k^r}\leq R_w^r$, r_{ρ^m,k^m} and r_{ρ^r,k^r} denote numbers of factors used to approximate the correlation among elements in $\boldsymbol{\rho}_{k^m}^m$ and $\boldsymbol{\rho}_{k^r}^r$, respectively. Upper triangles of $\boldsymbol{G}_{\rho^m,k^m}$ and $\boldsymbol{G}_{\rho^r,k^r}$ are restricted to zero for identification. $\boldsymbol{H}_{\rho^m,k^m}$ and $\boldsymbol{H}_{\rho^r,k^r}$ are diagonal matrices with diagonal elements $\boldsymbol{h}_{\rho^m,k^m}=(h_{\rho^m,k^m,1},\ldots,h_{\rho^m,k^m,R_w^m})'$, and $\boldsymbol{h}_{\rho^r,k^r}=(h_{\rho^r,k^r,1},\ldots,h_{\rho^r,k^r,R_w^r})'$. As demonstrated in Ong et al. (2018), a small number of factors ($r_{\rho^m,k^m}=3$ or 4) already provides satisfactory approximation to high-dimensional posteriors. Indeed, approximation accuracy can be improved if we increase the number of factors r_{ρ^m,k^m} used, which also raises the optimization difficulty; if we set $r_{\rho^m,k^m}=R_w^m$,

we are using a multivariate normal distribution with a full covariance structure, which yields the closest approximation.

With the factorized variational posterior for $\boldsymbol{\rho}$, updating $\boldsymbol{\phi}_{\rho}$ reduces to the iterative update of $\boldsymbol{\phi}_{\rho^m,k^m}$ and $\boldsymbol{\phi}_{\rho^r,k^r}$. To update $\boldsymbol{\phi}_{\rho^m,k^m}$, we need to compute the gradient estimates, $\widehat{\nabla_{\boldsymbol{m}_{\rho^m,k^m}}\mathcal{L}}$, $\widehat{\nabla_{\boldsymbol{G}_{\rho^m,k^m}}\mathcal{L}}$, and $\widehat{\nabla_{\boldsymbol{H}_{\rho^m,k^m}}\mathcal{L}}$, and then update $\boldsymbol{m}_{\rho^m,k^m}$, $\boldsymbol{G}_{\rho^m,k^m}$, and $\boldsymbol{H}_{\rho^m,k^m}$ sequentially according to the formula. Similarly, for $\boldsymbol{\phi}_{\rho^r,k^r}$, compute $\widehat{\nabla_{\boldsymbol{m}_{\rho^r,k^r}}\mathcal{L}}$, $\widehat{\nabla_{\boldsymbol{G}_{\rho^r,k^r}}\mathcal{L}}$, and $\widehat{\nabla_{\boldsymbol{H}_{\rho^r,k^r}}\mathcal{L}}$, then update $\boldsymbol{m}_{\rho^r,k^r}$, $\boldsymbol{G}_{\rho^r,k^r}$, and $\boldsymbol{H}_{\rho^r,k^r}$. The gradients are computed using the reparametrization approach, which was introduced to control variance for the Monte Carlo gradient. It is applicable when the generic model parameter $\boldsymbol{\theta}$ can be represented as $\boldsymbol{\theta} = t(\boldsymbol{\phi}, \boldsymbol{\upsilon})$, where $\boldsymbol{\upsilon}$ denotes a random vector with a known fixed distribution $f(\boldsymbol{\upsilon})$. For instance, suppose $\boldsymbol{\theta}$ follows a multivariate normal variational distribution, then it can be reparametrized by the mean vector $\boldsymbol{m}$, the lower Cholesky factor $\boldsymbol{L}$ of its covariance matrix, and a standard normal random vector $\boldsymbol{\upsilon}$ as $\boldsymbol{\theta} = \boldsymbol{m} + \boldsymbol{L}\boldsymbol{\upsilon}$. After reparametrization, the ELBO objective function can be rewritten as

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}}}\left[\log p(\boldsymbol{y}, \boldsymbol{\theta}) - \log \widetilde{p}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\right] = \mathbb{E}_{f}\left[\log p(\boldsymbol{y}, t(\boldsymbol{\phi}, \boldsymbol{\upsilon})) - \log \widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi}, \boldsymbol{\upsilon}))\right],$$

and the gradient becomes

$$\nabla_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{f}\left\{\frac{\mathrm{d}t(\boldsymbol{\phi}, \boldsymbol{\upsilon})}{\mathrm{d}\boldsymbol{\phi}}' \nabla_{\boldsymbol{\theta}}\left[\log p(\boldsymbol{y}, t(\boldsymbol{\phi}, \boldsymbol{\upsilon})) - \log \widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi}, \boldsymbol{\upsilon}))\right] - \nabla_{\boldsymbol{\phi}} \log \widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi}, \boldsymbol{\upsilon}))\right\}.$$

The variance of the gradient estimate can be further reduced by dropping the last term in the equation above (Ong et al., 2018); the final formula for estimating gradient is given as:

$$\nabla_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{f}\left\{\frac{\mathrm{d}t(\boldsymbol{\phi}, \boldsymbol{\upsilon})}{\mathrm{d}\boldsymbol{\phi}}' \nabla_{\boldsymbol{\theta}}\left[\log p(\boldsymbol{y}, t(\boldsymbol{\phi}, \boldsymbol{\upsilon})) - \log \widetilde{p}_{\boldsymbol{\phi}}(t(\boldsymbol{\phi}, \boldsymbol{\upsilon}))\right]\right\},$$

which is computed with random samples generated from $f(\boldsymbol{\upsilon})$.

With the reparametrization approach, $\boldsymbol{\rho}^m_{k^m}$ and $\boldsymbol{\rho}^r_{k^r}$ can be represented as

$$\boldsymbol{\rho}^m_{k^m} = \boldsymbol{m}_{\rho^m,k^m} + \boldsymbol{G}_{\rho^m,k^m}\boldsymbol{v}_{\rho^m,k^m,1} + \boldsymbol{h}_{\rho^m,k^m} \circ \boldsymbol{v}_{\rho^m,k^m,2},$$

and

$$\boldsymbol{\rho}^r_{k^r} = \boldsymbol{m}_{\rho^r,k^r} + \boldsymbol{G}_{\rho^r,k^r}\boldsymbol{v}_{\rho^r,k^r,1} + \boldsymbol{h}_{\rho^r,k^r} \circ \boldsymbol{v}_{\rho^r,k^r,2},$$

where $\boldsymbol{v}_{\rho^m,k^m,1}$, $\boldsymbol{v}_{\rho^m,k^m,2}$, $\boldsymbol{v}_{\rho^r,k^r,1}$, and $\boldsymbol{v}_{\rho^r,k^r,2}$ are independent standard normal random vectors of dimensions r_{ρ^m,k^m}, R^m_w, r_{ρ^r,k^r}, R^r_w respectively. The gradient with respect to $\boldsymbol{m}_{\rho^m,k^m}$ is

