DISSERTATION

BAYESIAN METHODS FOR ENVIRONMENTAL EXPOSURES: MIXTURES AND MISSING DATA

Submitted by

Lauren Hoskovec

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2022

Doctoral Committee:

Advisor: Ander Wilson

Jennifer Hoeting
Dan Cooley
Sheryl Magzamen

ABSTRACT

BAYESIAN METHODS FOR ENVIRONMENTAL EXPOSURES:

MIXTURES AND MISSING DATA

Air pollution exposure has been linked to increased morbidity and mortality. Estimating the association between air pollution exposure and health outcomes is complicated by simultaneous exposure to multiple pollutants, referred to as a multipollutant mixture. In a multipollutant mixture, exposures may have both independent and interactive effects on health. In addition, observational studies of air pollution exposure often involve missing data. In this dissertation, we address challenges related to model choice and missing data when studying exposure to a mixture of environmental pollutants. First, we conduct a formal simulation study of recently developed methods for estimating the association between a health outcome and exposure to a multipollutant mixture. We evaluate methods on their performance in estimating the exposure-response function, identifying mixture components associated with the outcome, and identifying interaction effects. Other studies have reviewed the literature or compared performance on a single data set; however, none have formally compared such a broad range of new methods in a simulation study. Second, we propose a statistical method to analyze multiple asynchronous multivariate time series with missing data for use in personal exposure assessments. We develop an infinite hidden Markov model for multiple time series to impute missing data and identify shared time-activity patterns in exposures. We estimate hidden states that represent latent environments presenting a unique distribution of a mixture of environmental exposures. Through our multiple imputation algorithm, we impute missing exposure data conditional on the hidden states. Finally, we conduct an individual-level study of the association between long-term exposure to air pollution and COVID-19 severity in a Denver, Colorado, USA cohort. We develop a Bayesian multinomial logistic regression model for data with partially missing categorical outcomes. Our model uses Polya-gamma data augmentation, and we propose a visu-

alization approach for inference on the odds ratio. We conduct one of the first individual-level studies of air pollution exposure and COVID-19 health outcomes using detailed clinical data and individual-level air pollution exposure data.

ACKNOWLEDGEMENTS

I recognize a number of people who have supported me on my educational journey. First, my family has been an infinite source of encouragement. I share my success and accomplishments with my husband, Luke Hoskovec, and am forever thankful for his unconditional love. I thank my parents, John and JoEtta Heck, for their love and support, and my sister, Rachel Heck, who is my inspiration.

My advisor, Dr. Ander Wilson, deserves utmost recognition. From the beginning of my journey, his teaching and mentoring have strengthened my skills a researcher, programmer, and writer. I am particularly grateful for his compelling encouragement, and for helping me build the confidence I needed to reach this milestone.

I recognize my committee members for their service: Thanks to Dr. Sheryl Magzamen for her fresh perspective from the environmental epidemiology field and for her contributions to my first and third projects, to Dr. Dan Cooley for his exceptional instruction throughout my graduate school journey, and to Dr. Jennifer Hoeting for her technical expertise and inspiring career and life advice.

I would like to thank Dr. Matthew Koslovsky for his input on my second project. His instruction and advice greatly improved my writing and programming skills. Thanks to Dr. Sheena Martenies for her collaborations and for helping me gain experience as an applied statistician. I also thank Dr. Mevin Hooten and Dr. John Tipton for their mentoring at the beginning of my Statistics education, and for encouraging me to pursue ecological and environmental statistics. I thank my high school math teacher Mr. Eric Gutjahr for ensuring I was sufficiently challenged.

Friends I have had and made along the way brought much joy to this academic achievement. Thanks to Katie Zagnoli for her sense of humor through challenging coursework, and to TK Santos for sharing life's adventures and constantly believing in me.

DEDICATION

I would like to dedicate this dissertation to my daughter Jocelyn, who brings endless joy to my life and gives meaning to all that I do.

TABLE OF CONTENTS

Chapter 1

Introduction

Exposure to air pollution is associated with increased morbidity and mortality, and presents a major global environmental health risk (Di et al., 2017a,b; Dockery et al., 1993; Global Burden of Diseases 2019 Risk Factors Collaborators, 2020; Health Effects Institute, 2018). Studies focusing on air pollution include exposure assessments and health effects estimation. Exposure assessments are used to estimate an individual's exposure to pollutants over a period of time (Finazzi and Paci, 2019; Koehler et al., 2019), and can later be used for health effects studies. Air pollution exposures have been associated with numerous health endpoints, including chronic diseases such as chronic obstructive pulmonary disease (Pan et al., 2018) and asthma (Benka-Coker et al., 2020), as well as acute diseases such as influenza (Landguth et al., 2020) and other respiratory illnesses (Cui et al., 2003; Dockery and Pope, 1994). Estimating air pollution-health associations is complicated by the reality that individuals are jointly exposed to a mixture of pollutants at a time. In addition, observational studies are often rife with missing observations, which may include missing exposure data or missing health data. In this dissertation, we aim to fill several gaps in the literature induced by these challenges.

1.1 Multipollutant Mixtures

At any point in time, an individual may be exposed to a number of environmental pollutants. A multipollutant mixture is defined as the joint exposure to three or more pollutants. The study of health effects associated with multipollutant mixtures is a top priority for environmental health scientists (NIEHS, 2018). Estimating health outcomes associated with multipollutant mixtures is challenging due to possible nonlinear effects and interactions, small effect sizes, and high correlation among pollutants. Recently, many statistical methods have been developed to address these challenges.

Statistical methods for multipollutant mixtures vary widely (Davalos et al., 2017; Hamra and Buckley, 2018; Sun et al., 2013; Taylor et al., 2016). The simplest models are additive models, such as generalized linear regression models. The next set of models permit interactions, typically in a generalized linear model framework. Usually, these models only include pairwise interactions to maintain adequate statistical power and interpretability. Regularization and shrinkage approaches have been applied to both additive models and models with interactions to improve estimation in the context of a large number of possibly correlated exposures (Carbajal-Arroyo et al., 2011; Herring, 2010; Lenters et al., 2016; Roberts and Martin, 2005; Winquist et al., 2014). Further methods for multipolluant mixtures include dimension reduction techniques. Dimension reduction models, such as principle components analysis (Roberts and Martin, 2006b) and Bayesian profile regression (Molitor et al., 2010), transform the exposure data, which may be high-dimensional, to a smaller subset to reduce the parameter space. Last, when the relationship between exposures and the health endpoint is hypothesized to be complex, nonparametric methods are favorable because they relax assumptions on the shape of the exposure-response function, therefore allowing complex interactions and nonlinear effects in the modeling framework. Nonparametric methods for multipollutant mixtures analyses include Bayesian kernel machine regression (Bobb et al., 2015) and classification and regression trees (Gass et al., 2014, 2015).

Epidemiological studies of the effects of multipollutant mixtures on human health have the potential to answer a variety of questions depending on the type of data and modeling approach (Braun et al., 2016). First, researchers may be interested in understanding the synergistic effect of exposures or predicting health outcomes associated with exposures. In these situations, the modeling approach may focus on estimating the form of the exposure-response function. Second, the primary goal may be identifying which components of the mixture truly have an effect on the health outcome. Along the same line, interest may lie in identifying which components of the mixture do not impact the health outcome. Identifying mixture components motivates a variable selection approach to analysis. Last, joint exposures may motivate iden-

tification and estimation of interaction effects among exposures. In this case, a model must incorporate interactions and perhaps also include a variable selection approach on the interaction terms. Defining the scientific objective of a study is the first step in selecting an appropriate statistical method for analysis.

As the study of multipollutant mixtures gains popularity and numerous statistical methods are developed, a gap remains in the literature regarding evaluation of the proposed methods and guidance on choosing a method. In Chapter 2, we investigate recently developed statistical methods for multipollutant mixtures. We evaluate the methods in a formal simulation study where we test each method's ability to answer epidemiological questions of interest.

1.2 Missing Exposure Data

Air pollution exposure assessments can be conducted on a variety of scales. Exposures can be measured for individuals, residences, or regions at resolutions ranging from seconds to days, weeks, or longer. Advances in technology currently allow ambient air pollution exposures to be measured on a personal level at very high resolution. Personal exposure monitoring devices are small enough to carry, and exposures to ambient environmental pollutants can be measured at 10-second intervals (Good et al., 2016; Koehler et al., 2019). Not only does high resolution personal exposure monitoring permit a more precise view of an individual's exposures over time, but the exposures are measured exactly where an individual is located, whether it be indoors, outdoors, at their home, or away from home. Even in a single location, such as home, different exposures to air pollution are expected when sleeping as opposed to other activities such as cooking on an indoor stove. Hence, personal exposure monitors offer improvements over area monitors in individual-level exposure assessments because personal monitors can detect exposure changes associated with an individual's specific location and activity.

Missing data is a common problem in studies where repeated measures are taken over time, such as in exposure assessments. Missing personal exposure data may be due to a number of factors including device malfunction, participant noncompliance, and exposures below the de-

vice's limit of detection. In short-term personal exposure monitoring, missing exposure data is a significant obstacle to maximizing use of the data. In a single day, missing minutes of exposure data may have a large influence on the entire daily exposure pattern. Complete case-only analyses and single imputation approaches can introduce bias and underestimate variability in the exposure data, particularly when the amount of missing data is high (Engels and Diehr, 2003; Junger and Ponce de Leon, 2015).

In a localized area, daily exposure patterns may be similar among multiple individuals. For example, different individuals sleep at home, commute to work, and cook meals at similar times each day. These shared activity patterns are likely to elicit similar exposure levels. In the presence of missing exposure data, shared activity patterns can help inform missing observations. In Chapter 3, we develop a Bayesian infinite hidden Markov model (Beal and Rasmussen, 2002) to impute missing multivariate exposure data for multiple people based on estimated latent states. The latent states represent unobserved time-activity patterns that are associated with unique distributions of exposures and may be shared among multiple people. We apply our method to an analysis of the Fort Collins Commuter Study (Good et al., 2016; Koehler et al., 2019) data to impute missing exposure data for three pollutants. In our analysis, we identify shared and unique time-activity patterns associated with exposures among multiple people during typical workdays in Fort Collins, Colorado, USA.

1.3 Missing Health Outcomes

Individual-level observational data are almost certain to contain some missing observations. When the primary goal is to estimate a health outcome associated with exposures, complete health outcome data are typically required. Commonly, individuals with missing health outcomes are dropped from the analysis, reducing sample size and power, and potentially introducing bias. In some situations, health outcomes may be partially missing, as opposed to fully missing. For example, time-to-event data may be censored or multinomial outcomes may

be missing for only some categories. In such situations, statistical methods that can analyze data with partially missing health outcomes are desirable.

During the global pandemic of coronavirus disease 2019 (COVID-19), scientists looking to explain spatial differences in disease transmission and severity identified the study of the effects of air pollution exposure on COVID-19 endpoints as a critically important area of research (Bhaskar et al., 2020). Though numerous ecological analyses suggest an association between increased air pollution exposure and negative COVID-19 endpoints (Bhaskar et al., 2020; Comunian et al., 2020; Copat et al., 2020; Frontera et al., 2020; Setti et al., 2020a), individual-level studies are needed to clarify the effects of air pollution and determine a causal link (Brandt and Mersha, 2021).

Individual-level COVID-19 health outcomes can take a variety of forms including infection status, death, and peak disease severity. Peak disease severity for individuals with confirmed COVID-19 can be subdivided into the following mutually exclusive categories: asymptomatic, symptomatic, hospitalized, admitted to an intensive care unit, placed on a mechanical ventilator, or death. The City and County of Denver, Colorado identified 57,027 individuals with confirmed COVID-19 between March 6, 2020 and February 28, 2021. During surges in the pandemic, there was inadequate staff capacity to follow up on all cases. For some individuals, it was known that they were symptomatic, but unknown if they were admitted to a hospital or intensive care unit, or placed on a mechanical ventilator. For others, it was known that they were not hospitalized or worse, but unknown if they were asymptomatic or symptomatic. The State of Colorado maintained accurate recordings of deaths caused by COVID-19. This verified death status was available for all 57,027 cases. As a result, all individuals had at least partial information regarding their peak COVID-19 severity classification.

Partial outcome information can be used to impute missing outcomes and inform estimation of the exposure-response relationship. In Chapter 4, we develop a Bayesian multinomial logistic regression model and multiple imputation approach for data with partially missing categorical outcomes. We estimate the association between long-term exposure to a mixture of

two ambient air pollutants and temperature and COVID-19 peak severity in a Denver, Colorado COVID-19 cohort. Just over 37% of cases had complete health outcome data, while nearly 63% of cases had partially missing health outcome data. The development of a modeling approach for partially missing health outcome data allows for a substantial increase in sample size, which leads to improved estimation and power over a complete case analysis.

1.4 Outline

The remainder of this dissertation is organized as follows. In Chapter 2, we conduct a formal simulation study evaluating five contemporary methods for estimating the association between exposure to multipollutant mixtures and health outcomes. We evaluate the methods on their ability to answer three specific epidemiological questions in a variety of exposure-response function scenarios. Our simulation study shows that the best method depends on both the data-generating mechanism and the primary research objective. We apply each method to estimate the association between lung function and a mixture of air pollutants and pesticides in children with asthma in Fresno, California.

In Chapter 3, we develop an infinite hidden Markov model for analyzing multiple time series of multivariate air pollution exposure data, where exposure values may be observed, missing at random, or below the limit of detection. We estimate hidden state structures among multiple time series to identify shared and unique activity patterns that give rise to pollutant exposures and to inform missing exposure data imputation. In simulation and validation studies, we demonstrate the estimation and imputation gains from our proposed method over independent analyses of multiple time series, a model with no temporal structure, and fixed-state approaches. We apply our proposed method to an analysis of 50 sampling days in the Fort Collins Commuter Study. We impute missing exposure data for three pollutants and identify time-activity patterns associated with the latent states.

We develop a Bayesian multinomial logistic regression model for data with partially missing categorical outcomes in Chapter 4. We develop a multiple imputation approach to im-

pute missing health outcome data and demonstrate the estimation gains from our proposed approach over complete case analyses in a variety of scenarios. We apply our method to a Denver, Colorado COVID-19 cohort to estimate the association between long-term exposure to fine particulate matter, ozone, and temperature and COVID-19 peak severity in a sample of 55,273 cases confirmed between March 6, 2020 and February 28, 2021. We propose a visualization approach to make inference on the odds ratio for each severity category that is associated with exposures. We find that fine particulate matter is associated with an increased risk of severe COVID-19. Our analysis also suggests possible interaction effects among the exposures.

We provide a summary in Chapter 5. We describe how our work has filled several gaps in the literature regarding statistical methods for air pollution, and we propose future directions for related research.

Chapter 2

Model choice for estimating the association between exposure to chemical mixtures and health outcomes: A simulation study

2.1 Introduction

Individuals are continuously exposed to complex mixtures of environmental chemicals. Mounting evidence from epidemiological studies links environmental exposures to increased morbidity and mortality (Di et al., 2017a,b; Dockery and Pope, 1994; Dockery et al., 1993; Pan et al., 2018). Traditional epidemiological studies have focused on a single pollutant and additive models with a small number of exposures; however, studying pollutants in isolation can lead to biased estimates (Slama and Vrijheid, 2015; Weisskopf et al., 2018) and does not reflect the reality that people are jointly exposed to mixtures of pollutants. Hence, interest is rapidly growing in studying health outcomes associated with simultaneous exposure to mixtures of pollutants (i.e. multipollutant mixtures) (Dominici et al., 2010; Samet, 2005). The National Institute for Environmental Health Sciences (NIEHS) identified the study of multipollutant mixtures as a goal in its 2012-2017 strategic plan while noting that this will require novel quantitative approaches (NIEHS, 2012). As such, numerous statistical methods have been proposed. There is a need to identify the most appropriate statistical methods currently available for estimating health outcomes associated with exposure to multipollutant mixtures (Hamra and Buckley, 2018; Taylor et al., 2016).