$$\nabla_{\boldsymbol{m}_{\rho^m,k^m}}\mathcal{L} = \mathbb{E}_f\Bigg\{\nabla_{\boldsymbol{\rho}^m_{k^m}}\left[\log p(\boldsymbol{m}_{\rho^m,k^m} + \boldsymbol{G}_{\rho^m,k^m}\boldsymbol{v}_{\rho^m,k^m,1} + \boldsymbol{h}_{\rho^m,k^m} \circ \boldsymbol{v}_{\rho^m,k^m,2})\right] + \left(\boldsymbol{G}_{\rho^m,k^m}\boldsymbol{G}'_{\rho^m,k^m} + \boldsymbol{H}^2_{\rho^m,k^m}\right)^{-1}\left(\boldsymbol{G}_{\rho^m,k^m}\boldsymbol{v}_{\rho^m,k^m,1} + \boldsymbol{h}_{\rho^m,k^m} \circ \boldsymbol{v}_{\rho^m,k^m,2}\right)\Bigg\}.$$

The gradient with respect to $\boldsymbol{G}_{\rho^m,k^m}$ is

$$\nabla_{\boldsymbol{G}_{\rho^m,k^m}}\mathcal{L} = \mathbb{E}_f\Bigg\{\nabla_{\boldsymbol{\rho}^m_{k^m}}\left[\log p(\boldsymbol{m}_{\rho^m,k^m} + \boldsymbol{G}_{\rho^m,k^m}\boldsymbol{v}_{\rho^m,k^m,1} + \boldsymbol{h}_{\rho^m,k^m} \circ \boldsymbol{v}_{\rho^m,k^m,2})\right]\boldsymbol{v}'_{\rho^m,k^m,1} + \left(\boldsymbol{G}_{\rho^m,k^m}\boldsymbol{G}'_{\rho^m,k^m} + \boldsymbol{H}^2_{\rho^m,k^m}\right)^{-1}\left(\boldsymbol{G}_{\rho^m,k^m}\boldsymbol{v}_{\rho^m,k^m,1} + \boldsymbol{h}_{\rho^m,k^m} \circ \boldsymbol{v}_{\rho^m,k^m,2}\right)\boldsymbol{v}'_{\rho^m,k^m,1}\Bigg\}.$$

The gradient with respect to $\boldsymbol{h}_{\rho^m,k^m}$ is

$$\nabla_{\boldsymbol{G}_{\rho^m,k^m}}\mathcal{L} = \mathbb{E}_f\Bigg\{\mathrm{diag}\Big(\nabla_{\boldsymbol{\rho}^m_{k^m}}\left[\log p(\boldsymbol{m}_{\rho^m,k^m} + \boldsymbol{G}_{\rho^m,k^m}\boldsymbol{v}_{\rho^m,k^m,1} + \boldsymbol{h}_{\rho^m,k^m} \circ \boldsymbol{v}_{\rho^m,k^m,2})\right] \cdot \boldsymbol{v}'_{\rho^m,k^m,2} + \left(\boldsymbol{G}_{\rho^m,k^m}\boldsymbol{G}'_{\rho^m,k^m} + \boldsymbol{H}^2_{\rho^m,k^m}\right)^{-1}\left(\boldsymbol{G}_{\rho^m,k^m}\boldsymbol{v}_{\rho^m,k^m,1} + \boldsymbol{h}_{\rho^m,k^m} \circ \boldsymbol{v}_{\rho^m,k^m,2}\right)\boldsymbol{v}'_{\rho^m,k^m,2}\Big)\Bigg\}.$$

The expectations are computed numerically by generating samples $\boldsymbol{v}_{\rho^m,k^m,1}$, and $\boldsymbol{v}_{\rho^m,k^m,2}$ from respective standard normal distributions. Gradients with respect to $\boldsymbol{m}_{\rho^r,k^r}$, $\boldsymbol{G}_{\rho^r,k^r}$,

and $\boldsymbol{h}_{\rho^r,k^r}$ can be derived in the same fashion and computed numerically with samples of $\boldsymbol{v}_{\rho^r,k^r,1}$, and $\boldsymbol{v}_{\rho^r,k^r,2}$.

The updating of $\boldsymbol{\phi}_\beta$ is similar to that of $\boldsymbol{\phi}_\rho$. For the K^m sets of R_β^c- and R_β^p-dimensional emission model parameters $\{\boldsymbol{\beta}_{k^m}^c\}$ and $\{\boldsymbol{\beta}_{k^m}^p\}$, as well as the K^r sets of R_β^{pl}-dimensional emission model parameters $\{\boldsymbol{\beta}_{k^r}^{pl}\}$, we assign them with similar factorized multivariate normal variational posteriors as

$$\begin{aligned}
\prod_{k^m=1}^{K^m} \widetilde{p}(\boldsymbol{\beta}_{k^m}^c) &= \prod_{k^m=1}^{K^m} \mathcal{N}\left(\boldsymbol{\beta}_{k^m}^c; \boldsymbol{m}_{\beta^c,k^m}, \boldsymbol{G}_{\beta^c,k^m}\boldsymbol{G}'_{\beta^c,k^m} + \boldsymbol{H}^2_{\beta^c,k^m}\right), \\
\prod_{k^m=1}^{K^m} \widetilde{p}(\boldsymbol{\beta}_{k^m}^p) &= \prod_{k^m=1}^{K^m} \mathcal{N}\left(\boldsymbol{\beta}_{k^m}^p; \boldsymbol{m}_{\beta^p,k^m}, \boldsymbol{G}_{\beta^p,k^m}\boldsymbol{G}'_{\beta^p,k^m} + \boldsymbol{H}^2_{\beta^p,k^m}\right), \\
\prod_{k^r=1}^{K^r} \widetilde{p}(\boldsymbol{\beta}_{k^r}^{pl}) &= \prod_{k^r=1}^{K^r} \mathcal{N}\left(\boldsymbol{\beta}_{k^r}^{pl}; \boldsymbol{m}_{\beta^{pl},k^r}, \boldsymbol{G}_{\beta^{pl},k^r}\boldsymbol{G}'_{\beta^{pl},k^r} + \boldsymbol{H}^2_{\beta^{pl},k^r}\right),
\end{aligned}$$