Studying health outcomes associated with exposure to multipollutant mixtures is complicated by small effect sizes, highly correlated exposures, possible nonlinear and interaction effects, and often small sample sizes. In this context, traditional regression methods are often inadequate as they may yield biased or unstable estimates (Witte and Greenland, 1996) and

have low power to detect effects, especially in the case of nonlinear associations and interactions. Common methods designed for variable selection tend to incorrectly select predictors when many predictors are highly correlated (Barrera-Gómez et al., 2017) and classical model selection techniques ignore uncertainty in both the selected model and selected mixture components when estimating the exposure-response function (Clyde, 2000; Hoeting et al., 1999).

In a broad literature review, Davalos et al. (2017) identified five classes of methods currently used in mixtures analyses: additive main effects (AME), effect measure modification (EMM), unsupervised dimension reduction (UDR), supervised dimension reduction (SDR), and nonparametric (NP). AME and EMM methods are typically regression based. AME methods only allow additive effects, while EMM methods include multiplicative interactions. Hierarchical and penalized regression methods have been applied to AME and EMM models to identify important mixture components and improve precision (Carbajal-Arroyo et al., 2011; Herring, 2010; Lenters et al., 2016; Roberts and Martin, 2005; Winquist et al., 2014). The next two groups are dimension reduction techniques (UDR and SDR), which transform the exposure data to reduce the dimension of the predictor and, therefore, the required parameter space. UDR methods such as k-means (Austin et al., 2012; Zanobetti et al., 2014) transform exposure data without regard to the health outcome (Pearce et al., 2014, 2015, 2016; Sacks et al., 2012). SDR methods, including supervised principle components analysis (Roberts and Martin, 2006b), let the outcome inform exposure data transformation (Carrico et al., 2015; Nikolov et al., 2007; Pachon et al., 2012; Roberts and Martin, 2006a; Sun et al., 2013; Wold et al., 1984). Finally, NP methods like Bayesian kernel machine regression (Bobb et al., 2015) are flexible data-driven techniques for estimating a complex exposure-response function that may include interactions and nonlinear effects (Gass et al., 2014, 2015).

Choosing an appropriate statistical model depends on the research objectives (Braun et al., 2016; Taylor et al., 2016) and requires understanding the empirical performance of methods. Recent studies have compared several methods in subsets of the model classes described above. Among those evaluated include linear regression AME (Agier et al., 2016) and EMM (Barrera-

Gómez et al., 2017) methods, principle components analysis (Sun et al., 2013), structural equation models (Chiu et al., 2018), Bayesian kernel machine regression (Chiu et al., 2018), and Bayesian semiparametric regression (Antonelli et al., 2020). These studies highlight challenges induced by highly correlated data in estimating complex exposure-response functions and characterizing uncertainty. To our knowledge, there has been no formal evaluation of methods from all five classes identified by Davalos et al. (2017) in a single simulation study.

Evaluating the empirical performance of methods across a wide spectrum of model classes is important as it guides researchers in choosing across classes of models and aids in interpreting results and understanding the limitations of epidemiological studies using these methods. In addition, the existing literature is sparse with regards to a comparison among Bayesian methods, which are favorable in the multipollutant setting as they can incorporate prior information and fully characterize uncertainty (Gibson et al., 2019; Hamra and Buckley, 2018; Sun et al., 2013). To this end, we focus on a comparison of Bayesian methods across a variety of model classes in this paper. By comparing performance across classes of models, researchers can also gain insight into promising future directions for statistical methods development.

Motivated by research linking mixtures of air pollutant and pesticide exposures to child respiratory health, we conducted a simulation study to compare contemporary methods developed for estimating the association between health outcomes and exposure to multipollutant mixtures. We considered one method from each of the five classes identified by Davalos et al. (2017) and evaluated each method in three data-generating scenarios. The data-generating scenarios cover a range of linear to nonlinear functions of multiple pollutants with synergistic effects on the response in order to test each method in its ability to estimate both simple and complex exposure-response functions that may be encountered in practice.

In contrast to many recent studies that have compared methods from a conceptual standpoint or compared their performance in the analysis of a single data set, the primary contribution of our work is to compare diverse methods in a simulation study addressing a variety of research questions. Specifically, we quantified four aspects of model performance correspond-

ing to previously identified epidemiological questions of interest: 1) how well does the model estimate the exposure-response function, 2) can the model identify important mixture components, 3) can the model identify components not associated with the outcome, and 4) can the model identify interactions among exposures (Braun et al., 2016).

A secondary contribution of our work is to provide software for the tested methods that currently lack software. Our simulation study describes the strengths and weaknesses of each method and available software encourages practitioners to use the most appropriate methods in a given application. Software in the form of the `R` package `mmpack` (Hoskovec, 2019) is available at github.com/lvhoskovec/mmpack to reproduce the simulation. Further, the software allows researchers to easily conduct a simulation study using the same methods and simulated exposure-response functions but substituting in their own exposure data which will have a different correlation structure and may result in different model performance. Hence, researchers can determine which methods are most appropriate for their own study. Finally, we applied each method to a data analysis of a cohort study investigating the relationship between air pollutant and pesticide exposures and lung function in children with asthma. We describe the differences in results among the methods, highlighting the importance of model choice.

2.2 Data

2.2.1 Health Data

This study was approved by the Institutional Review Board of Colorado State University, Protocol Number 19-9437H. This was a secondary data analysis from a closed cohort with all personal identifying information stripped from the database. We used data from Fresno Asthmatic Children's Environment Study (FACES). The study design, including recruitment, eligibility criteria, and measurement procedures, is described elsewhere (Gale et al., 2012; Mann et al., 2010; Margolis et al., 2009; Mortimer et al., 2008; Noth et al., 2011; Padula et al., 2015). FACES includes data for children aged 6-11 years with asthma symptoms at the time of enrollment and living within a 20 kilometer radius of one of Fresno's EPA air quality monitoring sites. The health out-

come of interest was baseline forced expiratory volume in the first second (FEV_1) measured via spirometry. We regressed FEV_1 on age, sex, height and ethnicity and used the residuals as the outcome in our data analysis (Benka-Coker et al., 2020; Raanan et al., 2016; Van Sickle et al., 2011). Age, sex, height, and ethnicity are well-known predictors of FEV_1 so we remove all variation from these predictors before looking into the effects of air pollution and pesticide exposure on FEV_1. Other covariates have not been as well studied regarding their association with FEV_1 and are included in the model as potential confounding variables. Complete exposure, health, and covariate data were available for 153 children.

The data contain information on covariates and potential confounding variables. We included average temperature and precipitation over three months, the temporal scale of the pesticide exposure data, prior to baseline as covariates. Subject-specific covariates include body mass index (BMI, kg/m^2) and indicators for: self-reported residence within one block of a freeway, any smoking in the home, positive atopy or allergy test, modified Global Initiative for Asthma (GINA) score ≥ 3 at baseline, household income greater than \$30K/year, mother having post-secondary education, child not covered by insurance, and season of baseline spirometry test. Temperature, precipitation, and BMI were scaled to have mean 0 and variance 1. Approximately 1% of the covariate data was missing, including any smoking in the home (16%), household income (3%), and mother having post-secondary education (1%). As all covariates with missing data were binary variables, we singly imputed the missing values with 0 and then added a dummy variable for each covariate with any missing data that indicated which values of that covariate were missing. We summarize the demographic characteristics of the sample in Appendix A.1 (Table A.1).

2.2.2 Air Pollution and Pesticide Data

We obtained air pollution data from the EPA Air Quality System Data Mart. Air pollutant concentrations were calculated as 24-hour averages for particles ≤ 2.5 μm in aerodynamic diameter ($PM_{2.5}$) and particles ≤ 10 μm in aerodynamic diameter (PM_{10}), 8-hour daily maximum

levels for ozone (O_3) and one-hour daily maximum levels for nitrogen dioxide (NO_2) (Mann et al., 2010). Concentrations were taken from the air monitoring site closest to each child's residence and exposure levels were summarized as averages over three months prior to baseline spirometry tests to be consistent with available pesticide exposure data. Due to right-skewed distributions, air pollutant exposures were square-root transformed and then scaled to have mean 0 and variance 1.

We obtained data on the date, location, and amount (kilograms) of applied agricultural pesticides from the California Pesticide Use Report (PUR) (California Department of Pesticide Regulation, 2015). Based on previous evidence linking pesticide exposure to respiratory illness (Bulathsinghala and Shaw, 2014; Colovic et al., 2013), we considered three pesticide classes: carbamates (C), methyl bromide (MeBr), and organophosphates (OP). Pesticide exposures were estimated using the `purexposure` (Severson, 2019) package in `R`. We applied inverse distance weighting to the total reported pesticide amount over three months prior to baseline spirometry tests (as PUR reports are aggregated quarterly) to estimate exposures within a 3km buffer of each child's residence. Pesticide exposures were also highly skewed and so were square-root transformed and then scaled to have mean 0 and variance 1.

Exposure data summary statistics are shown in Table 2.1, and Spearman correlations among exposures are shown in Table 2.2. Strong Spearman correlation existed between NO_2 and $PM_{2.5}$ (ρ = 0.88) and between NO_2 and PM_{10} (ρ = 0.72). Moderate Spearman correlation existed between $PM_{2.5}$ and PM_{10} (ρ = 0.67), O_3 and NO_2 (ρ = -0.63), O_3 and $PM_{2.5}$ (ρ = -0.54), O_3 and OP (ρ = 0.53), and OP and NO_2 (ρ = -0.53).

Table 2.1: Pesticide and air pollutant exposure data summary statistics. Table shows mean, standard deviation (SD), minimum, 25th percentile, median, 75th percentile, and maximum concentration for each exposure.

	mean	SD	min	25^{th}	median	75^{th}	max
C × 10^6 (kg/3km^2)	0.15	0.33	0.00	0.00	0.00	0.15	2.35
MeBr × 10^6 (kg/3km^2)	3.88	9.90	0.00	0.00	0.00	0.00	48.92
OP × 10^6 (kg/3km^2)	0.93	1.08	0.00	0.00	1.11	1.17	5.40
O_3 (ppb)	0.04	0.01	0.01	0.03	0.04	0.04	0.06
NO_2 (ppb)	15.48	3.26	9.49	12.64	14.42	17.96	23.07
$PM_{2.5}$ (μg/m^3)	16.35	9.80	6.66	10.14	11.23	18.20	40.21
PM_{10} (μg/m^3)	37.89	10.68	19.55	30.30	32.49	47.23	65.94

Table 2.2: Spearman correlation among all pairs of air pollutant and pesticide exposures.

	MeBr	OP	O_3	NO_2	$PM_{2.5}$	PM_{10}
C	0.27	0.12	0.09	0.08	0.06	0.01
MeBr		-0.08	0.02	0.07	-0.03	-0.13
OP			0.53	-0.53	-0.38	-0.24
O_3				-0.63	-0.54	-0.22
NO_2					0.88	0.72
$PM_{2.5}$						0.67

2.3 Statistical Methods

Our primary interest was to estimate the association between exposures to p pollutants $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and a continuous outcome y_i, while controlling for q potential confounders $\mathbf{w}_i = (w_{i1}, \ldots, w_{iq})^T$ in a sample $i = 1, \ldots, n$. We considered five recently proposed methods. The first two are the AME model nonparametric Bayes shrinkage with main effects only (NPBr) and the EMM model nonparametric Bayes shrinkage with main effects and all pairwise multiplicative interactions (NPB) as proposed by Herring (2010). The next two models are unsupervised (UPR) and supervised Bayesian profile regression (SPR) as proposed by Molitor et al. (2010). The fifth is the NP model Bayesian kernel machine regression (BKMR) (Bobb et al., 2015). We chose

these methods since they represent the five classes identified by Davalos et al. (2017) and are recently developed Bayesian methods for estimating health outcomes associated with exposure to multipollutant mixtures. These five methods cover a variety of exposure-response function shapes, handle multicollinearity in various ways, and include options for variable selection. BKMR is presented exactly as proposed by Bobb et al. (2015); NPB and SPR have been modified to accommodate a continuous outcome with normal residuals rather than the logistic model originally proposed by Herring (2010) and Molitor et al. (2010), respectively; and NPBr and UPR are further modifications of those previously introduced methods. For a baseline comparison, we also included a normal linear model with main effects only (LM) and with all pairwise interactions (LM-int), both estimated with least squares. All models considered in this paper have the form

$$y_i = h(\mathbf{x}_i) + \mathbf{w}_i^T \boldsymbol{\gamma} + \epsilon_i, \tag{2.1}$$

where ϵ_i are independent $\mathrm{N}(0, \sigma^2)$ and $h(\mathbf{x}_i)$ represents the exposure-response function. All models were fit in R version 3.6.0 (R Core Team, 2018).

2.3.1 Nonparametric Bayes Shrinkage

Nonparametric Bayes shrinkage (Herring, 2010) was originally introduced as a logistic regression EMM model and was adapted to the linear regression setting used here. We consider two variations. NPB, originally proposed by Herring (2010), is an EMM model including main effects and all pairwise interactions, where

$$h(\mathbf{x}_i) = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} x_{ij}x_{ik}\zeta_{jk}. \tag{2.2}$$

NPBr is a reduced AME model not originally proposed in Herring (2010) that includes only main effects. In NPBr, the exposure-response function is

$$h(\mathbf{x}_i) = \sum_{j=1}^{p} x_{ij}\beta_j. \tag{2.3}$$

The prior distributions on the intercept (γ_0), regression coefficients for covariates ($\boldsymbol{\gamma}$), and error precision (σ^{-2}) are

$$\begin{aligned} \gamma_0 &\sim N(\mu_0, \kappa_0^2) && (2.4) \\ \boldsymbol{\gamma}|\boldsymbol{\mu}_\gamma, \kappa^2 &\sim \mathrm{N}(\boldsymbol{\mu}_\gamma, \kappa^2 \mathbf{I}) && (2.5) \\ \sigma^{-2} &\sim \mathrm{Gamma}(\alpha_\sigma, \beta_\sigma). && (2.6) \end{aligned}$$

Both models place a Dirichlet Process (DP) prior on the exposure regression coefficients. The base distribution of the DP is a finite mixture of a normal distribution and a point mass at 0 to induce sparsity in the model. Hence, some coefficients are set exactly to 0, effectively selecting out variables that do not contribute to the health outcome. Correlated exposures can be clustered and assigned equal regression coefficients to reduce variance (Dunson et al., 2008; Herring, 2010). This effectively reparameterizes the model to have a single effect for the sum of two correlated predictors and is particularly advantageous in situations where it is difficult to differentiate the effects of two highly correlated predictors. The DP prior for main effects is constructed as

$$\begin{aligned} \beta_j|D_1 &\sim D_1, \qquad j = 1, \ldots p && (2.7) \\ D_1|\alpha_1, D_{01} &\sim \mathrm{DP}(\alpha_1 D_{01}) \\ D_{01}|\pi_{01}, G_1 &= \pi_{01}\delta_0 + (1-\pi_{01})G_1 \\ G_1|\mu_1, \phi_1^2 &\equiv \mathrm{N}(\mu_1, \phi_1^2), \end{aligned}$$

where δ_0 represents the Dirac delta function at 0. The model is completed with standard hyper-priors $\alpha_1 \sim \mathrm{Gamma}(\alpha_{\alpha 1}, \beta_{\alpha 1})$, $\pi_{01} \sim \mathrm{Beta}(\alpha_{\pi 1}, \beta_{\pi 1})$, $\mu_1 \sim \mathrm{N}(0, \sigma^2_{\mu 1})$, and $\phi_1^{-2} \sim \mathrm{Gamma}(\alpha_{\phi 1}, \beta_{\phi 1})$.

The DP prior for interactions is similarly constructed. Specifically,

$$\begin{aligned}
\zeta_{jk}|D_2 &\sim D_2, \qquad j = 1,\dots p-1 \,\&\, k = j+1,\dots,p \\
D_2|\alpha_2, D_{02} &\sim \text{DP}(\alpha_2 D_{02}) \\
D_{02}|\pi_{02}, G_2 &= \pi_{02}\delta_0 + (1-\pi_{02})G_2 \\
G_2|\mu_2, \phi_2^2 &\equiv \text{N}(\mu_2, \phi_2^2).
\end{aligned} \tag{2.8}$$

The hyperpriors on α_2, π_{02}, μ_2, and ϕ_2^{-2} come from the same families specified for the main effects. The distributions on the main effects and interactions are independent a priori.

To sample from the posterior distribution, we introduce latent allocation variables. For main effects, we introduce the variable S such that $S_j = c \Leftrightarrow \beta_j = \theta_c$. Hence, the variable S is used to cluster regression coefficients. Let K denote the number of clusters, or unique regression coefficients, including the null group. We describe the DP prior on each β_j by taking $\lim_{K\to\infty}$ of

$$\begin{aligned}
\beta_j|S_j = c, \theta_c &\sim \delta_{\theta_c} \\
\theta_c &\sim G_1 \\
G_1 &\equiv \text{N}(\mu_1, \phi_1^2) \\
S_j|\pi_{01}, \boldsymbol{\pi} &\sim \text{Categorical}\,(\pi_{01}, (1-\pi_{01})\pi_1, \dots, (1-\pi_{01})\pi_{K-1}), \quad \sum_{c=1}^{K-1} \pi_c = 1 \\
\pi_{01} &\sim \text{Beta}(\alpha_{\pi 1}, \beta_{\pi 1}) \\
\boldsymbol{\pi}|\alpha_1 &\sim \text{Dirichlet}\Big(\frac{\alpha_1}{K-1}, \dots, \frac{\alpha_1}{K-1}\Big) \\
\alpha_1 &\sim \text{Gamma}(\alpha_{\alpha 1}, \beta_{\alpha 1}),
\end{aligned} \tag{2.9}$$

where δ_{θ_c} denotes the Dirac delta function at the value θ_c, and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{K-1})$ are the assignment probabilities for the $K-1$ non-null clusters for main effects. Let $\mathbf{S}^{(j)}$ denote all allocation variables except the j^{th} one. Let $c = 0$ denote the null cluster and $c \neq 0$ denote a non-null cluster. Let $p_0^{(j)}$ represent the number of main effect regression coefficients assigned to the null

cluster excluding the j^{th} coefficient, and let $p_c^{(j)}$ represent the number of main effect regression coefficients assigned to the c^{th} cluster excluding the j^{th} coefficient. The prior distribution for the allocation variable S_j, conditional on $\mathbf{S}^{(j)}$, is

$$P(S_j = c|\mathbf{S}^{(j)}) = \begin{cases} \left(\frac{p_0^{(j)}+\alpha_{\pi 1}}{p+\alpha_{\pi 1}+\beta_{\pi 1}-1}\right), & c = 0 \\ \left(\frac{p-p_0^{(j)}+\beta_{\pi 1}-1}{p+\alpha_{\pi 1}+\beta_{\pi 1}-1}\right)\left(\frac{p_c^{(j)}}{p+\alpha_1-p_0^{(j)}-1}\right), & c \neq 0 \cap p_c^{(j)} \neq 0 \\ \left(\frac{p-p_0^{(j)}+\beta_{\pi 1}-1}{p+\alpha_{\pi 1}+\beta_{\pi 1}-1}\right)\left(\frac{\alpha_1}{p+\alpha_1-p_0^{(j)}-1}\right), & c \neq 0 \cap p_c^{(j)} = 0. \end{cases} \tag{2.10}$$

In (2.10), the first line represents the assignment probability to the null cluster, the second line represents the assignment probability to an existing occupied cluster, and the third line represents the assignment probability to a new unoccupied cluster.

For interactions in NPB, we introduce an allocation variable Q such that $Q_{jk} = z \Leftrightarrow \zeta_{jk} = \psi_z$. Let M denote the number of clusters for the interaction regression coefficients, including the null group. We describe the DP prior on each ζ_{jk} by taking $\lim_{M\to\infty}$ of

$$\begin{aligned} \zeta_{jk}|\mathbf{Q}_{jk} = z, \psi_z &\sim \delta_{\psi_z} \\ Q_{jk}|\pi_{02}, \boldsymbol{\pi}^* &\sim \text{Categorical}\left(\pi_{02}, (1-\pi_{02})\pi_1^*, \ldots, (1-\pi_{02})\pi_{M-1}^*\right), \quad \sum_{z=1}^{M-1} \pi_z^* = 1 \\ \psi_z &\sim G_2 \\ G_2 &\equiv \text{N}(\mu_2, \phi_2^2) \\ \pi_{02} &\sim \text{Beta}(\alpha_{\pi 2}, \beta_{\pi 2}) \\ \boldsymbol{\pi}^*|\alpha_2 &\sim \text{Dirichlet}\left(\frac{\alpha_2}{M-1}, \ldots, \frac{\alpha_2}{M-1}\right) \\ \alpha_2 &\sim \text{Gamma}(\alpha_{\alpha 2}, \beta_{\alpha 2}), \end{aligned} \tag{2.11}$$

where $\boldsymbol{\pi}^* = \left(\pi_1^*, \ldots, \pi_{M-1}^*\right)$ are the assignment probabilities for the $M-1$ non-null clusters for interactions. Let $\mathbf{Q}^{(jk)}$ denote all allocation variables except Q_{jk}. Let r be the number of 2-way interactions $\left(r = \frac{p(p-1)}{2}\right)$. Then $r_0^{(jk)}$ is the number of interaction regression coefficients assigned to the null cluster excluding the interaction term between the j^{th} and k^{th} pollutants

(i.e. the jk^{th} term), and $r_z^{(jk)}$ is the number of interaction regression coefficients assigned to the z^{th} cluster excluding the jk^{th} coefficient. The prior distribution on the allocation variable Q_{jk}, conditional on $\mathbf{Q}^{(jk)}$, is

$$Pr(Q_{jk}=z|\mathbf{Q}^{(jk)}) = \begin{cases} \left(\frac{r_0^{(jk)}+\alpha_{\pi 2}}{r+\alpha_{\pi 2}+\beta_{\pi 2}-1}\right), & z=0 \\ \left(\frac{r-r_0^{(jk)}+\beta_{\pi 2}-1}{r+\alpha_{\pi 2}+\beta_{\pi 2}-1}\right)\left(\frac{r_z^{(jk)}}{r+\alpha_2-r_0^{(jk)}-1}\right), & z\neq 0 \cap r_z^{(jk)}\neq 0 \\ \left(\frac{r-r_0^{(jk)}+\beta_{\pi 2}-1}{r+\alpha_{\pi 2}+\beta_{\pi 2}-1}\right)\left(\frac{\alpha_2}{r+\alpha_2-r_0^{(jk)}-1}\right), & z\neq 0 \cap r_z^{(jk)}=0. \end{cases} \tag{2.12}$$

We use a blocked Gibbs sampler for efficient mixing of the posterior distribution. Let $\boldsymbol{\beta}$ denote the $p\times 1$ vector of main effect regression coefficients and let $\boldsymbol{\theta}$ denote the $(K-1)\times 1$ vector of unique non-null main effect regression coefficients. At each iteration of the MCMC sampler, we create a $p\times(K-1)$ binary allocation matrix, $\mathbf{T}_1$, that identifies to which θ_c each β_j belongs. If a regression coefficient falls in the null cluster, then it has zeros for all elements in its row. Similarly for interactions, let $\boldsymbol{\zeta}$ denote the vector of r interaction regression coefficients and let $\boldsymbol{\psi}$ denote the $(M-1)\times 1$ vector of unique non-null interaction regression coefficients. At each iteration, we create the $r\times(M-1)$ binary allocation matrix, $\mathbf{T}_2$, that identifies to which ψ_z each ζ_{jk} belongs. Finally, let $\mathbf{X}$ denote the $n\times p$ matrix of exposure data, $\mathbf{Z}$ the $n\times r$ matrix of pairwise multiplicative interactions between elements in $\mathbf{X}$, and $\mathbf{W}$ the $n\times q$ matrix of covariate data. We define the block parameters

$$\begin{aligned} \mathbf{A} &= (\mathbf{XT_1}, \mathbf{ZT_2}, \mathbf{1}, \mathbf{W}) \\ \boldsymbol{\Delta}^T &= (\boldsymbol{\theta}, \boldsymbol{\psi}, \gamma_0, \boldsymbol{\gamma}) \\ \boldsymbol{\Sigma}^{-1} &= \text{diag}\left(\phi_1^{-2},\cdots,\phi_1^{-2},\, \phi_2^{-2},\cdots,\phi_2^{-2}, \kappa_0^{-2}, \kappa^{-2},\cdots,\kappa^{-2}\right) \\ \mathbf{m}^T &= (\mu_1,\ldots,\mu_1,\mu_2,\ldots,\mu_2,\mu_0,\boldsymbol{\mu}_\gamma). \end{aligned}$$

The parameter $\boldsymbol{\Sigma}^{-1}$ is a $(K+M+q-1)$-dimensional diagonal matrix where the first $K-1$ diagonals are ϕ_1^{-2}, the next $M-1$ diagonals are ϕ_2^{-2}, the $K+M-1$ diagonal is κ_0^{-2}, and the final q

diagonals are κ^{-2}. The parameter $\mathbf{m}$ is a $(K+M+q-1)$-dimensional vector where the first $K-1$ elements are μ_1, the next $M-1$ elements are μ_2, the $K+M-1$ element is μ_0, and the final q elements are defined by the vector $\boldsymbol{\mu}_\gamma$. We sample all exposure (main effect and interaction, if applicable) and covariate regression coefficients as a block from

$$\boldsymbol{\Delta}|\cdot \sim \mathrm{N}\left((\sigma^{-2}\mathbf{A}^T\mathbf{A}+\boldsymbol{\Sigma}^{-1})^{-1}(\sigma^{-2}\mathbf{A}^T\mathbf{y}+\boldsymbol{\Sigma}^{-1}\mathbf{m}),(\sigma^{-2}\mathbf{A}^T\mathbf{A}+\boldsymbol{\Sigma}^{-1})^{-1}\right).$$

We sample the error precision from

$$\sigma^{-2}|\cdot \sim \mathrm{Gamma}\Big(\alpha_\sigma+\frac{n}{2},\beta_\sigma+\frac{1}{2}\left(\mathbf{y}-\mathbf{A}\boldsymbol{\Delta}\right)^T\left(\mathbf{y}-\mathbf{A}\boldsymbol{\Delta}\right)\Big).$$

We sample the DP base distribution parameters for main effects and interactions from

$$\begin{aligned}
\mu_1|\cdot &\sim \mathrm{N}\left(\frac{\phi_1^{-2}\sum_{c=1}^{K-1}\theta_c}{(K-1)\phi_1^{-2}+\sigma_{\mu 1}^{-2}},\frac{1}{(K-1)\phi_1^{-2}+\sigma_{\mu 1}^{-2}}\right)\\
\phi_1^{-2}|\cdot &\sim \mathrm{Gamma}\left(\alpha_{\phi 1}+\frac{K-1}{2},\beta_{\phi 1}+\frac{1}{2}(\boldsymbol{\theta}-\mu_1\mathbf{1})^T(\boldsymbol{\theta}-\mu_1\mathbf{1})\right)\\
\mu_2|\cdot &\sim \mathrm{N}\left(\frac{\phi_2^{-2}\sum_{c=1}^{M-1}\psi_c}{(M-1)\phi_2^{-2}+\sigma_{\mu 2}^{-2}},\frac{1}{(M-1)\phi_2^{-2}+\sigma_{\mu 2}^{-2}}\right)\\
\phi_2^{-2}|\cdot &\sim \mathrm{Gamma}\left(\alpha_{\phi 2}+\frac{M-1}{2},\beta_{\phi 2}+\frac{1}{2}(\boldsymbol{\psi}-\mu_2\mathbf{1})^T(\boldsymbol{\psi}-\mu_2\mathbf{1})\right).
\end{aligned}$$

We sample the DP concentration parameter α_1 for main effects from

$$\begin{aligned}
\eta &\sim \mathrm{Beta}(\alpha_1+1,p)\\
\pi_\eta &= \frac{(\alpha_{\alpha 1}+K-1)/(p(\beta_{\alpha 1}-\log(\eta))}{(\alpha_{\alpha 1}+K-1)/(p(\beta_{\alpha 1}-\log(\eta))+1}\\
\alpha_1|\cdot &\sim \pi_\eta * \mathrm{Gamma}\left(\alpha_{\alpha 1}+K,\beta_{\alpha 1}+\log(\eta)\right)+\\
&\quad (1-\pi_\eta)*\mathrm{Gamma}\left(\alpha_{\alpha 1}+K-1,\beta_{\alpha 1}-\log(\eta)\right).
\end{aligned}$$

We sample the DP concentration parameter α_2 for interactions from

$$\begin{aligned}
\eta_2 &\sim \text{Beta}(\alpha_2+1, r) \\
\pi_{\eta 2} &= \frac{(\alpha_{\alpha 2}+M-1)/(r(\beta_{\alpha 2}-\log(\eta_2))}{(\alpha_{\alpha 2}+M-1)/(r(\beta_{\alpha 2}-\log(\eta_2))+1} \\
\alpha_2|\cdot &\sim \pi_{\eta 2} * \text{Gamma}\left(\alpha_{\alpha 2}+M, \beta_{\alpha 2}+\log(\eta_2)\right) + \\
&\quad (1-\pi_{\eta 2}) * \text{Gamma}\left(\alpha_{\alpha 2}+M-1, \beta_{\alpha 2}-\log(\eta_2)\right).
\end{aligned}$$

To update the allocation variables, **S**, for main effects, we calculate the proportional probabilities

$$\begin{aligned}
P(S_j=0|\text{rest}) &\propto \left(\frac{p_0^{(j)}+\alpha_\pi}{p+\alpha_\pi+\beta_\pi-1}\right) f(y|\boldsymbol{\beta}^{(j)}, \beta_j=0, \sigma^2) \\
P(S_j=c|\text{rest}) &\propto \left(\frac{p-p_0^{(j)}+\beta_\pi-1}{p+\alpha_\pi+\beta_\pi-1}\right)\left(\frac{p_c^{(j)}}{p+\alpha_1-p_0^{(j)}-1}\right) f(y|\beta^{(j)}, \beta_j=\theta_c, \sigma^2) \\
P(S_j=c^*|\text{rest}) &\propto \left(\frac{p-p_0^{(j)}+\beta_\pi-1}{p+\alpha_\pi+\beta_\pi-1}\right)\left(\frac{\alpha_1}{p+\alpha_1-p_0^{(j)}-1}\right)(2\pi)^{-\frac{n}{2}}(\sigma^{-2})^{\frac{n}{2}}(\phi^{-2})^{\frac{1}{2}} \\
&\times (\sigma^{-2}\sum_{i=1}^{n}(x_{ij})^2+\phi^{-2})^{-\frac{1}{2}} \\
&\times \exp\Big\{-\frac{1}{2}\Big[\sigma^{-2}\sum_{i=1}^{n}\left(y_i-\gamma_0-\mathbf{x}_i^{(j)^T}\boldsymbol{\beta}^{(j)}-\mathbf{z_i}^T\boldsymbol{\zeta}-\mathbf{w_i}^T\boldsymbol{\gamma}\right)^2+\phi^{-2}\mu^2- \\
&\qquad \frac{\left(\sigma^{-2}\sum_{i=1}^{n}(y_i-\gamma_0-\mathbf{x}_i^{(j)^T}\boldsymbol{\beta}^{(j)}-\mathbf{z_i}^T\boldsymbol{\zeta}-\mathbf{w_i}^T\boldsymbol{\gamma})(x_{ij})+\phi^{-2}\mu\right)^2}{\sigma^{-2}\sum_{i=1}^{n}(x_{ij})^2+\phi^{-2}}\Big]\Big\},
\end{aligned}$$

for all currently occupied clusters c and a single new unoccupied cluster c^*. We sample the allocation variables from a normalized multinomial distribution using the above probabilities. The allocation variables **Q** for interactions are updated similarly.