where $\boldsymbol{m}_{\beta^c,k^m}$, $\boldsymbol{m}_{\beta^p,k^m}$, $\boldsymbol{m}_{\beta^{pl},k^r}$, $\boldsymbol{G}_{\beta^c,k^m}$, $\boldsymbol{G}_{\beta^p,k^m}$, $\boldsymbol{G}_{\beta^{pl},k^r}$, $\boldsymbol{H}_{\beta^c,k^m}$, $\boldsymbol{H}_{\beta^p,k^m}$, and $\boldsymbol{H}_{\beta^{pl},k^r}$ are defined similarly. Numbers of factors are denoted as r_{β^c,k^m}, r_{β^p,k^m}, and r_{β^{pl},k^r}. The variational parameters $\boldsymbol{\phi}_{\beta^c,k^m} = \{\boldsymbol{m}_{\beta^c,k^m}, \boldsymbol{G}_{\beta^c,k^m}, \boldsymbol{H}_{\beta^c,k^m}\}$, $\boldsymbol{\phi}_{\beta^p,k^m} = \{\boldsymbol{m}_{\beta^p,k^m}, \boldsymbol{G}_{\beta^p,k^m}, \boldsymbol{H}_{\beta^p,k^m}\}$, and $\boldsymbol{\phi}_{\beta^{pl},k^r} = \{\boldsymbol{m}_{\beta^{pl},k^r}, \boldsymbol{G}_{\beta^{pl},k^r}, \boldsymbol{H}_{\beta^{pl},k^r}\}$ can be updated similarly as above. For other model parameters and their corresponding variational parameters:

- $\boldsymbol{\mu}^c$: $\boldsymbol{\phi}_{\mu^c} = \{\boldsymbol{m}_{\mu^c}, \boldsymbol{G}_{\mu^c}, \boldsymbol{H}_{\mu^c}\}$; $\boldsymbol{\mu}^p$: $\boldsymbol{\phi}_{\mu^p} = \{\boldsymbol{m}_{\mu^p}, \boldsymbol{G}_{\mu^p}, \boldsymbol{H}_{\mu^p}\}$;
  $\boldsymbol{\mu}^{pl}$: $\boldsymbol{\phi}_{\mu^{pl}} = \{\boldsymbol{m}_{\mu^{pl}}, \boldsymbol{G}_{\mu^{pl}}, \boldsymbol{H}_{\mu^{pl}}\}$;
- $\boldsymbol{\xi}_i^c$: $\boldsymbol{\phi}_{\xi^c,i} = \{\boldsymbol{m}_{\xi^c,i}, \boldsymbol{G}_{\xi^c,i}, \boldsymbol{H}_{\xi^c,i}\}$; $\boldsymbol{\xi}_i^p$: $\boldsymbol{\phi}_{\xi^p,i} = \{\boldsymbol{m}_{\xi^p,i}, \boldsymbol{G}_{\xi^p,i}, \boldsymbol{H}_{\xi^p,i}\}$;
  $\boldsymbol{\xi}_i^{pl}$: $\boldsymbol{\phi}_{\xi^{pl},i} = \{\boldsymbol{m}_{\xi^{pl},i}, \boldsymbol{G}_{\xi^{pl},i}, \boldsymbol{H}_{\xi^{pl},i}\}$;
- $\boldsymbol{\zeta}_i^m$: $\boldsymbol{\phi}_{\zeta^m,i} = \{\boldsymbol{m}_{\zeta^m,i}, \boldsymbol{G}_{\zeta^m,i}, \boldsymbol{H}_{\zeta^m,i}\}$; $\boldsymbol{\zeta}_i^r$: $\boldsymbol{\phi}_{\zeta^r,i} = \{\boldsymbol{m}_{\zeta^r,i}, \boldsymbol{G}_{\zeta^r,i}, \boldsymbol{H}_{\zeta^r,i}\}$;
- $\boldsymbol{\eta}^m$: $\boldsymbol{\phi}_{\eta^m} = \{\boldsymbol{m}_{\eta^m}, \boldsymbol{G}_{\eta^m}, \boldsymbol{H}_{\eta^m}\}$; $\boldsymbol{\eta}^r$: $\boldsymbol{\phi}_{\eta^r} = \{\boldsymbol{m}_{\eta^r}, \boldsymbol{G}_{\eta^r}, \boldsymbol{H}_{\eta^r}\}$;
- $\boldsymbol{\pi}^m$: $\boldsymbol{\phi}_{\pi^m} = \{\boldsymbol{m}_{\pi^m}, \boldsymbol{G}_{\pi^m}, \boldsymbol{H}_{\pi^m}\}$; $\boldsymbol{\pi}^r$: $\boldsymbol{\phi}_{\pi^r} = \{\boldsymbol{m}_{\pi^r}, \boldsymbol{G}_{\pi^r}, \boldsymbol{H}_{\pi^r}\}$.

Here, for convenience, we model and estimate the vectorized version of $\boldsymbol{\zeta}_i^m$ such that $\boldsymbol{m}_{\zeta^m,i}$ is a $K^m \times (K^m - 1)$-dimensional vector, and $\boldsymbol{G}_{\zeta^m,i}$ and $\boldsymbol{H}_{\zeta^m,i}$ are $K^m(K^m-1) \times r_{\zeta^m}$- and $K^m(K^m-1) \times K^m(K^m-1)$-dimensional matrices respectively. Similar vectorization and dimensionality apply to variational parameters $\boldsymbol{m}_{\zeta^r,i}$, $\boldsymbol{G}_{\zeta^r,i}$, $\boldsymbol{H}_{\zeta^r,i}$, $\boldsymbol{m}_{\eta^m}$, $\boldsymbol{G}_{\eta^m}$, $\boldsymbol{H}_{\eta^m}$, $\boldsymbol{m}_{\eta^r}$, $\boldsymbol{G}_{\eta^r}$, and $\boldsymbol{H}_{\eta^r}$. Gradients with respect to these variational parameters can be derived and computed in the same fashion as those with respect to $\boldsymbol{\phi}_\rho$.

For a dataset with large number of individuals, N, we can speed up the computational efficiency via subsampling because of the independence among individuals. In specific, we randomly sample N_s out of the N individuals at each iteration and update parameters based on the subsample. Hoffman et al. (2013) labeled this subsampling-based variational Bayes estimation scheme the stochastic variational Bayes (SVB), which reduces the time complexity of the algorithm significantly from $O(NT)$ to $O(N_sT)$. The ELBO, $\mathcal{L}$, and the likelihood function, p, in the aforementioned gradient derivations are thus replaced by $\mathcal{L}_{N_s}$ and p_{N_s}, the ELBO and the likelihood function computed from the subsample of N_s sequences, respectively. The SVB method is guaranteed to converge to an local optimum; we terminate the algorithm after a sufficient number of iterations when convergence is attained (Blei et al., 2017; Ong et al., 2018; Hoffman et al., 2013).

The SVB Procedure

The SVB procedure consists of the following steps:

Step 1: Specify the number of iterations for the algorithm, *Iter*, and initialize all variational parameters.

The procedure iteratively performs the following steps until *Iter* is reached.

Step 2: Randomly sample N_s individuals from the total N without replacement.

Step 3: Update the posteriors for hidden states $\{\boldsymbol{Z}_{i,t_i}\}$ for $i \in \{N_s\}$ using current values

of variational parameters and the sampled N_s sequences. To address numerical underflow issues, the forward-backward probabilities are normalized for each time point t_i.

Step 4: Update shared (global) variational parameters, $\boldsymbol{\phi}_\mu = \{\boldsymbol{\phi}_{\mu^c}, \boldsymbol{\phi}_{\mu^p}, \boldsymbol{\phi}_{\mu^{pl}}\}$, $\boldsymbol{\phi}_\beta = \{\boldsymbol{\phi}_{\beta^c,k^m}, \boldsymbol{\phi}_{\beta^p,k^m}, \boldsymbol{\phi}_{\beta^{pl},k^r} | k^m \in [K^m], k^r \in [K^r]\}$, $\boldsymbol{\phi}_\eta = \{\boldsymbol{\phi}_{\eta^m}, \boldsymbol{\phi}_{\eta^r}\}$, $\boldsymbol{\phi}_\rho = \{\boldsymbol{\phi}_{\rho^m,k^m}, \boldsymbol{\phi}_{\rho^r,k^r} | k^m \in [K^m], k^r \in [K^r]\}$, and $\boldsymbol{\phi}_\pi = \{\boldsymbol{\phi}_{\pi^m}, \boldsymbol{\phi}_{\pi^r}\}$ via SGA. The adaptive learning rates are determined using the *Adam* optimizer.

Step 5: Update individual (local) variational parameters for $i \in \{N_s\}$, $\boldsymbol{\phi}_{\xi,i} = \{\boldsymbol{\phi}_{\xi^c,i}, \boldsymbol{\phi}_{\xi^p,i}, \boldsymbol{\phi}_{\xi^{pl},i}\}$ and $\boldsymbol{\phi}_{\zeta,i} = \{\boldsymbol{\phi}_{\zeta^m,i}, \boldsymbol{\phi}_{\zeta^r,i}\}$ via SGA. The adaptive learning rates are determined using the *Adam* optimizer.

Step 6: Compute the ELBO value $\mathcal{L}_{N_s}$ using current values of variational parameters.

4.6.2 The *Adam* Optimizer

The *Adam* optimizer generates adaptive learning rates for SGA. Let $\phi_j^{(\tau)}$, $\psi_{\tau,j}$, and $\nabla l_j^{(\tau)}$ denote the j-th element of $\boldsymbol{\phi}^{(\tau)}$, $\boldsymbol{\psi}_\tau$, and $\nabla \mathcal{L}(\boldsymbol{\phi}^{(\tau)})$, respectively. The learning rate $\psi_{\tau,j}$ at the τth-iteration is computed as follows:

$$\psi_{\tau,j} = o_j \frac{\widehat{\chi}_{1,\tau,j}}{(\sqrt{\widehat{\chi}_{2,\tau,j}} + \epsilon_0)\nabla l_j^{(\tau)}},$$

where

$$\begin{aligned}
\widehat{\chi}_{1,\tau,j} &= \frac{\chi_{1,\tau,j}}{1-\epsilon_1^\tau}, \\
\widehat{\chi}_{2,\tau,j} &= \frac{\chi_{2,\tau,j}}{1-\epsilon_2^\tau}, \\
\chi_{1,\tau,j} &= \epsilon_1 \chi_{1,\tau-1,j} + (1-\epsilon_1)\nabla l_j^{(\tau)}, \\
\chi_{2,\tau,j} &= \epsilon_2 \chi_{2,\tau-1,j} + (1-\epsilon_2)\nabla l_j^{(\tau)^2},
\end{aligned}$$

and ϵ_0 ϵ_1, ϵ_2, and o_j are predefined hyperparameters. For this approach, we set starting values as $\chi_{1,0,j} = \chi_{2,0,j} = 0$ and hyperparameters as $\epsilon_0 = 10^{-8}$, $\epsilon_1 = 0.9$, $\epsilon_2 = 0.999$, $o_j = o_{1j} o_{2j}^\tau$, $o_{1j}, \in (10^{-4}, 10^{-2})$, and $o_{2j} \in (0.999, 0.9999)$, according to the suggestions in the literature (Shi et al., 2019; Kingma and Ba, 2015). For other details of the *Adam* optimizer, we refer readers to Kingma and Ba (2015).

4.6.3 The Forward-Backward Algorithm

The forward probabilities $a_{i,t_i}(k)$ and backward probabilities $b_{i,t_i}(k)$ are defined as: for $k \in [K]$,

$$\begin{aligned} a_{i,t_i}(k) &= p(\boldsymbol{Y}_{i,1:t_i}, \boldsymbol{Z}_{i,t_i} = k | \boldsymbol{X}_{i,1:t_i}, \boldsymbol{W}_{i,1:t_i}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\pi}), \\ b_{i,t_i}(k) &= p(\boldsymbol{Y}_{i,(t_i+1):T_i} | \boldsymbol{Z}_{i,t_i} = k, \boldsymbol{X}_{i,(t_i+1):T_i}, \boldsymbol{W}_{i,(t_i+1):T_i}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\pi}), \end{aligned} \tag{4.9}$$

where $K = K^m \times K^r$, and $\boldsymbol{Z}_{i,t_i} = k$ if and only if $Z^m_{i,t_t} = k^m$ for $k^m \in [K^m]$ and $Z^r_{i,t_i} = k^r$ for $k^r \in [K^r]$. Here we establish a one-to-one correspondence between (k^m, k^r) and k by letting $k = k^m \times (K^r - 1) + k^r$. The forward and backward probabilities can be computed iteratively using the following formulas:

$$\begin{aligned} \boldsymbol{a}'_{i,1} &= \boldsymbol{\pi}'_i \boldsymbol{D}_{i,1}, \\ \boldsymbol{a}'_{i,t_i+1} &= \boldsymbol{a}'_{i,t_i} \boldsymbol{E}_{i,t_i+1} \boldsymbol{D}_{i,t_i+1}, \\ \boldsymbol{b}_{i,T_i} &= \boldsymbol{1}, \\ \boldsymbol{b}_{i,t_i} &= \boldsymbol{E}_{i,t_i+1} \boldsymbol{D}_{i,t_i+1} \boldsymbol{b}_{i,t_i+1}, \end{aligned}$$

where $\boldsymbol{a}_{i,t_i} = (a_{i,t_i}(1), ..., a_{i,t_i}(K))'$, $\boldsymbol{b}_{i,t_i} = (b_{i,t_i}(1), ..., b_{i,t_i}(K))'$, $\boldsymbol{1}$ is a K-dimensional vector with all elements being 1, and $\boldsymbol{D}_{i,t_i}$ and $\boldsymbol{E}_{i,t_i}$ are diagonal and square matrices, respectively, as defined below. To prevent numerical underflow issues, $\boldsymbol{a}_{i,t_i}$ and $\boldsymbol{b}_{i,t_i}$ are normalized at each iteration. The diagonal elements of $\boldsymbol{D}_{i,t_i}$ are given as follows: for $k = 1, \ldots, K$,

$$d_{i,t_i,kk} = \exp\left\{ \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\mu}(\boldsymbol{\mu}) \widetilde{p}_{\boldsymbol{\phi}_\xi}(\boldsymbol{\xi}) \widetilde{p}_{\boldsymbol{\phi}_\beta}(\boldsymbol{\beta})} \left[\log p(\boldsymbol{Y}_{i,t_i} | \boldsymbol{Z}_{i,t_i} = k, \boldsymbol{X}_{i,t_i}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\beta}) \right] \right\}.$$