If the j^{th} main effect regression coefficient is assigned to a new cluster c^* then we draw a new coefficient θ^* from

$$\begin{aligned}
\mu_\theta &= \frac{\sigma^{-2}\sum_{i=1}^{n}(y_i - \gamma_0 - \mathbf{x}_i^{(j)^T}\boldsymbol{\beta}^{(j)} - \mathbf{z_i}^T\boldsymbol{\zeta} - \mathbf{w_i}^T\boldsymbol{\gamma})(x_{i_j}) + \phi_1^{-2}\mu_1}{\sigma^{-2}\sum_{i=1}^{n}(x_{i_j})^2 + \phi_1^{-2}} \\
\Sigma_\theta &= \frac{1}{\sigma^{-2}\sum_{i=1}^{n}(x_{i_j})^2 + \phi_1^{-2}} \\
\theta^*|\cdot &\sim \mathrm{N}(\mu_\theta, \Sigma_\theta).
\end{aligned}$$

Similarly, if the jk^{th} interaction regression coefficient is assigned to a new cluster z^* then we draw a new coefficient ψ^* from

$$\begin{aligned}
\mu_\psi &= \frac{\sigma^{-2}\sum_{i=1}^{n}(y_i - \gamma_0 - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{z_i}^{(jk)^T}\boldsymbol{\zeta}^{(jk)} - \mathbf{w_i}^T\boldsymbol{\gamma})(z_{i_{jk}}) + \phi_2^{-2}\mu_2}{\sigma^{-2}\sum_{i=1}^{n}(z_{i_{jk}})^2 + \phi_2^{-2}} \\
\Sigma_\psi &= \frac{1}{\sigma^{-2}\sum_{i=1}^{n}(z_{i_{jk}})^2 + \phi_2^{-2}} \\
\psi^*|\cdot &\sim \mathrm{N}(\mu_\psi, \Sigma_\psi).
\end{aligned}$$

Posterior inclusion probabilities (PIPs) are calculated for each mixture component as the posterior probability of the regression coefficient being assigned a non-zero value. Both NPBr and NPB were fit using the `R` package `mmpack` (Hoskovec, 2019).

2.3.2 Bayesian Profile Regression

Bayesian profile regression is a dimension reduction approach that classifies pollutant exposure profiles, $\mathbf{x}_i$, into a parsimonious set of clusters using a DP mixture model (DPMM) (Molitor et al., 2010, 2011). Each cluster represents a group of observations with similar exposure levels across the vector of pollutants. The health outcome is regressed on cluster indicators to estimate health risks associated with each cluster. We model the health outcome y_i as

$$y_i|z_i = c, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2 \sim \mathrm{N}(\theta_c + \mathbf{w}_i^T\boldsymbol{\gamma}, \sigma^2),$$

where $z_i = c$ is a latent variable indicating that exposure profile i is assigned to cluster c, and θ_c is a cluster-specific intercept. Hence, the exposure-response function is constant for all subjects in the same cluster. That is,

$$h(\mathbf{x}_i) = \theta_c \tag{2.13}$$

if profile $\mathbf{x}_i$ is assigned to cluster c. Conditional on cluster assignment, the model for an individual exposure profile is

$$\begin{aligned} \mathbf{x}_i | z_i = c, \boldsymbol{\mu}_c, \Sigma_c &\sim \mathrm{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \\ \boldsymbol{\mu}_c &\sim \mathrm{N}(\boldsymbol{\nu}_0, \boldsymbol{\Lambda}_0) \\ \boldsymbol{\Sigma}_c^{-1} &\sim \mathrm{Wish}_\mathrm{p}(\mathbf{R}, r). \end{aligned} \tag{2.14}$$

The DPMM for cluster assignment places a truncated stick-breaking prior on the assignment probabilities to each cluster. The stick-breaking process and cluster assignment model are

$$\begin{aligned} V_1, \ldots, V_{C-1} | \alpha &\sim \mathrm{Beta}(1, \alpha),\ V_C = 1 \\ \alpha &\sim \mathrm{Gamma}(\alpha_\alpha, \beta_\alpha) \\ P(z_i = c) = \psi_c &= V_c \prod_{h=1}^{c-1} (1 - V_h) \\ z_i &\sim \mathrm{Categorical}(\boldsymbol{\psi}). \end{aligned} \tag{2.15}$$

Subject to a maximum of C clusters, the DPMM allows the number of non-empty clusters to be estimated from the data. To identify the most optimal partitioning of the data, we follow the approach described in Dahl (2006) and Molitor et al. (2010). First, we construct an $n \times n$ score matrix at each iteration with a 1 in the i, j location if individuals i and j belong to the same cluster and a 0 otherwise. Then we calculate a probability matrix $\mathbf{S}$ by averaging the score matrices. The most optimal clustering is the clustering from the MCMC iteration that has a score matrix with minimum least squared distance to the probability matrix $\mathbf{S}$. We calculate

model averaged estimates of the cluster-specific parameters θ_c to incorporate the uncertainty present in the best clustering (Molitor et al., 2010).

We complete the model with the following prior distributions. For clusters $c = 1, \ldots, C$, the prior distribution on the cluster intercepts is

$$\begin{aligned} \theta_c|\kappa_c^{-2} &\sim \text{N}(0, \kappa_c^2) \\ \kappa_c^{-2} &\sim \text{Gamma}(\alpha_\kappa, \beta_\kappa). \end{aligned}$$

The prior distributions on the covariate regression coefficients ($\gamma_l, l = 1, \ldots, q$), and error precision ($\sigma^{-2}$) are

$$\begin{aligned} \gamma_l|\phi_l^{-2} &\sim \text{N}(0, \phi_l^2) \qquad l = 1, \ldots q \\ \phi_l^{-2} &\sim \text{Gamma}(\alpha_\phi, \beta_\phi) \\ \sigma^{-2} &\sim \text{Gamma}(\alpha_\sigma, \beta_\sigma). \end{aligned}$$

We consider two variations of profile regression. The first, supervised profile regression (SPR), originally introduced by Molitor et al. (2010) belongs to the SDR class of methods since cluster assignments are influenced by the health outcome. The second is an unsupervised adaptation (UPR) not originally proposed by Molitor et al. (2010) that belongs to the UDR class. The difference between the two variations manifests when the latent cluster assignment variable z_i is updated. In the supervised case, we jointly model the response and cluster assignments. Feedback between the health outcome model and the profile assignment model allow the health outcomes to influence cluster assignment. In SPR, the full conditional for z_i depends on both the likelihood of exposures and the likelihood of the response and is calculated as

$$P(z_i = c|\mathbf{x}_i, y_i, \cdot) = \frac{\psi_c f(\mathbf{x}_i|z_i = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \sigma^2)}{\sum_{c=1}^{C} \psi_c f(\mathbf{x}_i|z_i = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \sigma^2)}. \tag{2.16}$$

Hence, individuals with similar exposure profiles may be assigned to different clusters if they have different health outcomes.

The unsupervised case involves a two-step procedure where we first estimate cluster assignments independently of the response and then model the response conditional on cluster assignment. In UPR, the full conditional for z_i depends only on the exposure likelihood and is calculated as

$$P(z_i = c|\mathbf{x}_i, \cdot) = \frac{\psi_c f(\mathbf{x}_i|z_i = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{c=1}^{C} \psi_c f(\mathbf{x}_i|z_i = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}. \tag{2.17}$$

Since the response does not inform cluster assignment in UPR, there may be high uncertainty in the estimates of the cluster indicators θ_c if individuals with similar exposure profiles have very different health outcomes but are assigned to the same cluster.

We implement a blocked Gibb's sampler for posterior computation. At each iteration of the MCMC sampler, we create the $n \times C$ binary matrix $\mathbf{Z}$ indicating to which cluster each individual exposure profile belongs. We can then write the response model as

$$\mathbf{y}|\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2 \sim \mathrm{N}\left(\mathbf{Z}\boldsymbol{\theta} + \mathbf{W}\boldsymbol{\gamma}, \sigma^2\mathbf{I}\right),$$

where $\mathbf{y}$ is the vector of health outcomes for individuals $i = 1, \ldots, n$, $\boldsymbol{\theta}$ is the $C \times 1$ vector of cluster-intercepts, $\mathbf{W}$ is the $n \times q$ matrix of covariate data, and $\boldsymbol{\gamma}$ is the $q \times 1$ matrix of covariate regression coefficients. We define the block parameters

$$\begin{aligned} \mathbf{A} &= (\mathbf{Z}, \mathbf{W}) \\ \boldsymbol{\delta}^T &= (\boldsymbol{\theta}, \boldsymbol{\gamma}) \\ \tau^{-2} &= \mathrm{diag}\left(\kappa_1^{-2}, \cdots, \kappa_C^{-2}, \phi_1^{-2}, \cdots, \phi_q^{-2}\right), \end{aligned}$$

where τ^{-2} is a $(C+q)$-dimensional diagonal matrix of precision parameters for the cluster intercepts and covariates. Additionally, n_c is the number of individual exposure profiles currently

allocated to cluster c and $\bar{\mathbf{x}}_c = (\frac{1}{n_c}\sum_{i=1}^{n_c} x_{i1}, \ldots, \frac{1}{n_c}\sum_{i=1}^{n_c} x_{ip})$ is the vector of empirical exposure means for individuals currently in cluster c. We sample cluster intercepts and covariate regression coefficients as a block from

$$\boldsymbol{\delta}|\cdot \sim N\left((\sigma^{-2}\mathbf{A}^T\mathbf{A}+\tau^{-2}\mathbf{I})^{-1}(\sigma^{-2}\mathbf{A}^T\mathbf{y}), (\sigma^{-2}\mathbf{A}^T\mathbf{A}+\tau^{-2}\mathbf{I})^{-1}\right).$$

For $c = 1, \ldots, C$, we sample the cluster-specific exposure means and variances from

$$\begin{aligned}
\boldsymbol{\mu}_c|\cdot &\sim N\left([n_c\boldsymbol{\Sigma}_c^{-1}+\boldsymbol{\Lambda}_0^{-1}]^{-1}[n_c\boldsymbol{\Sigma}_c^{-1}\bar{\mathbf{x}}_c+\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\nu}_0], [n_c\boldsymbol{\Sigma}_c^{-1}+\boldsymbol{\Lambda}_0^{-1}]^{-1}\right) \\
\boldsymbol{\Sigma}_c^{-1}|\cdot &\sim \text{Wish}_p\left([\mathbf{R}^{-1}+\sum_{i:z_i=c}(\mathbf{x}_i-\boldsymbol{\mu}_c)(\mathbf{x}_i-\boldsymbol{\mu}_c)^T]^{-1}, n_c+r\right).
\end{aligned}$$

We sample the stick-breaking process parameters from

$$\begin{aligned}
V_c|\cdot &\sim \text{Beta}\left(n_c+1, \alpha+n-\sum_{k=1}^{c} n_k\right), \quad c = 1, \ldots, C-1 \\
\alpha|\cdot &\sim \text{Gamma}\left(\alpha_\alpha+C-1, \beta_\alpha-\sum_{c=1}^{C-1}\log(1-V_c)\right).
\end{aligned}$$

We sample the precision parameters from

$$\begin{aligned}
\kappa_c^{-2}|\cdot &\sim \text{Gamma}\left(\alpha_\kappa+\frac{C}{2}, \beta_\kappa+\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}\right), \quad c = 1, \ldots, C \\
\phi_l^{-2}|\cdot &\sim \text{Gamma}\left(\alpha_\phi+\frac{q}{2}, \beta_\phi+\frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{\gamma}\right), \quad l = 1, \ldots, q \\
\sigma^{-2}|\cdot &\sim \text{Gamma}\left(\alpha_\sigma+\frac{n}{2}, \beta_\sigma+\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\delta})^T(\mathbf{y}-\mathbf{A}\boldsymbol{\delta})\right).
\end{aligned}$$

The model has been extended to include variable selection to identify mixture components actively contributing to cluster assignment (Chung and Dunson, 2009; Liverani et al., 2015; Papathomas et al., 2012). Briefly, binary random variables are introduced that indicate whether the mean for a mixture component within a cluster is unique to that cluster or common among all clusters. Hence, mixture components that are selected into the model are interpreted as being

informative in partitioning the exposure data into clusters, but are not necessarily related to the health outcome. For exposures $j = 1, \ldots, p$, the variable selection model is

$$\begin{aligned}
\mu^*_{c,j} &= \pi_{c,j}\mu_{c,j} + (1-\pi_{c,j})\bar{x}_j \\
\pi_{c,j}|\rho_j &\sim \text{Bernoulli}(\rho_j) \\
\rho_j|\omega_j &\sim I(\omega_j = 0)\delta_0(\rho_j) + I(\omega_j = 1)\text{Beta}(\alpha_\rho, \beta_\rho) \\
\omega_j &\sim \text{Bernoulli}(0.5),
\end{aligned}$$

where the binary variable $\pi_{c,j}$ determines if exposure j informs the clustering of exposures profiles into cluster c and $\bar{x}_j$ is the empirical mean for exposure j. If $\pi_{c,j} = 0$ then the empirical mean of exposure j replaces the cluster-specific mean for exposure j for all clusters; hence, exposure j does not inform the clustering. When variable selection is implemented, $\boldsymbol{\mu}^*_c = (\mu^*_{c,1}, \ldots, \mu^*_{c,p})$ replaces $\boldsymbol{\mu}_c$ for $c = 1, \ldots, C$ in the likelihood in (2.14) and in all parameter updates that depend on $\boldsymbol{\mu}_c$.

To sample the parameters for variable selection, we introduce the following notation: $n_{\{\pi_j=1\}}$ is the number of clusters in which the variable selection indicator variable $\pi_{j,c}$ for the j^{th} exposure is equal to 1, C^* is the number of non-empty clusters, $\sigma^2_{c,j}$ is the $(j,j)^{th}$ element of $\boldsymbol{\Sigma}_c$, λ_j is the j^{th} diagonal element of $\boldsymbol{\Lambda}$, ν_j is the j^{th} element of $\boldsymbol{\nu}$, and $\boldsymbol{\Pi}_c$ is a diagonal matrix with diagonal elements $\pi_{c,1}, \ldots, \pi_{c,p}$. We integrate over $\mu_{c,j}$ to update $\pi_{c,j}$ for $c = 1, \ldots, C$ and $j = 1, \ldots, p$.

We update the variable selection parameters by

$$
\begin{aligned}
P(\omega_j = 1|\sum_c \pi_{c,j} > 0, \cdot) &= 1 \\
P(\omega_j = 0|\sum_c \pi_{c,j} = 0, \cdot) &\propto P(\omega_j = 0) = 0.5 \\
P(\omega_j = 1|\sum_c \pi_{c,j} = 0, \cdot) &\propto \Big(\frac{\Gamma(\alpha_\rho + \beta_\rho)\Gamma(\beta_\rho + C^*)}{\Gamma(\beta_\rho)\Gamma(\alpha_\rho + \beta_\rho + C^*)}\Big)(0.5) \\
P(\rho_j = 0|\omega_j = 0, \cdot) &= 1 \\
\rho_j|\omega_j = 1, \cdot &\sim \text{Beta}(n_{\{\pi_j=1\}} + \alpha_\rho, C^* - n_{\{\pi_j=1\}} + \beta_\rho) \\
P(\pi_{c,j} = 1|\cdot) &\propto (2\pi)^{-\frac{n_c}{2}} (\sigma^2_{c,j})^{-\frac{n_c}{2}} \rho_j \lambda_j^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\Big[\frac{\sum_{i:z_i=c} x^2_{i,j}}{\sigma^2_{c,j}} + \frac{\nu_j^2}{\lambda_j}\Big]\right\} \times \\
&\quad \exp\left\{\frac{1}{2}\Big(\frac{n_c}{\sigma^2_{c,j}} + \frac{1}{\lambda_j}\Big)\left(\frac{\frac{\sum_{i:z_i=c} x_{i,j}}{\sigma^2_{c,j}} + \frac{\nu_j}{\lambda_j}}{\frac{n_c}{\sigma^2_{c,j}} + \frac{1}{\lambda_j}}\right)^2\right\}\left(\frac{n_c}{\sigma^2_{c,j}} + \frac{1}{\lambda_j}\right)^{-\frac{1}{2}} \\
P(\pi_{c,j} = 0|\cdot) &\propto \prod_{i:z_i=c} f(x_{ij}|\bar{x}_j, \sigma^2_{c,j})(1 - \rho_j) \\
\boldsymbol{\mu}^*_c|\cdot &\sim \text{N}\big(\tilde{\boldsymbol{\Sigma}}\left[\boldsymbol{\Lambda}^{-1}\boldsymbol{\nu} + n_c \boldsymbol{\Pi}_c \boldsymbol{\Sigma}_c^{-1}(\bar{\mathbf{x}}_c - (I - \boldsymbol{\Pi}_c)\bar{\mathbf{x}})\right], \tilde{\boldsymbol{\Sigma}}\big),
\end{aligned}
$$

where $\tilde{\boldsymbol{\Sigma}} = \left[n_c \boldsymbol{\Pi}_c \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\Pi}_c + \boldsymbol{\Lambda}^{-1}\right]^{-1}$.

We fit SPR using the `R` package `PReMiuM` (Liverani et al., 2015) and UPR using the `R` package `mmpack` (Hoskovec, 2019) developed for this paper.

2.3.3 Bayesian Kernel Machine Regression

Bayesian kernel machine regression (BKMR) (Bobb et al., 2015) belongs to the NP class of methods and flexibly models the exposure-response function to allow for nonlinear associations and higher order interactions. In BKMR, $h(\mathbf{x})$ is a smooth function represented using a Gaussian kernel. The response is modeled as

$$
\begin{aligned}
y_i &\sim N(h_i + \mathbf{w}_i^T \boldsymbol{\gamma}, \sigma^2) \\
\mathbf{h} \equiv (h_1, \ldots, h_n)^T &\sim N(\mathbf{0}, \tau \mathbf{K}),
\end{aligned}
\tag{2.18}
$$

where $\mathbf{K}$ is the kernel matrix with (i, i') element $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp\{-\sum_{j=1}^{p} r_j(x_{ij} - x_{i'j})^2\}$, τ is a hyperparameter, and $\mathbf{r} = \left(r_1, \ldots, r_p\right)^T$ are variable selection parameters. Estimated health outcomes for individuals with similar exposure levels across the p predictors are shrunk towards each other, resulting in a smooth but flexible exposure-response function.

BKMR allows for both component-wise and hierarchical variable selection (HVS) to identify important mixture components. For inference in our simulation study and data analysis, we implemented component-wise variable selection and calculated PIPs for each exposure. In our data analysis, we also implemented HVS to address sensitivity of results. We partitioned the mixture components into groups of air pollutants ($PM_{2.5}$, PM_{10}, NO_2, and O_3) and pesticides (C, MeBr, and OP) and calculated PIPs for each group (group PIPs) and each component within a group, conditional on group inclusion (conditional PIPs). We fit BKMR using the `R` package `bkmr` (Bobb, 2017). We refer to Bobb et al. (2015) and Bobb et al. (2018) for details on implementing BKMR.

2.3.4 Simulation Study Design

We evaluated the proposed methods in a simulation study. We ensured a realistic correlation structure among the pollutants by using the observed exposure data from 153 individuals in the FACES data set in our simulation study. We also used the observed covariate data in our simulation study. Health responses were simulated for three exposure-response scenarios, denoted h_k, $k = 1, 2, 3$, as $y_i = h_k(\mathbf{x}_i) + \mathbf{w}_i^T \boldsymbol{\gamma} + \varepsilon_i$, with $\varepsilon_i \sim \mathrm{N}(0, 1)$. The covariate coefficients $\gamma_1, \ldots, \gamma_q$ were simulated as independent $\mathrm{N}(0, 1)$.

The first scenario, h_1 (linear), is an EMM model. For exposures x_j, $j = 1, \ldots, 4$, the exposure-response function is

$$h_1(\mathbf{x}) = x_1 - x_2 + x_3 - x_4 + 0.7x_1x_2 - 0.5x_3x_4. \tag{2.19}$$

Second, h_2 (nonlinear) includes nonlinear sigmoidal functions of three pollutants and a multiplicative interaction between two of those pollutants:

$$h_2(\mathbf{x}) = \frac{2}{1+\exp(-3x_1)} + \frac{2}{1+\exp(-5x_2)} - \frac{2}{1+\exp(-5x_3)} - 0.4x_1x_2. \tag{2.20}$$

Last, h_3 (fixed profiles) groups individuals into four distinct clusters based on dichotomous cut-offs for two pollutants. We assign a constant health effect to individuals in the same cluster:

$$h_3(\mathbf{x}) = \begin{cases} -2, & x_1 \leq \text{median}(x_1) \text{ and } x_2 \leq \text{median}(x_2) \\ -1, & x_1 \leq \text{median}(x_1) \text{ and } x_2 > \text{median}(x_2) \\ 0, & x_1 > \text{median}(x_1) \text{ and } x_2 \leq \text{median}(x_2) \\ 2, & x_1 > \text{median}(x_1) \text{ and } x_2 > \text{median}(x_2). \end{cases} \tag{2.21}$$

We selected these three exposure-response scenarios to cater to different methods in our simulation study. The linear scenario plays to NPBr and NPB, the nonlinear scenario plays to BKMR, and the fixed profiles scenario plays to UPR and SPR. We hypothesize that the methods to which each scenario caters will perform best in that scenario. We are interested in evaluating how methods perform in exposure-response scenarios for which they were not explicitly developed.

We simulated 200 data sets for each scenario and fit all five Bayesian methods plus LM and LM-int. As results can be sensitive to which pollutants, x_j, $j = 1,\ldots,4$, are included in $h(\mathbf{x})$, we randomly selected pollutants to be the active components in each simulated data set. All pollutants, even those not selected as one of the active components, are included as inputs in the estimated models. By randomly selecting which exposures are the active components of the mixture, each simulated data set has a different correlation structure among the active exposures, which adds robustness to our simulation study results.

To measure the grouping structure of the data generated by the fixed profiles scenario, we calculated the Calinski-Harabasz index (Caliński and Harabasz, 1974), the silhouette statistic

(Rousseeuw, 1987), and the number of clusters to maximize the gap width (Tibshirani et al., 2001). Since the active mixture components differ for each simulated data set, the clustering of the data in the fixed profiles scenario differs for each data set. Across the 200 data sets used in our simulation study, the median Calinski-Harabasz index was 22.54, the median silhouette was 0.15, and the median number of clusters to maximize the gap width was 6. The distribution of each of these statistics can be found in Appendix A.3 (Table A.2). In general, the fixed profiles scenario did not always generate a strong grouping structure with this data, but instead represents a wide variety of clustering schemes.

We evaluated exposure-response function estimation using root mean squared error (RMSE) and interval coverage (Cvg). RMSE was calculated as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[h(\mathbf{x}_i)-\hat{h}(\mathbf{x}_i)\right]^2}$ and coverage was calculated as the percent of $h(\mathbf{x}_i)$'s covered by 95% credible or confidence intervals. RMSE measures the variation between estimated and true values of the exposure-response function. Coverage measures how often the 95% credible or confidence intervals for the estimated exposure-response function capture the true mixture effect. A method with high RMSE and low coverage fails to capture the overall mixture effect. In this way, RMSE and coverage measure the ability of each method to capture the overall mixture effect.

We summarized variable selection through true and false selection rates. In the Bayesian methods, we consider a variable with a PIP above 0.5 as selected into the model (Barbieri and Berger, 2004). In LM and LM-int, a variable is selected if the 95% confidence interval for the respective regression coefficient does not contain 0. We calculated true selection rate (TSR) as the proportion of mixture components active in the exposure-response function as main effects that were selected into the model as main effects, and false selection rate (FSR) as the proportion of mixture components not in the exposure-response function as main effects that were selected into the model as main effects. All seven exposures are included in the models as inputs, but the active mixture components are those that define the exposure-response function for each simulated data set. For scenario 1, the active main effects are the randomly selected exposures denoted by x_1, x_2, x_3, and x_4; for scenario 2, the active main effects are x_1, x_2, and

x_3; and for scenario 3 the active main effects are x_1 and x_2. In most methods (NPBr, UPR, SPR, BKMR, and LM), TSR and FSR are calculated only for main effects. In NPB we can calculate PIPs for interactions and in LM-int we can calculate confidence intervals for the interaction effects. Hence, we also evaluate variable selection rates for interactions in NPB and LM-int. We calculate true selection rate for interactions (TSR_{int}) as the proportion of the exact pairwise interactions active in the exposure response function that were selected into the model as interactions, and false selection rate for interactions (FSR_{int}) as the proportion of interactions that were not active in the exposure-response function that were selected into the model as interactions. In scenario 1, the true active interactions are x_1x_2 and x_3x_4 and in scenarios 2 and 3 the only active interaction is x_1x_2.

We assessed convergence for a few simulated data sets by visualizing trace plots and comparing results from multiple chains. We found evidence of convergence by 20,000 iterations for all methods. To ensure convergence across all simulated data sets, we based inference on 25,000 samples after a burn-in of 25,000 samples.

We conducted three additional simulation studies to further assess method performance. First, we considered a null scenario, $h_4(\mathbf{x})$, where none of the exposures are associated with the response. Second, we considered a complex mixture scenario, $h_5(\mathbf{x})$, where we simulated data for seven additional pollutants to have a total of 14 mixture components. Third, we applied our original simulation study design to a larger sample of size $n = 1000$ for each data set. Details on the null, complex mixture, and large sample size simulations can be found in Appendix A.4.

2.3.5 Data Analysis

We conducted a data analysis on 153 individuals with complete data in the FACES data set. We used regression-adjusted FEV_1 as the outcome. Pesticide and air pollutant exposures and covariate data were identical to that in our simulation study (Tables 2.1, 2.2, and A.1). We fit the same models as in the simulation study. Prior specifications for the Bayesian models are listed in Appendix A.2.

2.4 Results

2.4.1 Simulation Study Results

Simulation results are shown in Table 2.4. Standard errors are shown in Appendix A.3 (Tables A.3, A.4, and A.5). We show the computational time for each method to run for 5000 iterations in Table 2.3.

Table 2.3: Computational time for each method to run 5000 iterations on MacBook Pro in R version 3.6.1. Time is reported in seconds. Results reflect 10 evaluations of each method.

method	minimum	mean	maximum
NPBr	6.90	7.03	7.17
NPB	24.73	24.95	25.23
BKMR	219.43	222.96	235.41
UPR	57.82	58.66	59.50
SPR	90.34	92.47	98.65

Overall BKMR and NPB were the best performing methods with BKMR performing slightly better in the nonlinear and fixed profiles scenarios. Regarding RMSE for the exposure-response function, BKMR (RMSE = 0.55) and NPB (RMSE = 0.54) tied for lowest in the linear scenario. In the nonlinear scenario, BKMR (RMSE = 0.59) pulled slightly ahead of NPB (RMSE = 0.69), while in the fixed profiles scenario, BKMR (RMSE = 0.69) outperformed all other methods by a substantial margin. UPR had the highest RMSE in all three scenarios with SPR having the second highest RMSE.

In addition to having the lowest RMSE in all three scenarios, BKMR consistently had interval coverage closest to the nominal level. LM-int also had interval coverage near the nominal level in all three scenarios and NPB performed well in the linear scenario. BKMR (Cvg = 0.96), NPB (Cvg = 0.95), and LM-int (Cvg = 0.95) all achieved the nominal coverage level (0.95) in the linear scenario. In the nonlinear scenario, BKMR (Cvg = 0.92) and LM-int (Cvg = 0.91) came closest to the nominal level, with NPB next best but trailing behind (Cvg = 0.86). BKMR (Cvg = 0.91) and

LM-int (Cvg = 0.91) had the highest coverage by far in the fixed profiles scenario. Again, UPR and SPR performed poorly with the lowest coverage in all three scenarios.

The story is more complex when it comes to variable selection. While BKMR had the highest TSR in all three scenarios, it also had the highest FSR. Again, NPB performed very well in the linear scenario but not as well in the other scenarios, while UPR and SPR had consistently poor selection rates. Regarding TSR, BKMR (TSR = 1.00) and NPB (TSR = 0.92) performed best in the linear scenario. BKMR had the highest TSR in the nonlinear scenario (TSR = 0.96), where the next best methods, NPBr, NPB, and LM, all had mean TSR just under 0.80. BKMR is singled out with the best TSR in the fixed profiles scenario (TSR = 0.97). UPR, SPR, and LM-int tended to have low TSR in all three scenarios.

A low false selection rate indicates a model does not erroneously classify exposures as associated with the outcome when they are not. Here, BKMR had some of the highest FSR across the three scenarios. In the linear scenario, LM-int (FSR = 0.04) and NPB (FSR = 0.10) had the lowest FSR. LM-int also had the lowest FSR in the nonlinear scenario (FSR = 0.08). In the fixed profiles scenario, NPBr, NPB, LM, LM-int all had similar FSR at or below 0.14. Along with BKMR, SPR had high FSR in all three scenarios.

When considering overall variable selection performance, NPB takes the top spot in the linear scenario, with high TSR and low FSR. No method was able to simultaneously achieve dominant TSR and FSR in the nonlinear or fixed profiles scenarios.

Only NPB and LM-int directly parameterized variable selection for interactions in an easily interpretable manner. Interpretable variable selection for interactions is itself an advantage of these approaches over the other methods. In the linear scenario, NPB (TSR_{int} = 0.59) had higher TSR_{int} than LM-int (TSR_{int} = 0.32). Both methods had poor TSR_{int} in the nonlinear and fixed profiles scenarios, with values at or below 0.25. Regarding FSR_{int}, both methods performed well in all three scenarios, with FSR_{int} consistently at or below 0.11.

The additional simulations (Appendix A.4) produced similar results, with NPB and BKMR being consistently top-performing methods in terms of estimating the exposure-response func-

Table 2.4: Summary of method performance in three data-generating scenarios. Table shows means across all data sets for: root mean squared error (RMSE), coverage (Cvg), true selection rate for main effects (TSR), false selection rate for main effects (FSR), true selection rate for interactions (TSR_{int}), and false selection rate for interactions (FSR_{int}). Top-performing methods will have low RSME, coverage near the nominal level (0.95), high TSR and low FSR. For each measure and exposure-response scenario, results from the top-performing method(s) are listed in bold.

Method	RMSE	Cvg	TSR	FSR	TSR_{int}	FSR_{int}
$h_1(\mathbf{x})$: linear with multiplicative interactions						
NPBr	1.02	0.73	0.85	0.35	–	–
NPB	**0.54**	**0.95**	0.92	0.10	**0.59**	**0.02**
UPR	2.01	0.56	0.25	0.26	–	–
SPR	1.59	0.54	0.63	0.53	–	–
BKMR	**0.55**	**0.96**	**1.00**	0.39	–	–
LM	1.01	0.73	0.84	0.29	–	–
LM-int	0.73	**0.95**	0.68	**0.04**	0.32	0.04
$h_2(\mathbf{x})$: nonlinear with multiplicative interactions						
NPBr	0.77	0.80	0.79	0.22	–	–
NPB	0.69	0.86	0.78	0.16	**0.25**	**0.01**
UPR	1.42	0.56	0.27	0.24	–	–
SPR	1.27	0.58	0.68	0.58	–	–
BKMR	**0.59**	**0.92**	**0.96**	0.48	–	–
LM	0.78	0.81	0.78	0.17	–	–
LM-int	0.89	**0.91**	0.54	**0.08**	0.20	0.07
$h_3(\mathbf{x})$: constant function of fixed profiles						
NPBr	1.11	0.66	0.66	**0.11**	–	–
NPB	1.02	0.75	0.68	0.13	0.06	**0.02**
UPR	1.41	0.55	0.27	0.25	–	–
SPR	1.38	0.54	0.68	0.59	–	–
BKMR	**0.69**	**0.91**	**0.97**	0.64	–	–
LM	1.13	0.70	0.69	0.14	–	–
LM-int	0.99	**0.91**	0.56	0.14	**0.12**	0.11

tion and identifying active mixture components. In the null scenario (Table A.6), NPBr and NPB had lowest FSR, meaning these methods were the best at not selecting any mixture components into the model when none were associated with the response. Results from the complex mixture scenario (Table A.7) largely mirrored those from the linear scenario. In the large sample size simulation (Table A.8), BKMR and NPB remained top-performing. TSR was high for all methods. UPR and SPR also had high FSR, meaning they often selected all of the mixture components into the model.