Elements in $\boldsymbol{E}_{i,t_i}$ are given as follows: for $k_1, k_2 = 1, \ldots, K$

$$e_{i,t_i,k_1k_2} = \exp\left\{ \mathbb{E}_{\widetilde{p}_{\boldsymbol{\phi}_\zeta}(\zeta) \widetilde{p}_{\boldsymbol{\phi}_\eta}(\boldsymbol{\eta}) \widetilde{p}_{\boldsymbol{\phi}_\rho}(\boldsymbol{\rho})} \left[\log p(\boldsymbol{Z}_{i,t_i} = k_2 | \boldsymbol{Z}_{i,t_i-1} = k_1, \boldsymbol{W}_{i,t_i}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\rho}) \right] \right\}.$$

The above two expectations are computed numerically using Monte Carlo samples from relevant variational posteriors. Note that $\boldsymbol{D}_{i,t_i}$ and $\boldsymbol{E}_{i,t_i}$ can be viewed as variational estimates of emission probabilities and transition probability matrices respectively. Marginal posteriors $\widetilde{p}(\boldsymbol{Z}_{i,t_i})$ and $\widetilde{p}(\boldsymbol{Z}_{i,t_i-1}, \boldsymbol{Z}_{i,t_i})$ can be computed as follows:

$$\begin{aligned}\widetilde{p}(\boldsymbol{Z}_{i,t_i} = k) &\propto a_{i,t_i}(k) b_{i,t_i}(k) \\ &= \frac{a_{i,t_i}(k) b_{i,t_i}(k)}{\sum_j a_{i,t_i}(j) b_{i,t_i}(j)},\end{aligned}$$

and

$$\begin{aligned}\widetilde{p}(\boldsymbol{Z}_{i,t_i-1} = k_1, \boldsymbol{Z}_{i,t_i} = k_2) &\propto a_{i,t_i-1}(k_1) e_{i,t_i,k_1k_2} d_{i,t_i,k_2k_2} b_{i,t_i}(k_2) \\ &= \frac{a_{i,t_i-1}(k_1) e_{i,t_i,k_1k_2} d_{i,t_i,k_2k_2} b_{i,t_i}(k_2)}{\sum_{j_1} \sum_{j_2} a_{i,t_i-1}(j_1) e_{i,t_i,j_1j_2} d_{i,t_i,j_2j_2} b_{i,t_i}(j_2)}.\end{aligned}$$

4.6.4 Model Selection Tables

Table 4.12: CNHMM model selection results for selecting numbers of states; metrics include the DIC, in-sample mean squared error (ISMSE) and prediction accuracy (ISPA), and out-of-sample mean squared error (OSMSE) and prediction accuracy (OSPA).

DIC				
K^m	1	2	3	4
K^r				
1	98093.10	62306.71	68109.99	66207.12
2	56783.00	62127.00	59664.64	62052.25
3	57368.12	58457.01	**54127.10**	55970.27
4	57609.51	54562.28	54411.67	57205.30
ISMSE				
K^m	1	2	3	4
K^r				
1	0.165	0.066	0.067	0.060
2	0.062	0.062	0.067	0.062
3	0.061	0.065	**0.059**	0.072
4	0.060	0.063	0.066	0.065
ISPA				
K^m	1	2	3	4
K^r				
1	0.848	0.978	0.971	0.975
2	0.978	0.980	0.967	0.981
3	0.978	0.980	**0.978**	0.979
4	0.977	0.978	0.978	0.978
OSMSE				
K^m	1	2	3	4
K^r				
1	0.357	0.193	0.234	0.206
2	0.208	0.215	0.216	0.188
3	0.199	0.201	**0.169**	0.172
4	0.212	0.208	0.230	0.191
OSPA				
K^m	1	2	3	4
K^r				
1	0.641	0.921	0.885	0.910
2	0.877	0.923	0.923	0.921
3	0.925	0.926	**0.933**	0.928
4	0.912	0.923	0.910	0.917

Table 4.13: FHMM model selection results for selecting numbers of states; metrics include the DIC, in-sample mean squared error (ISMSE) and prediction accuracy (ISPA), and out-of-sample mean squared error (OSMSE) and prediction accuracy (OSPA).

DIC				
K^m	1	2	3	4
K^r				
1	98093.10	62306.71	68109.99	66207.12
2	56783.00	57206.20	58604.66	**56982.58**
3	57368.12	65591.20	65854.34	69868.21
4	57609.51	59951.82	71976.29	56805.46
ISMSE				
K^m	1	2	3	4
K^r				
1	0.165	0.066	0.067	0.060
2	0.062	0.079	0.072	**0.061**
3	0.061	0.087	0.071	0.074
4	0.060	0.059	0.073	0.060
ISPA				
K^m	1	2	3	4
K^r				
1	0.848	0.978	0.971	0.975
2	0.978	0.978	0.978	**0.979**
3	0.978	0.970	0.976	0.974
4	0.977	0.978	0.972	0.979
OSMSE				
K^m	1	2	3	4
K^r				
1	0.357	0.193	0.234	0.206
2	0.208	0.238	0.222	**0.177**
3	0.199	0.270	0.244	0.254
4	0.212	0.188	0.251	0.175
OSPA				
K^m	1	2	3	4
K^r				
1	0.641	0.921	0.885	0.910
2	0.877	0.918	0.913	**0.930**
3	0.925	0.851	0.885	0.878
4	0.912	0.918	0.869	0.923

Table 4.14: NHMM model selection results for selecting numbers of states; metrics include the DIC, in-sample mean squared error (ISMSE) and prediction accuracy (ISPA), and out-of-sample mean squared error (OSMSE) and prediction accuracy (OSPA).