2.4.2 Data Analysis Results

The results from our analysis of the FACES data set varied across the methods. First we consider the traditional models LM and LM-int. LM showed evidence for main effects of NO_2 ($\hat{\beta}$: -0.32, CI: -0.54, -0.10) and PM_{10} ($\hat{\beta}$: 0.19, CI: 0.02, 0.35). LM-int showed evidence for main effects of MeBr ($\hat{\beta}$: 0.17, CI: 0.05, 0.29), NO_2 ($\hat{\beta}$: -0.68, CI: -1.10, -0.25), and PM_{10} ($\hat{\beta}$: 0.50, CI: 0.08, 0.93) and an interaction between C and $PM_{2.5}$ ($\hat{\beta}$: 0.28, CI: 0.01, 0.54) (Table 2.5). The results from the linear models indicating a protective effect of PM_{10} are counter-intuitive as there is an extensive literature on the deleterious health effects of PM on lung function. None of the other methods found evidence of protective effects for any of the exposures.

Next we consider the five contemporary methods. NPBr did not identify any exposures with PIPs above 0.5. The exposure with the highest PIP was NO_2 (PIP = 0.47), which was estimated to be negatively associated with FEV_1 ($\hat{\beta}$: -.08, CI: -0.35, 0.00). In NPB, NO_2 was selected (PIP = 0.60) and was also negatively associated with FEV_1 ($\hat{\beta}$: -0.12, CI: -0.36, 0.00) (Table 2.6). No other main effects or interactions were selected by either method (Table A.10).

In BKMR, NO_2 was selected as an important mixture component with a PIP of 0.96 (Table A.11). No other exposures had PIPs above 0.5. Results were similar using the HVS formulation (Table A.12). NO_2 had a negative and nonlinear association with FEV_1 (Figure 2.1). To identify interactions, we plot the posterior distribution of the exposure-response function for each pair of exposures, holding all other exposures constant at their median values, and visually inspect

Table 2.5: Results from analysis of FACES data set using LM and LM-int. Table includes effect estimates ($\hat{\beta}$), 95% confidence intervals, and associated p-values for all main effects in LM and LM-int plus the interaction effects in LM-int with p-values ≤ 0.10. The regression coefficient $\hat{\beta}$ is the expected change in FEV_1 for a 1 standard deviation increase in the square root transformed exposures.

	LM			LM-int		
	$\hat{\beta}$	95% CI	p-value	$\hat{\beta}$	95% CI	p-value
C	0.04	(-0.03 , 0.11)	0.24	0.05	(-0.08 , 0.19)	0.44
MeBr	0.00	(-0.06 , 0.07)	0.96	0.17	(0.05 , 0.29)	0.01
OP	0.05	(-0.03 , 0.13)	0.24	0.02	(-0.17 , 0.22)	0.80
O_3	-0.06	(-0.20 , 0.07)	0.36	-0.13	(-0.32 , 0.06)	0.17
NO_2	-0.32	(-0.54 , -0.10)	0.01	-0.68	(-1.10 , -0.25)	0.00
$PM_{2.5}$	-0.01	(-0.20 , 0.17)	0.90	-0.11	(-0.48 , 0.26)	0.55
PM_{10}	0.19	(0.02 , 0.35)	0.03	0.50	(0.08 , 0.93)	0.02
C:$PM_{2.5}$	–	–	–	0.28	(0.01 , 0.54)	0.04
OP:PM_{10}	–	–	–	0.31	(-0.01 , 0.62)	0.05
NO_2:PM_{10}	–	–	–	0.33	(-0.05 , 0.72)	0.09

Table 2.6: Results from analysis of FACES data set using NPBr and NPB. Table shows estimates ($\hat{\beta}$), 95% credible intervals, and posterior inclusion probabilities (PIP) for main effect exposures in NPB and NPBr. The regression coefficient $\hat{\beta}$ is the expected change in FEV_1 for a 1 standard deviation increase in the square root transformed exposures. All interaction effects in NPB had posterior inclusion probabilities below 0.12.

	NPBr			NPB		
	$\hat{\beta}$	95% CI	PIP	$\hat{\beta}$	95% CI	PIP
C	0.00	(0.00 , 0.04)	0.07	0.00	(0.00 , 0.03)	0.07
MeBr	0.00	(-0.02 , 0.00)	0.06	0.00	(-0.01 , 0.00)	0.06
OP	0.02	(0.00 , 0.12)	0.21	0.01	(0.00 , 0.11)	0.16
O_3	0.00	(-0.08 , 0.02)	0.11	-0.01	(-0.12 , 0.01)	0.11
NO_2	-0.08	(-0.35 , 0.00)	0.47	-0.12	(-0.36 , 0.00)	0.60
$PM_{2.5}$	0.00	(-0.08 , 0.06)	0.13	0.00	(-0.09 , 0.05)	0.12
PM_{10}	0.02	(0.00 , 0.21)	0.21	0.02	(-0.01 , 0.20)	0.19

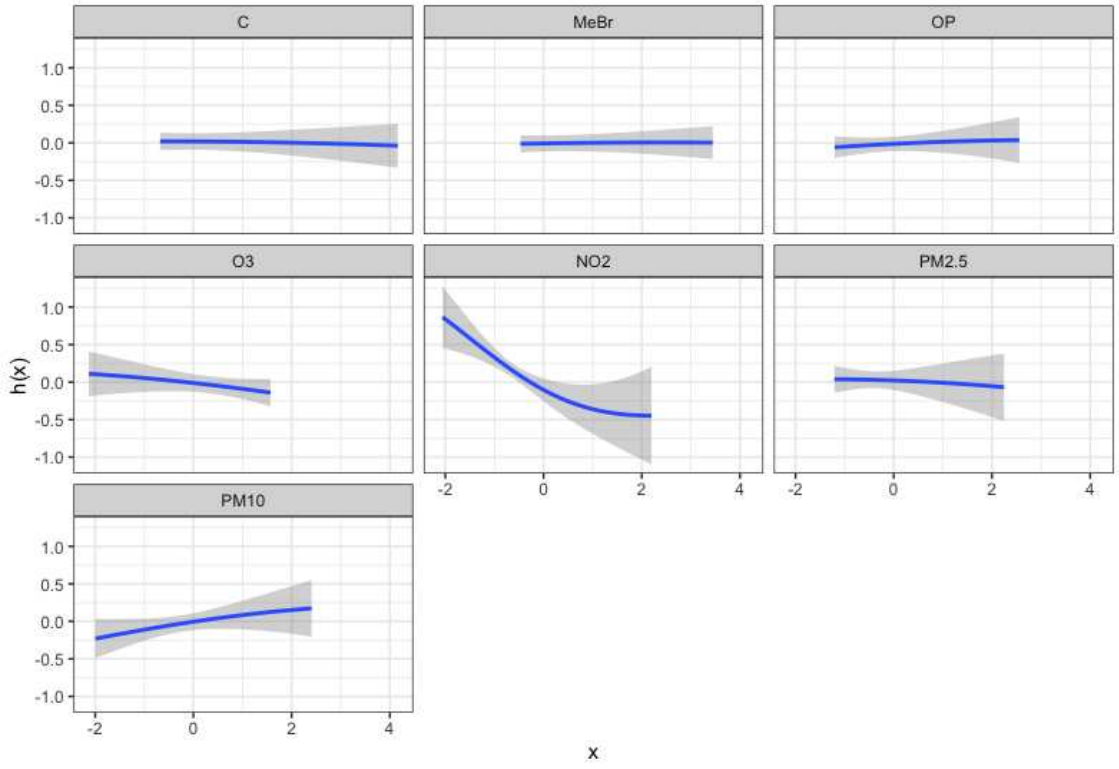


Figure 2.1: Results from analysis of FACES data set using BKMR. Figure shows the univariate relationship between each exposure and FEV_1, holding all other exposures at their median value.

changes in the response as both exposures change. In doing so we found no notable interactions among exposures (Figure A.1).

As clustering algorithms, UPR and SPR reveal a different kind of story. UPR revealed four clusters as the best partitioning of the data. Each cluster had similar estimated health effects (Figure 2.2a); hence, despite partitioning the exposure space there was no meaningful association between the exposure profiles and the health outcome. Figure 2.2b-e shows the empirical exposure means for individuals assigned to each cluster. The first cluster of n = 25 individuals was distinguished by higher than average exposure to MeBr. Cluster 2 (n = 33) had low exposure to OP and O_3 and high exposure to NO_2 and $PM_{2.5}$ relative to the average. The third cluster (n = 9) was characterized by relatively high exposure to OP and low exposure to O_3. Individuals in cluster 4 (n = 86) had nearly average exposure to most pollutants except MeBR, which was notably below average; in addition, O_3 exposure was slightly above and $PM_{2.5}$ exposure was

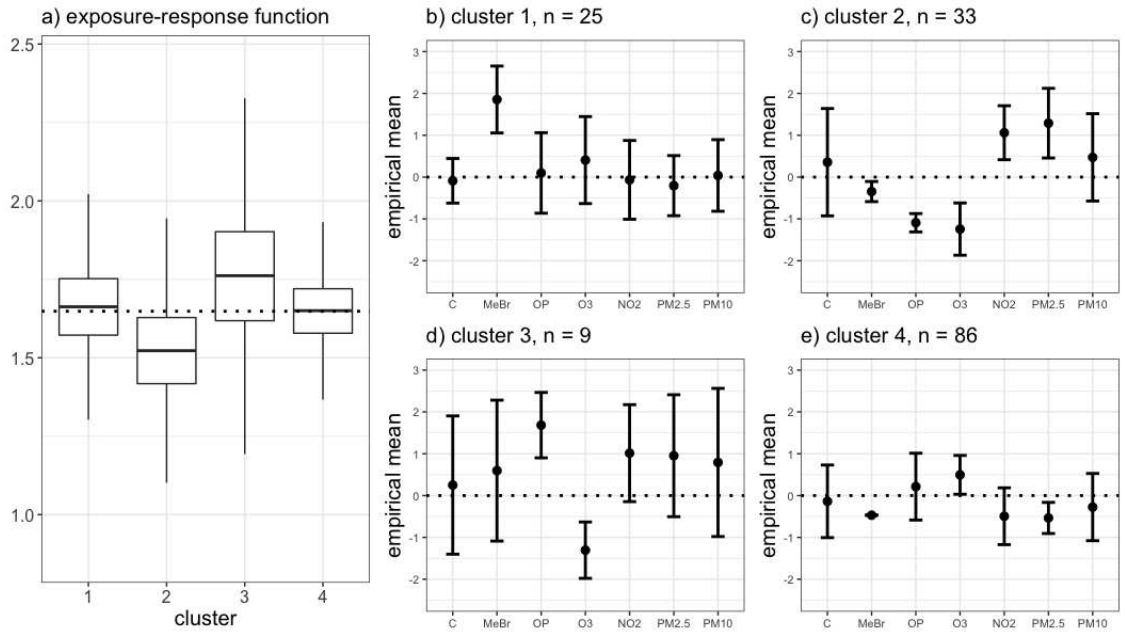


Figure 2.2: Results from analysis of FACES data set using UPR. Panel (a) shows the distribution of the model averaged estimated exposure-response function (θ_c) for each cluster identified in the best clustering by UPR. The dotted line represents the overall mean estimated exposure-response function across all clusters. Panels (b-e) show the empirical exposure means of the individuals assigned to each cluster in the best clustering, with 1 standard deviation error bars. The dotted lines are drawn at 0, the mean of the standardized exposure data.

slightly below average. UPR selected OP (PIP = 0.57), O_3 (PIP = 0.54), NO_2 (PIP = 0.61), and $PM_{2.5}$ (PIP = 0.56) as important mixture components (Table A.13).

SPR also revealed four clusters as the best partitioning of the data. The estimated exposure-response function for cluster 3, the smallest cluster ($n = 9$), had a 0.97 posterior probability of being greater than the overall mean estimated exposure-response function (Figure 2.3a). The cluster sample sizes and associated empirical exposure means were very similar to those in UPR (Figure 2.3b-e), with the labels switched for clusters 1 and 4. In both UPR and SPR, cluster 3 was the smallest cluster and had an estimated mean health effect higher than average, but there was more uncertainty around the health effect in UPR likely due to the two-stage approach for estimation. SPR selected five important mixture components: MeBr (PIP = 0.71), OP (PIP = 0.51), O_3 (PIP = 0.75), NO_2 (PIP = 0.67), and $PM_{2.5}$ (PIP = 0.63) (Table A.13). We found the

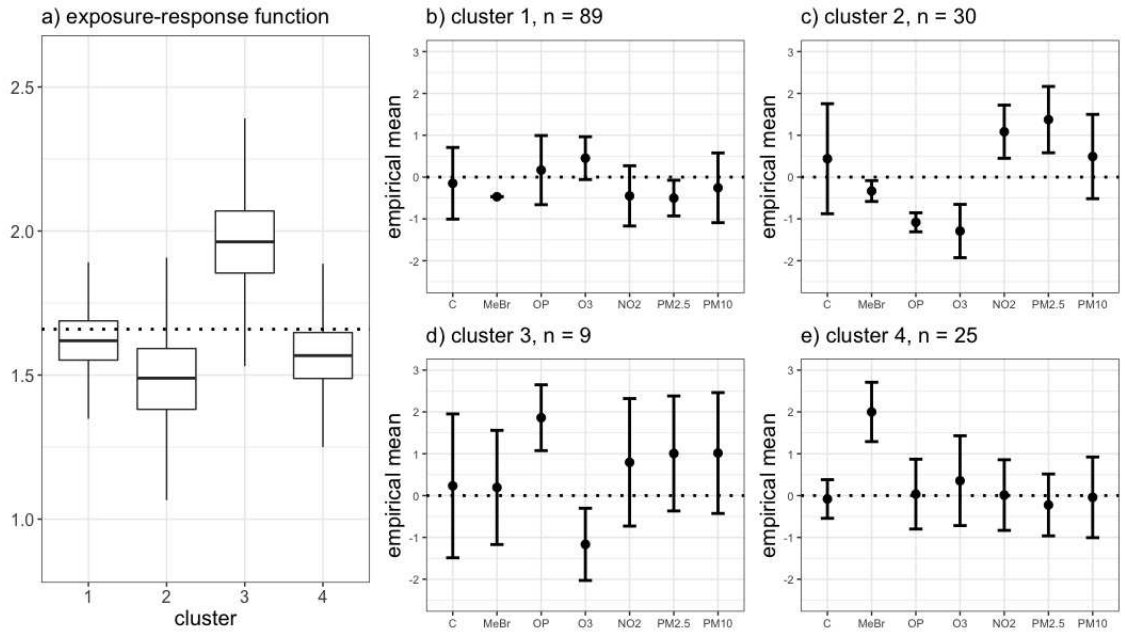


Figure 2.3: Results from analysis of FACES data set using SPR. Panel (a) shows the distribution of the model averaged estimated exposure-response function (θ_c) for each cluster identified in the best clustering by SPR. The dotted line represents the overall mean estimated exposure-response function across all clusters. Panels (b-e) show the empirical exposure means of the individuals assigned to each cluster in the best clustering, with 1 standard deviation error bars. The dotted lines are drawn at 0, the mean of the standardized exposure data.

clustering and PIPs in UPR and SPR to be sensitive to prior choice particularly for the cluster-specific precision matrix and error precision.

2.5 Discussion

Interest is rapidly growing in estimating the association between exposure to mixtures of environmental chemicals and health outcomes. As a result, new statistical approaches have been developed for studying health outcomes associated with exposure to mixtures. The purpose of this paper was to evaluate and compare recently developed methods for mixtures and determine which research questions they answer well and in which scenarios. We limited our study to contemporary Bayesian methods since they are under-studied, under-utilized, and may have the ability to answer multiple research questions. Our results highlight the advantages of the flexible modeling and Bayesian framework of BKMR and NPB in estimating the

exposure-response function precisely and identifying mixture components most strongly associated with the health outcome.