K	DIC	ISMSE	ISPA	OSMSE	OSPA
2	62354.04	0.063	0.965	0.191	0.921
3	**55618.99**	**0.060**	**0.977**	**0.179**	**0.923**
4	61522.02	0.063	0.978	0.198	0.926
5	61644.24	0.060	0.978	0.190	0.933
6	67494.01	0.060	0.978	0.184	0.936

Bibliography

Abarbanel, H. D., Brown, R., and Kennel, M. B. (1992). Local lyapunov exponents computed from observed data. *Journal of Nonlinear Science*, 2(3):343–365.

Aicher, C., Ma, Y.-A., Foti, N. J., and Fox, E. B. (2019). Stochastic gradient mcmc for state space models. *SIAM Journal on Mathematics of Data Science*, 1(3):555–587.

Alie, D., Mahoor, M. H., Mattson, W. I., Anderson, D. R., and Messinger, D. S. (2011). Analysis of eye gaze pattern of infants at risk of autism spectrum disorder using markov models. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 282–287.

Altman, R. M. (2007). Mixed hidden markov models: an extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. Arlington: American Psychiatric Publishing, 5th edition.

Andrews, M., Luo, X., Fang, Z., and Ghose, A. (2016). Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science*, 35(2):218–233.

Ansari, A., Li, Y., and Zhang, J. Z. (2018). Probabilistic topic model for hybrid recommender systems: A stochastic variational bayesian approach. *Marketing Science*, 37(6):987–1008.

Ansari, A., Mela, C. F., and Neslin, S. A. (2008). Customer channel migration. *Journal of Marketing Research*, 45(1):60–76.

Ansari, A., Montoya, R., and Netzer, O. (2012). Dynamic learning in behavioral games: A hidden markov mixture of experts approach. *Quantitative Marketing and Economics*, 10(4):475–503.

Arnold, L. (1998). Random dynamical systems. In *Springer Monographs in Mathematics*. Springer-Verlag Berlin Heidelberg.

Ascarza, E., Netzer, O., and Hardie, B. G. S. (2018). Some customers would rather leave without saying goodbye. *Marketing Science*, 37(1):54–77.

Avery, J., Steenburgh, T. J., Deighton, J., and Caravella, M. (2012). Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time. *Journal of Marketing*, 76(3):96–111.

Barendregt, J. J. and Veerman, J. L. (2010). Categorical versus continuous risk factors and the calculation of potential impact fractions. *Journal of Epidemiology & Community Health*, 64(3):209–212.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Boulding, W., Staelin, R., Ehret, M., and Johnston, W. J. (2005). A customer relationship management roadmap: What is known, potential pitfalls, and where to go. *Journal of Marketing*, 69(4):155–166.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC press.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Science & Business Media, USA.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error In Nonlinear Models: A Modern Perspective*. CRC press.

Chawarska, K., Macari, S., and Shic, F. (2012). Context modulates attention to social scenes in toddlers with autism. *Journal of Child Psychology and Psychiatry*, 53(8):903–913.

Collet, P. and Leonardi, F. (2014). Loss of memory of hidden markov models and lyapunov exponents. *The Annals of Applied Probability*, 24(1):422–446.

Copeland, K. T., Checkoway, H., McMichael, A. J., and Holbrook, R. H. (1977). Bias due to missclassification in the estimation of relative risk. *American Journal of Epidemiology*, 105(5):488–495.

Dalen, I., Buonaccorsi, J. P., Sexton, J. A., Laake, P., and Thoresen, M. (2009). Correction for misclassification of a categorized exposure in binary regression using replication data. *Statistics in Medicine*, 28(27):3386–3410.

Deubner, D. C., Wilkinson, W. E., Helms, M. J., Tyroler, H. A., and Hames, C. G. (1980). Logistic model estimation of death attributable to risk factors for cardiovascular disease in evans county, georgia. *American Journal of Epidemiology*, 112(1):135–143.

Dowling, G. R. and Uncles, M. (1997). Do customer loyalty programs really work? *Sloan management review*, 38:71–82.

Drescher, K. and Becher, H. (1997). Estimating the Generalized Impact Fraction from Case-Control Data. *Biometrics*, 53(3):1170.

Du, R. Y. and Kamakura, W. A. (2006). Household life cycles and lifestyles in the united states. *Journal of Marketing Research*, 43(1):121–132.

Dwyer, F. R., Schurr, P. H., and Oh, S. (1987). Developing buyer-seller relationships. *Journal of Marketing*, 51(2):11–27.

Eide, G. E. and Heuch, I. (2001). Attributable fractions: fundamental concepts and their visualization. *Statistical Methods in Medical Research*, 10(3):159–193. PMID: 11446147.

Ertekin, N., Shulman, J. D., and Chen, H. A. (2019). On the profitability of stacked discounts: Identifying revenue and cost effects of discount framing. *Marketing Science*, 38(2):317–342.

Fader, P. S., Hardie, B. G., and Lee, K. L. (2005). Rfm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4):415–430.

Feinberg, F. M., Salisbury, L. C., and Ying, Y. (2016). When random assignment is not enough: Accounting for item selectivity in experimental research. *Marketing Science*, 35(6):976–994.

Feng, X., Li, Y., Lin, X., and Ning, Y. (2020). Mobile targeting in industrial marketing: Connecting with the right businesses. *Industrial Marketing Management*, 86:65–76.

Fong, N. M., Fang, Z., and Luo, X. (2015). Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research*, 52(5):726–735.

Foti, N., Xu, J., Laird, D., and Fox, E. (2014). Stochastic variational inference for hidden markov models. In *Advances in Neural Information Processing Systems*, pages 3599–3607.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Giovannucci, E., Rimm, E. B., Ascherio, A., Stampfer, M. J., Colditz, G. A., and Willett, W. C. (1995). Alcohol, Low-Methionine-Low-Folate Diets, and Risk of Colon Cancer in Men. *JNCI: Journal of the National Cancer Institute*, 87(4):265–273.

Giovannucci, E., Rimm, E. B., Stampfer, M. J., Colditz, G. A., Ascherio, A., and Willett, W. C. (1994). Intake of fat, meat, and fiber in relation to risk of colon cancer in men. *Cancer Research*, 54(9):2390–2397.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Goldberg, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association*, 70(351a):561–567.

Graham, P. (2000). Bayesian inference for a generalized population attributable fraction: the impact of early vitamin a levels on chronic lung disease in very low birthweight infants. *Statistics in Medicine*, 19(7):937–956.

Grant, A. W. and Schlesinger, L. A. (1995). Realize your customers' full profit potential. *Harvard Business Review*, 73(5):59–72.

Haenlein, M., Kaplan, A. M., and Schoder, D. (2006). Valuing the real option of abandoning unprofitable customers when calculating customer lifetime value. *Journal of Marketing*, 70(3):5–20.

Heaps, S. E., Boys, R. J., and Farrow, M. (2015). Bayesian modelling of rainfall data by using non-homogeneous hidden markov models and latent gaussian variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):543–568.

Hoff, P. D. (2009). *A First Course In Bayesian Statistical Methods*, volume 580. Springer.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Holsclaw, T., Greene, A. M., Robertson, A. W., and Smyth, P. (2017). Bayesian nonhomogeneous markov models via pólya-gamma data augmentation with applications to rainfall modeling. *Annals of Applied Statistics*, 11(1):393–426.

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., and Verhoef, P. C. (2017). No future without the past? predicting churn in the face of customer privacy. *International Journal of Research in Marketing*, 34(1):154–172.

Hsieh, C.-C. and Walter, S. D. (1988). The effect of non-differential exposure misclassification on estimates of the attributable and prevented fraction. *Statistics in Medicine*, 7(10):1073–1085.

Hughes, A. M. (2000). *Strategic database marketing: the masterplan for starting and managing a profitable, customer-based marketing program*, volume 12. McGraw-Hill New York, NY.

Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1):15–30.

Ip, E., Zhang, Q., Rejeski, J., Harris, T., and Kritchevsky, S. (2013). Partially ordered mixed hidden markov model for the disablement process of older adults. *Journal of the American Statistical Association*, 108(502):370–384.

Johnson, C. Y., Flanders, W. D., Strickland, M. J., Honein, M. A., and Howards, P. P. (2014). Potential sensitivity of bias analysis results to incorrect assumptions of nondifferential or differential binary exposure misclassification. *Epidemiology*, 25(6):902–909.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Kang, K., Cai, J., Song, X., and Zhu, H. (2019). Bayesian hidden markov models for delineating the pathology of alzheimer's disease. *Statistical Methods in Medical Research*, 28(7):2112–2124.

Kani, A., DeSarbo, W. S., and Fong, D. K. H. (2018). A factorial hidden markov model for the analysis of temporal change in choice models. *Customer Needs and Solutions*, 5(3):162–177.

Kappe, E., Stadler Blank, A., and DeSarbo, W. S. (2018). A random coefficients mixture hidden markov model for marketing research. *International Journal of Research in Marketing*, 35(3):415–431.

Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., and Scassellati, B. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *Journal of Autism and Developmental Disorders*, 43(5):1038–1049.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Klin, A., Jones, W., Schultz, R., Volkmar, F., and Cohen, D. (2002). Defining and quantifying the social phenotype in autism. *American Journal of Psychiatry*, 159(6):895–908. PMID: 12042174.

Kotler, P. and Keller, K. L. (2016). *Marketing Management*. Pearson Italia Spa.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Kumar, V. and Reinartz, W. (2012). *Strategic Customer Relationship Management Today*, pages 3–20. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kumar, V., Venkatesan, R., Bohling, T., and Beckmann, D. (2008). Practice prize report—the power of clv: Managing customer lifetime value at ibm. *Marketing Science*, 27(4):585–599.

Le Gland, F. and Mevel, L. (2000a). Basic properties of the projective product with application to products of column-allowable nonnegative matrices. *Mathematics of Control, Signals and Systems*, 13(1):41–62.

Le Gland, F. and Mevel, L. (2000b). Exponential forgetting and geometric ergodicity in hidden markov models. *Mathematics of Control, Signals and Systems*, 13(1):63–93.

Lee, C., Ofek, E., and Steenburgh, T. J. (2018). Personal and social usage: The origins of active customers and ways to keep them engaged. *Management Science*, 64(6):2473–2495.

Lewis, M. (2004). The influence of loyalty programs and short-term promotions on customer retention. *Journal of Marketing Research*, 41(3):281–292.

Lewis, M. (2006). Customer acquisition promotions and customer asset value. *Journal of Marketing Research*, 43(2):195–203.

Liechty, J., Pieters, R., and Wedel, M. (2003). Global and local covert visual attention: Evidence from a bayesian hidden markov model. *Psychometrika*, 68(4):519–541.

Liu, Y. (2007). The long-term impact of loyalty programs on consumer purchase behavior and loyalty. *Journal of Marketing*, 71(4):19–35.

Luo, X., Andrews, M., Fang, Z., and Phang, C. W. (2014). Mobile targeting. *Management Science*, 60(7):1738–1756.

Ma, L., Sun, B., and Kekre, S. (2015a). The squeaky wheel gets the grease - an empirical analysis of customer voice and firm intervention on twitter. *Marketing Science*, 34(5):627–645.

Ma, L., Sun, B., and Kekre, S. (2015b). The squeaky wheel gets the grease—an empirical analysis of customer voice and firm intervention on twitter. *Marketing Science*, 34(5):627–645.

Ma, Y.-A., Foti, N. J., and Fox, E. B. (2017). Stochastic gradient mcmc methods for hidden markov models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2265–2274. JMLR.org.

Mavadati, S. M., Feng, H., Gutierrez, A., and Mahoor, M. H. (2014). Comparing the gaze responses of children with autism and typically developed individuals in human-robot interaction. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 1128–1133.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Meligkotsidou, L. and Dellaportas, P. (2011). Forecasting with non-homogeneous hidden markov models. *Statistics and Computing*, 21(3):439–449.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.

Montgomery, A. L., Li, S., Srinivasan, K., and Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4):579–595.

Montoya, R., Netzer, O., and Jedidi, K. (2010). Dynamic allocation of pharmaceutical detailing and sampling for long-term profitability. *Marketing Science*, 29(5):909–924.

Moon, S., Kamakura, W. A., and Ledolter, J. (2007). Estimating promotion response when competitive promotions are unobservable. *Journal of Marketing Research*, 44(3):503–515.

Morgenstern, H. and Bursic, E. S. (1982). A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *Journal of Community Health*, 7(4):292–309.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Neslin, S. A., Taylor, G. A., Grantham, K. D., and McNeil, K. R. (2013). Overcoming the "recency trap" in customer relationship management. *Journal of the Academy of Marketing Science*, 41(3):320–337.

Netzer, O., Lattin, J. M., and Srinivasan, V. (2008). A hidden markov model of customer relationship dynamics. *Marketing Science*, 27(2):185–204.

Ong, V. M.-H., Nott, D. J., and Smith, M. S. (2018). Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, 27(3):465–478.

Padilla, N., Montoya, R., and Netzer, O. (2020). Heterogeneity in hmms: Allowing for heterogeneity in the number of states. *Working Paper*.

Parasuraman, A. (1997). Reflections on gaining competitive advantage through customer value. *Journal of the Academy of marketing Science*, 25(2):154–161.

Petersen, J. A. and Kumar, V. (2015). Perceived risk, product returns, and optimal resource allocation: Evidence from a field experiment. *Journal of Marketing Research*, 52(2):268–285.

Pirikahu, S., Jones, G., Hazelton, M. L., and Heuer, C. (2016). Bayesian methods of confidence interval construction for the population attributable risk from cross-sectional studies. *Statistics in Medicine*, 35(18):3117–3130.