Overall, BKMR was a top-performing method. In each of the scenarios, BKMR estimated the exposure-response function with coverage closest to the nominal level (0.95) and lowest RMSE. Despite being a more flexible approach based on Gaussian processes, BKMR had lower RMSE in the linear scenario than NPBr, LM, and LM-int, all of which assume linearity. This is likely because NPBr and LM do not account for interactions and LM-int can result in inflated standard errors in the presence of correlated data. BKMR identified active mixture components with the greatest frequency, but also included inactive components more often than other methods. Although we did not evaluate variable selection rates for interactions in BKMR in our simulation, BKMR can identify linear or nonlinear interactions among exposures through visualization or summarizing the posterior distribution of the exposure-response function. A drawback to BKMR is that results are not as easily interpreted as in NPB or the linear models, though there are currently efforts to enhance interpretation and a suite of visualization approaches that aid in interpretation. BKMR is an appealing choice for mixtures because it makes minimal assumptions on the shape of the exposure-response function and includes a sophisticated variable selection algorithm for identifying important mixture components.

NPB was top-performing in the linear scenario regarding estimating the exposure-response function, identifying both active and inactive mixture components, and identifying interactions. NPB performed well even when the exposure-response function was mildly nonlinear, but lacks the flexibility of BKMR for the fixed profiles scenario, which is highly nonlinear. The AME method NPBr poorly estimated the exposure-response function in the linear scenario, likely from not accounting for interactions. An advantage of NPB is its ease of interpretation, which is similar to interpreting a linear regression model. NPB estimates PIPs and effect sizes for all main effect and interaction terms, providing precise information regarding the contribution of each exposure to the mixture and its effect on the health outcome.

The profile regression methods, UPR and SPR, poorly addressed the research questions of interest in all three scenarios. Two explanations for this include lack of a clustering structure in the exposure data and a weak signal, both of which inhibit these methods from accurately estimating the multipollutant exposure-response function. Further, UPR and SPR do not have the ability to identify or estimate interactions or tease out individual effects of the pollutants within a mixture. These methods may not be appropriate for studies in which the primary objectives are to estimate the multipollutant exposure-response function and identify driving mixture components. As clustering methods, UPR and SPR are likely to perform better on data that has a strong grouping structure. Since we used a single data set in our simulation study, the results of our simulation should not be interpreted as representative of performance on all data structures. A particular advantage of UPR and SPR is that the number of clusters need not be pre-specified.

The linear model with interactions, LM-int, had coverage above 0.91 in all three scenarios, but had higher RMSE and lower TSR than BKMR and NPB. LM-int and NPB are both EMM methods, and NPB outperformed LM-int in the linear EMM scenario. LM and LM-int have the advantage of being easy to implement and interpret, but these methods estimated the exposure-response function with more uncertainty than the top-performing methods and generally lacked the ability to select truly active mixture components, likely due to high correlation among exposures.

In our application to the FACES data set, four methods (LM, LM-int, NPB, and BKMR) identified NO_2 as an important mixture component negatively associated with the health outcome. In addition, LM and LM-int estimated PM_{10} to have a positive association with FEV_1, and PM_{10} was positively correlated with NO_2. Further, the magnitude of the effect estimate for NO_2 in LM and LM-int was several times larger than that estimated in NPB, and the confidence intervals were also larger, reflecting more uncertainty. UPR and SPR also identified NO_2 as an important mixture component, but we cannot determine the sign of effect using these methods. Instead, UPR and SPR have the ability to estimate how the overall mixture is associated with the health

outcome. UPR revealed four clusters with similar estimated health effects; hence, patterns in the exposure data were not associated with FEV_1. In SPR, the smallest cluster was associated with higher average FEV_1 than the other clusters, suggesting an association between a relatively rare mixture of exposures and the health outcome. Alternatively, this small cluster may reflect a strong influence from the health outcome in the clustering using a supervised learner. Meanwhile, BKMR was able to describe a nonlinear association between NO_2 and FEV_1.

Using missing indicators may have introduced some bias in the effect estimates. Additionally, all Bayesian methods are sensitive to prior specification and results may vary with more or less informative priors. PIPs are particularly sensitive to prior specification in all methods, so changing prior hyperparameters may lead to changes in TSR and FSR. We implemented all models using the default priors as specified by the authors to obtain an objective comparison of these methods.

Along with the primary research question, the best performing method is likely to depend on the exposure data. We used observed exposure data so our results are highly relevant to realistic settings. Our simulation results can be generalized to small data sets with a limited number of localized exposures, which is a frequent scenario in epidemiological studies.

In analyses of environmental mixtures and human health, model choice depends on the assumed exposure-response relationship and the primary questions of interest. NPB and BKMR are recently proposed methods that outperformed traditional regression models and offer promising tools for mixtures analyses. We recommend NPB when the exposure-response function is assumed to be approximately linear and a primary goal is accurately identifying which are the active and inactive components of the mixture. NPB is also highly interpretable and explicitly tests for interactions. We recommend BKMR if the exposure-response function is assumed to take on a complex form and the primary goal is estimating the form of the exposure-response function while at the same time identifying important mixture components. Our results suggest that UPR and SPR do not reliably answer our specified research questions, but may be applicable for different research questions such as pattern recognition. We further encourage users

to take advantage of our R package `mmpack` (Hoskovec, 2019) to replicate the simulation and determine how each method performs on their own data. Results will likely be different on different data sets. In particular, the profile regression methods may perform better on data that exhibit a stronger clustering structure in the fixed profiles scenario. We include the clustering statistics as part of the summary of the fixed profile scenario output so users can see how much grouping structure is in their own data. Replicating the simulation on their own data will enable users to choose the best method for their data and specific research question.

Chapter 3

Infinite Hidden Markov Models for Multiple Multivariate Time Series with Missing Data

3.1 Introduction

Exposure to indoor and outdoor air pollution are leading environmental risk factors for morbidity and mortality worldwide (Global Burden of Diseases 2019 Risk Factors Collaborators, 2020). Recent technological advances allow personal monitors to be used to collect time-resolved ambient pollutant exposure data at the individual level. As opposed to collecting data from local air quality monitoring sites, personal monitoring results in more accurate assessments of exposure to air pollutants because these monitors move with an individual through various indoor and outdoor microenvironments such as home, work, and transit. Along with the advantages, time-resolved personal exposure data also evoke several modeling challenges, including strong temporal dependence, missing observations, and exposure values below the monitoring device's limit of detection (LOD).

Our work is motivated by the Fort Collins Commuter Study (FCCS). The FCCS assessed personal exposure to ambient air pollutants during normal workdays in Fort Collins, Colorado, USA (Good et al., 2016; Koehler et al., 2019). Exposures were assessed for multiple people on different days, creating multiple asynchronous multivariate time series. Shared patterns in movement and exposures exist due to locality and repeated sampling days for the same individual, and may be informed by covariates collected during the study such as time of day or individual-level factors. As is typical in personal exposure monitoring studies, some exposure data were missing due to device malfunction, participant noncompliance, or values too low to be detected by the monitoring device.

Several model-based approaches have been proposed to impute missing air pollution data observed on a daily time scale or at larger temporal resolutions. Hopke et al. (2001) and Krall et al. (2015) proposed imputation approaches based on Bayesian multivariate normal models. Hopke et al. (2001) accounted for time series structure with smoothly-varying means through an integrated moving average, but both models assume a constant variance over time. Houseman and Virji (2017) proposed an imputation model that uses splines to account for temporal trends, but this model breaks down with high autocorrelations. No models have been proposed to impute missing multivariate exposure data observed from personal monitors that account for rapid changes in the exposure distribution as people transition between environments (e.g. indoors to outdoors).

We conceptualize environments and activities as unobserved, or latent, discrete states through which individuals transition over time, with each state giving rise to a unique distribution of multivariate exposure data. To model the complexity of these data, we propose an infinite hidden Markov model (iHMM) framework (Beal and Rasmussen, 2002). Unlike traditional hidden Markov models (HMMs) (Rabiner and Juang, 1986), iHMMs allow for a countably infinite number of hidden states in the model by leveraging Bayesian nonparametric prior formulations, such as the Dirichlet process and extensions thereof (Beal and Rasmussen, 2002; Fox et al., 2011; Montañez et al., 2015; Teh et al., 2006), the beta process (Fox et al., 2014), and the probit stick-breaking process (PSBP) (Rodríguez and Dunson, 2011; Sarkar et al., 2012). A natural extension of these models is to modify transition probabilities based on available covariate information (Altman, 2007; Sarkar et al., 2012) or to incorporate application-specific prior beliefs, for example an increased propensity of lingering in a given state (Fox et al., 2011; Hensley and Djuric, 2017; Montañez et al., 2015). While HMMs are often developed to handle multiple time series (Altman, 2007; Dias et al., 2015; Langrock et al., 2013), iHMMs are typically not designed for this setting, with few exceptions (Fox et al., 2014). Notably, we are unaware of any iHMM methods that allow for multiple multivariate time series that are covariate-dependent. Further, there are

no existing iHMMs that can impute both data that are missing at random (MAR) and below the LOD.

In this manuscript, we develop a covariate-dependent iHMM for the analysis of multiple multivariate time series with missing data. We model the hidden state transition distribution with a covariate-dependent PSBP to inform transitions and identify shared patterns among multiple time series. By developing a fully Bayesian computational approach, we handle multiple imputation naturally by sampling from the posterior predictive distribution of the missing data conditional on the observed data and the estimated hidden states. Our primary inferential goals are to impute missing observations and identify a hidden state structure representing time-activity patterns associated with personal exposures.

3.2 Fort Collins Commuter Study

The FCCS followed 45 individuals for between 1 and 13 non-consecutive days each and measured their exposure to fine particulate matter ($PM_{2.5}$) mass ($\mu g/m^3$), carbon monoxide (CO) (parts per million), and black carbon (BC) ($\mu g/m^3$) at 10-second resolution for 24-hour periods. Using GPS data and time-activity diaries, each time point was manually classified into one of five microenvironments: home, work, transit, eateries, and other. The FCCS aimed to identify patterns in exposure to multiple pollutants that were associated with microenvironments.

We considered a subset of the FCCS data. Specifically, we considered only those individuals who had at least 5 repeated sampling days with less than 10% total missing observations on each day. This resulted in 50 sampling days including 9 individuals. We averaged the data to 5-minute intervals. If the 5-minute interval contained at least 90% observed data, then the exposure value for that interval was considered observed and calculated as the mean of the observed data within the interval. Observed data were log-transformed and scaled so each exposure had mean 0 and variance 1. Otherwise, the exposure value was considered missing and either denoted MAR or below the LOD based on the mode of the missing data type within the

interval. In the FCCS, data classified as below the LOD were below the minimum value reported by the device (Koehler et al., 2019). Data specified as MAR were missing due to device malfunction or participant noncompliance. Hence, MAR data may be below or above the LOD. The missing data type was known, and we assumed the LODs to be fixed at the minimum value the device reports. The log-transformed LODs for BC, CO, and $PM_{2.5}$ were -3.57, -3.87, and -1.14, respectively. In this analysis, approximately 0.3% of observations were MAR and 3% were below the LOD.

In addition to exposure data, the FCCS data contain covariate information that may inform the latent time-activity patterns. These variables include time of day, microenvironments, and individual identifiers that link repeated sampling days for a single individual.

3.3 Model

We first present the model for multivariate exposure data conditional on the hidden states. We then present the model for the hidden states. We describe the missing data model next. Last, we discuss posterior computation.

3.3.1 Multivariate Exposure Data Model

Let $\mathbf{y}_{ist}$ be a p-dimensional vector of exposures measured at time points $t = 1,\ldots,T_{is}$ for individuals $i = 1,\ldots,n$ on sampling days $s = 1,\ldots,S_i$. Then $\mathbf{Y}_{is,1:T_{is}}$ is the $p \times T_{is}$ matrix of multivariate time series data for T_{is} equally-spaced time points for individual i on sampling day s. Our approach allows for varying time series lengths, but for presentation purposes, we assume all time series have length T and omit the is subscript. We assume the data for each pollutant are centered and scaled to have mean 0 and variance 1. Let z_{ist} denote the hidden state for individual i on sampling day s at time t, where $z_{ist} = k$ if individual i on sampling day s is in state k at time t. We define the vector $\mathbf{z}_{is,1:T} = (z_{is1},\ldots,z_{isT})$ as the hidden state trajectory for individual i on sampling day s, which has the first-order Markov property such that $p(z_{ist}|\mathbf{z}_{is,1:t-1}) = p(z_{ist}|z_{is,t-1})$. We assume $\mathbf{y}_{ist}$ are conditionally independent of exposure

data measured for any $i' \neq i$, $s' \neq s$, or $t' \neq t$ given the hidden states. The distribution of the multivariate emission data at a single time point is

$$f(\mathbf{y}_{ist}|\mathbf{Y}_{is,1:t-1}, \mathbf{z}_{is,1:t}) = f(\mathbf{y}_{ist}|z_{ist}), \tag{3.1}$$

where any parameters associated with the hidden state are indexed by z_{ist}, and global parameters are implicit. We will refer to (3.6) as the emission distribution. We assume a Gaussian emission distribution with state-specific mean and variance

$$f(\mathbf{y}_{ist}|z_{ist} = k) \quad \equiv \quad \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.2}$$

where $\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathrm{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda}\boldsymbol{\Sigma}_k\right)$ and $\boldsymbol{\Sigma}_k \sim$ Inverse-Wishart$(\nu, \mathbf{I}_p)$. The hyperparameters $\boldsymbol{\mu}_0$, λ, and ν are fixed. We set $\boldsymbol{\mu}_0 = \mathbf{0}$ since the data are centered and scaled, and we set $\lambda = 10$ to reflect the assumption that state-specific means are less variable than the data within a state. We set $\nu = p + 2$, so $\mathrm{E}(\boldsymbol{\Sigma}_k) = \mathbf{I}_p$ a priori.

3.3.2 Hidden State Model

We model hidden states for each time point as

$$z_{ist}|z_{is,t-1} \quad \sim \quad \text{Categorical}\left(\boldsymbol{\pi}_{z_{is,t-1}}\right), \tag{3.3}$$

where $\boldsymbol{\pi}_{z_{is,t-1}}$ is the vector of probabilities for transitioning out of state $z_{is,t-1}$ into each of the possibly infinite hidden states. We model $\boldsymbol{\pi}_{z_{is,t-1}}$ using a covariate-dependent PSBP (Rodríguez and Dunson, 2011; Sarkar et al., 2012). Let $\mathbf{x}_{ist}$ represent a vector of covariates measured for individual i on sampling day s at time t. The covariates we consider are either smooth basis functions of time of day or indicator variables for the microenvironment classification of time points. Let $\pi_{jk}(\mathbf{x}_{ist}) = P(z_{ist} = k|z_{is,t-1} = j, \mathbf{x}_{ist})$ be the probability of transitioning from state j to state k at time t given the covariates $\mathbf{x}_{ist}$ for individual i on sampling day s at time t. We

construct the transition distribution probabilities as

$$\pi_{jk}(\mathbf{x}_{ist}) = \Phi(\alpha_{jk} + \mathbf{x}'_{ist}\boldsymbol{\beta}_k + \mathbf{x}'_{ist}\boldsymbol{\gamma}_{ik}) \prod_{l<k} \left[1 - \Phi(\alpha_{jl} + \mathbf{x}'_{ist}\boldsymbol{\beta}_l + \mathbf{x}'_{ist}\boldsymbol{\gamma}_{il})\right], \tag{3.4}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. In (3.4), α_{jk} is an intercept term controlling dependency between states at consecutive time points, $\mathbf{x}'_{ist}\boldsymbol{\beta}_k$ controls the propensity of being in state k at time t based on covariates measured at time t, and $\boldsymbol{\gamma}_{ik}$ are subject-specific effects that inform the propensity for individual i to be in state k at time t. Specifically, $\boldsymbol{\gamma}_{ik}$ allows for subject-level deviation from the overall population effect of covariates when considering repeated sampling days for multiple subjects as in the FCCS.