Platz, E. A., Willett, W. C., Colditz, G. A., Rimm, E. B., Spiegelman, D., and Giovannucci, E. (2000). Proportion of Colon Cancer Risk That Might Be Preventable in a Cohort of Middle-Aged US Men. *Cancer Causes and Control 11:*, 11(7):579–588.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

Reinartz, W. J. and Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, 64(4):17–35.

Reinartz, W. J. and Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, 67(1):77–99.

Rhee, S. and McIntyre, S. (2008). Including the effects of prior and recent contact effort in a customer scoring model for database marketing. *Journal of the Academy of Marketing Science*, 36(4):538–551.

Rimm, E., Giovannucci, E., Willett, W., Colditz, G., Ascherio, A., Rosner, B., and Stampfer, M. (1991). Prospective study of alcohol consumption and risk of coronary disease in men. *The Lancet*, 338(8765):464 – 468. Originally published as Volume 2, Issue 8765.

Rimm, E. B., Giovannucci, E. L., Stampfer, M. J., Colditz, G. A., Litin, L. B., and Willett, W. C. (1992). Reproducibility and Validity of an Expanded Self-Administered Semiquantitative Food Frequency Questionnaire among Male Health Professionals. *American Journal of Epidemiology*, 135(10):1114–1126.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8(9):1051–1069.

Rust, R. T. and Verhoef, P. C. (2005). Optimizing the marketing interventions mix in intermediate-term crm. *Marketing Science*, 24(3):477–489.

Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P., and Shic, F. (2018). Improving social skills in children with asd using a long-term, in-home social robot. *Science Robotics*, 3(21).

Schmittlein, D. C. and Peterson, R. A. (1994). Customer base analysis: An industrial purchase process application. *Marketing Science*, 13(1):41–67.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2014). Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Science*, 33(2):188–205.

Sherlock, C., Xifara, T., Telfer, S., and Begon, M. (2013). A coupled hidden markov model for disease interactions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(4):609–627.

Shi, L., Onofrey, J. A., Revilla, E. M., Toyonaga, T., Menard, D., Ankrah, J., Carson, R. E., Liu, C., and Lu, Y. (2019). A novel loss function incorporating imaging acquisition physics for pet attenuation map generation using deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 723–731. Springer.

Shic, F., Wang, Q., Macari, S. L., and Chawarska, K. (2019). The role of limited salience of speech in selective attention to faces in toddlers with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, 0(0).

Sinha, S., Mallick, B. K., Kipnis, V., and Carroll, R. J. (2010). Semiparametric bayesian analysis of nutritional epidemiology data in the presence of measurement error. *Biometrics*, 66(2):444–454.

Song, X., Xia, Y., and Zhu, H. (2017). Hidden markov latent variable models with multivariate longitudinal data. *Biometrics*, 73(1):313–323.

Spezia, L. (2006). Bayesian analysis of non-homogeneous hidden markov models. *Journal of Statistical Computation and Simulation*, 76(8):713–725.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Spiegelman, D., Hertzmark, E., and Wand, H. C. (2007). Point and interval estimates of partial population attributable risks in cohort studies: Examples and software. *Cancer Causes and Control*, 18(5):571–579.

Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95(449):51–61.

Touloupou, P., Finkenstädt, B., Besser, T. E., French, N. P., and Spencer, S. E. F. (2020a). Bayesian inference for multistrain epidemics with application to ESCHERICHIA COLI O157:H7 in feedlot cattle. *The Annals of Applied Statistics*, 14(4):1925–1944.

Touloupou, P., Finkenstädt, B., and Spencer, S. E. F. (2020b). Scalable bayesian inference for coupled hidden markov and semi-markov models. *Journal of Computational and Graphical Statistics*, 29(2):238–249.

Venkatesan, R. and Farris, P. W. (2012). Measuring and managing returns from retailer-customized coupon campaigns. *Journal of Marketing*, 76(1):76–94.

Venkatesan, R., Kumar, V., and Bohling, T. (2007). Optimal customer relationship management using bayesian decision theory: An application for customer selection. *Journal of Marketing Research*, 44(4):579–594.

Walter, S. D. (1980). Prevention for multifactorial diseases. *American Journal of Epidemiology*, 112(3):409–416.

Wang, Q., Campbell, D. J., Macari, S. L., Chawarska, K., and Shic, F. (2018). Operationalizing atypical gaze in toddlers with autism spectrum disorders: a cohesion-based approach. *Molecular Autism*, 9(1):25.

Wang, Q., Wall, C. A., Barney, E. C., Bradshaw, J. L., Macari, S. L., Chawarska, K., and Shic, F. (2019). Promoting social attention in 3-year-olds with asd through gaze-contingent eye tracking. *Autism Research*, 0(0).

Wang, Y., Lewis, M., Cryder, C., and Sprigg, J. (2016). Enduring effects of goal achievement and failure within customer loyalty programs: A large-scale field experiment. *Marketing Science*, 35(4):565–575.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Wong, B. H., Peskoe, S. B., and Spiegelman, D. (2018). The effect of risk factor misclassification on the partial population attributable risk. *Statistics in Medicine*, 37(8):1259–1275.

Wong, B. H. W., Lee, J., Spiegelman, D., and Wang, M. (2020). Estimation and inference for the population attributable risk in the presence of misclassification. *Biostatistics*. kxz067.

Wübben, M. and v. Wangenheim, F. (2008). Instant customer base analysis: Managerial heuristics often "get it right". *Journal of Marketing*, 72(3):82–93.

Ye, X. (2018). *Stochastic dynamics: Markov chains, random transformations and applications*. PhD thesis, University of Washington.

Yi, G. Y., Ma, Y., Spiegelman, D., and Carroll, R. J. (2015). Functional and structural methods with mixed measurement error and misclassification in covariates. *Journal of the American Statistical Association*, 110(510):681–696. PMID: 26190876.

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *ArXiv e-prints*.

Zhang, J. and Yu, K. F. (1998). What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA*, 280(19):1690–1691.

Zhang, J. Z. and Chang, C.-W. (2021). Consumer dynamics: theories, methods, and emerging directions. *Journal of the Academy of Marketing Science*, 49:166–196.

Zhang, J. Z., Netzer, O., and Ansari, A. (2014). Dynamic targeted pricing in b2b relationships. *Marketing Science*, 33(3):317–337.

Zhang, X. A., Kumar, V., and Cosguner, K. (2017). Dynamically managing a profitable email marketing program. *Journal of Marketing Research*, 54(6):851–866.

Zhang, Y., Bradlow, E. T., and Small, D. S. (2015). Predicting customer value using clumpiness: From rfm to rfmc. *Marketing Science*, 34(2):195–208.

Zou, G. (2004). A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology*, 159(7):702–706.