We complete the model specification with hyperpriors $\alpha_{jk}|\sigma^2_\alpha \sim \text{N}(0, \sigma^2_\alpha)$ for $j \neq k$ and $\sigma^{-2}_\alpha \sim \text{Gamma}(1,1)$ to model transitions to different states and $\alpha_{jj}|m_\alpha, v_\alpha \sim \text{N}(m_\alpha, v_\alpha)$, $m_\alpha \sim \text{N}(0,1)$, and $v^{-1}_\alpha \sim \text{Gamma}(1,1)$ to model self-transitions. We place a hierarchical model on the self-transition mass α_{jj} to allow the data to inform the tendency to linger in a state or be transient, under the assumption that personal exposure data may elicit some hidden states that are short-lived and others that occur for long periods of time. Finally, $\boldsymbol{\beta}_k \sim \text{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\gamma}_{ik}|\kappa^2 \sim \text{N}(\mathbf{0}, \kappa^2\mathbf{I})$, and $\kappa^{-2} \sim \text{Gamma}(1,1)$.

3.3.3 Missing Data Model

The previous sections described our proposed model for exposure data with no missing values. We extend this model to accommodate missing exposure data by imputing values from the missing data model, which is the posterior predictive distribution of the missing data given the observed data. The missing data model is conditional on the estimated hidden states and corresponding emission distribution parameters, hence, we account for uncertainty in the estimated hidden states in our imputation.

At each time point, the vector of exposures may have any combination of data that are observed, MAR, or below the LOD. Denote $\mathbf{y}_{\text{obs}}$ as the set of data that is observed, $\mathbf{y}_{\text{MAR}}$ as the set of data that is MAR, and $\mathbf{y}_{\text{LOD}}$ as the set of data below the LOD. We first consider MAR data and

ignore data below the LOD. If all p exposures are MAR for individual i on sampling day s at time t, the missing data model is

$$\mathbf{y}_{ist,\text{MAR}}|z_{ist}=k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\sim \text{N}\left(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right). \tag{3.5}$$

When $\mathbf{y}_{ist}$ has some exposures that are observed and some that are MAR, we partition the complete data into its observed and missing parts as $\mathbf{y}_{ist}=(\mathbf{y}_{ist,\text{obs}},\mathbf{y}_{ist,\text{MAR}})$. The emission distribution for $\mathbf{y}_{ist}$ can then be written as

$$\begin{bmatrix}\mathbf{y}_{ist,\text{obs}}\\ \mathbf{y}_{ist,\text{MAR}}\end{bmatrix}\Bigg| z_{ist}=k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\sim \text{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{(k,\text{obs})}\\ \boldsymbol{\mu}_{(k,\text{MAR})}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{(k,\text{obs,obs})} & \boldsymbol{\Sigma}_{(k,\text{obs,MAR})}\\ \boldsymbol{\Sigma}_{(k,\text{MAR,obs})} & \boldsymbol{\Sigma}_{(k,\text{MAR,MAR})}\end{bmatrix}\right). \tag{3.6}$$

In this case, the missing data model is

$$\mathbf{y}_{ist,\text{MAR}}|\mathbf{y}_{ist,\text{obs}},z_{ist}=k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\sim \text{N}\left(\boldsymbol{\mu}_{(k,\text{MAR|obs})},\boldsymbol{\Sigma}_{(k,\text{MAR|obs})}\right), \tag{3.7}$$

where

$$\begin{aligned}\boldsymbol{\mu}_{(k,\text{MAR|obs})} &= \boldsymbol{\mu}_{(k,\text{MAR})}+\boldsymbol{\Sigma}_{(k,\text{MAR,obs})}\boldsymbol{\Sigma}^{-1}_{(k,\text{obs,obs})}\left(\mathbf{y}_{it,\text{obs}}-\boldsymbol{\mu}_{(k,\text{obs})}\right) &(3.8)\\ \boldsymbol{\Sigma}_{(k,\text{MAR|obs})} &= \boldsymbol{\Sigma}_{(k,\text{MAR,MAR})}+\boldsymbol{\Sigma}_{(k,\text{MAR,obs})}\boldsymbol{\Sigma}^{-1}_{(k,\text{obs,obs})}\boldsymbol{\Sigma}_{(k,\text{obs,MAR})}. &(3.9)\end{aligned}$$

The missing data model for data below the LOD is similar. We assume the LOD is fixed and known for each exposure. Data that fall below the LOD are censored at the LOD. At each time point, we partition the complete data into its observed and below LOD parts: $\mathbf{y}_{ist}=(\mathbf{y}_{ist,\text{obs}},\mathbf{y}_{ist,\text{LOD}})$. The emission distribution for $\mathbf{y}_{ist}$ can then be written as

$$\begin{bmatrix}\mathbf{y}_{ist,\text{obs}}\\ \mathbf{y}_{ist,\text{LOD}}\end{bmatrix}\Bigg| z_{ist}=k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\sim \text{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{(k,\text{obs})}\\ \boldsymbol{\mu}_{(k,\text{LOD})}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{(k,\text{obs,obs})} & \boldsymbol{\Sigma}_{(k,\text{obs,LOD})}\\ \boldsymbol{\Sigma}_{(k,\text{LOD,obs})} & \boldsymbol{\Sigma}_{(k,\text{LOD,LOD})}\end{bmatrix}\right). \tag{3.10}$$

Let $\mathbf{d} = (d_1, \ldots, d_p)'$ be the vector of lower limits of detection for components $j = 1, \ldots p$. Let D_{ist} be the support for data that is below the LOD for individual i on sampling day s at time t. For a single log-transformed exposure j that is below the LOD, $D_{ist} = (-\infty, \log(d_j))$. The missing data model for data below the LOD is

$$\mathbf{y}_{ist,\text{LOD}} | \mathbf{y}_{ist,\text{obs}}, z_{ist} = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \quad \sim \quad \text{TN}_{D_{ist}} \left(\boldsymbol{\mu}_{(k,\text{LOD}|\text{obs})}, \boldsymbol{\Sigma}_{(k,\text{LOD}|\text{obs})} \right), \tag{3.11}$$

where $\text{TN}_{D_{ist}}$ represents the truncated multivariate normal distribution restricted to the support D_{ist} and

$$\begin{aligned} \boldsymbol{\mu}_{(k,\text{LOD}|\text{obs})} &= \boldsymbol{\mu}_{(k,\text{LOD})} + \boldsymbol{\Sigma}_{(k,\text{LOD,obs})} \boldsymbol{\Sigma}^{-1}_{(k,\text{obs,obs})} \left(\mathbf{y}_{ist,\text{obs}} - \boldsymbol{\mu}_{(k,\text{obs})} \right) \\ \boldsymbol{\Sigma}_{(k,\text{LOD}|\text{obs})} &= \boldsymbol{\Sigma}_{(k,\text{LOD,LOD})} + \boldsymbol{\Sigma}_{(k,\text{LOD,obs})} \boldsymbol{\Sigma}^{-1}_{(k,\text{obs,obs})} \boldsymbol{\Sigma}_{(k,\text{obs,LOD})}. \end{aligned}$$

3.3.4 Posterior Computation

We implement a Metropolis-within-Gibbs algorithm to sample from the posterior distribution. After a burn-in period, the remaining samples are used for inference. The iterative steps of the MCMC sampler are outlined in Algorithm 1. Our computation approach closely mirrors that described in Sarkar et al. (2012). Software to fit our proposed approach exists in the `R` package `psbpHMM` (Hoskovec, 2021b), available at github.com/lvhoskovec/psbpHMM.

To sample the hidden state trajectories $\mathbf{z}_{is,1:T}$ for individuals $i = 1, \ldots, n$, sampling days $s = 1, \ldots, S_i$, and time points $t = 1, \ldots, T$, we implement beam sampling (Van Gael et al., 2008), a combination of slice sampling (Neal, 2003; Walker, 2007) and dynamic programming. Beam sampling allows each time series' entire hidden state trajectory to be sampled simultaneously, which improves mixing and convergence in highly dependent data. At each iteration, we introduce auxiliary slice variables u_{ist} for each individual i, sampling day s, and time point t that reduce the number of hidden states considered at that time point. Specifically, for z_{ist}, only states k such that $\pi_{z_{is,t-1},k} > u_{ist}$ are considered, where $\pi_{z_{is,t-1},k}$ is the transition probability

from state $z_{is,t-1}$ to state k and may depend on covariates $\mathbf{x}_{ist}$. For presentation purposes, we remove the dependence on covariates in the transition probabilities when describing posterior sampling for the hidden states. By limiting the number of states considered at each time point, we limit the number of possible hidden state trajectories from which we sample in each iteration. Slice sampling permits implementation of a forward-backward algorithm to sample entire hidden state trajectories at once.

We sample the auxiliary slice variables, u_{ist}, from the conditional distribution

$$u_{ist}|z_{ist}, z_{is,t-1}, \boldsymbol{\pi}_{is} \quad \sim \quad \text{Uniform}\left[0, \pi_{z_{is,t-1},z_{ist}}\right], \tag{3.12}$$

where $\boldsymbol{\pi}_{is}$ denotes the set of all transition probabilities for individual i on sampling day s.

Let $\mathbf{u}_{is,1:T}$ denote the slice variables for individual i on sampling day s for time points $t = 1,\ldots,T$. Let $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the state-specific emission distribution parameters for all latent states. To sample the hidden state trajectories $\mathbf{z}_{is,1:T}$ for $i = 1,\ldots,n$ and $s = 1,\ldots,S_i$, we use a forward-backward algorithm. Let $I(\cdot)$ denote the indicator function. In the forward step, we recursively compute the distribution of z_{ist} given $\mathbf{Y}_{is,1:t}, \mathbf{u}_{is,1:t}, \boldsymbol{\pi}_{is}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as

$$\begin{aligned}
p(z_{ist}|\mathbf{Y}_{is,1:t}, \mathbf{u}_{is,1:t}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= p(z_{ist}|\mathbf{y}_{ist}, u_{ist}, \mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\propto p(z_{ist}, \mathbf{y}_{ist}, u_{ist}|\mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \sum_{z_{is,t-1}} p(z_{ist}, \mathbf{y}_{ist}, u_{ist}, z_{is,t-1}|\mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \sum_{z_{is,t-1}} f(\mathbf{y}_{ist}|z_{ist}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(u_{ist}|z_{ist}, z_{is,t-1}, \boldsymbol{\pi}_{is}) p(z_{ist}|z_{is,t-1}, \boldsymbol{\pi}_{is}) \times \\
&\quad p(z_{is,t-1}|\mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= f(\mathbf{y}_{ist}|z_{ist}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sum_{z_{is,t-1}} I\left[0 < u_{ist} < \pi_{z_{is,t-1},z_{ist}}\right] \times \\
&\quad p(z_{is,t-1}|\mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= f(\mathbf{y}_{ist}|z_{ist}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sum_{z_{is,t-1}: u_{ist} < \pi_{z_{is,t-1},z_{ist}}} p(z_{is,t-1}|\mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}).
\end{aligned} \tag{3.13}$$

The use of slice sampling truncates the sum over $z_{is,t-1}$ to the finitely many values such that $0 < u_{ist} < \pi_{z_{is,t-1},z_{ist}}$ and $p(z_{is,t-1}|\mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) > 0$. For $t = 1$, we assume that $P(z_{is0} = 0) = 1$ for all i and s. Then

$$
\begin{aligned}
p(z_{is1}|\mathbf{y}_{is1}, u_{is1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto p(z_{is1}, \mathbf{y}_{is1}, u_{is1}|\boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= p(z_{is1}, \mathbf{y}_{is1}, u_{is1}, z_{is0}|\boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= f(\mathbf{y}_{is1}|z_{is1}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(z_{is1}|z_{is0}, \boldsymbol{\pi}_i) p(u_{is1}|z_{is0}, z_{is1}, \boldsymbol{\pi}_{is}) p(z_{is0}|\boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= f(\mathbf{y}_{is1}|z_{is1}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) I\left[0 < u_{is1} < \pi_{0,z_{is1}}\right].
\end{aligned} \tag{3.14}
$$

A new state may be proposed among the multiple time series. For identifiability, we only permit one new state to be proposed in each iteration of the MCMC sampler. If z_{ist} belongs to a new state k^* with nonzero probability, we evaluate $p(\mathbf{y}_{ist}|z_{ist} = k^*)$ by integrating over the possible values of $(\boldsymbol{\mu}_{k^*}, \boldsymbol{\Sigma}_{k^*})$ for the new state. Using the Normal-Inverse-Wishart model on $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, we have

$$
\begin{aligned}
f(\mathbf{y}_{ist}|z_{ist} = k^*) &= \int_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} f(\mathbf{y}_{ist}|z_{ist} = k^*, \boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_{k^*}) p(\boldsymbol{\mu}_{k^*}, \boldsymbol{\Sigma}_{k^*}) \mathrm{d}(\boldsymbol{\mu}_{k^*}, \boldsymbol{\Sigma}_{k^*}) \\
&= \left(\frac{\lambda}{(\pi)(1+\lambda)}\right)^{\frac{p}{2}} \left(\frac{\Gamma_p(\frac{\nu+1}{2})}{\Gamma_p(\frac{\nu}{2})}\right) \left(\frac{|\mathbf{I}_p|^{\frac{\nu}{2}}}{|R_{ist}^*|^{\frac{\nu+1}{2}}}\right),
\end{aligned} \tag{3.15}
$$

where $R_{ist}^* = \mathbf{I}_p + \lambda\boldsymbol{\mu}_0\boldsymbol{\mu}_0' + \mathbf{y}_{ist}\mathbf{y}_{ist}' - \left(\frac{1}{1+\lambda}\right)(\lambda\boldsymbol{\mu}_0 + \mathbf{y}_{ist})(\lambda\boldsymbol{\mu}_0 + \mathbf{y}_{ist})'$. In the backward step, we sample the latent sequence from

$$
p(z_{ist} = k|z_{is,t+1}, \mathbf{Y}_{is,1:t}, \mathbf{u}_{is,1:t}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \propto \tag{3.16}
$$

$$
\begin{cases}
p(z_{ist} = k|\mathbf{Y}_{is,1:t}, \mathbf{u}_{is,1:t}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), & t = T \\
p(z_{ist} = k|\mathbf{Y}_{is,1:t}, \mathbf{u}_{is,1:t}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(z_{is,t+1}|z_{ist} = k, u_{is,t+1}, \boldsymbol{\pi}_{is}), & 0 < t < T,
\end{cases}
$$

where

$$
\begin{aligned}
p(z_{is,t+1}|z_{ist},u_{is,t+1},\boldsymbol{\pi}_{is}) &\propto p(u_{is,t+1}|z_{is,t+1},z_{ist},\boldsymbol{\pi}_{is})p(z_{is,t+1}|z_{ist},\boldsymbol{\pi}_{is}) \\
&= \frac{I\left[0<u_{is,t+1}<\pi_{z_{ist},z_{is,t+1}}\right]}{\pi_{z_{ist},z_{is,t+1}}}\left[\pi_{z_{ist},z_{is,t+1}}\right] \\
&= I\left[0<u_{is,t+1}<\pi_{z_{ist},z_{is,t+1}}\right].
\end{aligned}
$$

To sample the parameters of the transition distribution, we follow a similar approach to that used in Bayesian probit regression (Chung and Dunson, 2009). Denote the currently occupied states $k = 1,\ldots,K$ and let $K+1$ denote a potential new state. Conditional on $z_{is,t-1} = j$ and $z_{ist} = k$, we introduce auxiliary variables w_{jistl} for each individual i, sampling day s, time point t, and state $l = 1,\ldots,k$ to represent the pieces of the PSBP for all individuals, sampling days, time points, and possible hidden state transitions. These auxiliary variables have the conditional probability

$$
w_{jistl}|z_{is,t-1}=j,z_{ist}=k,\alpha_{jl},\boldsymbol{\beta}_l,\mathbf{x}_{ist} \overset{\text{ind}}{\sim} \begin{cases} \text{TN}_{(0,\infty)}\left(\alpha_{jl}+\mathbf{x}'_{ist}\boldsymbol{\beta}_l+\mathbf{x}'_{ist}\boldsymbol{\gamma}_{il},1\right), & l=k \\ \text{TN}_{(-\infty,0)}\left(\alpha_{jl}+\mathbf{x}'_{ist}\boldsymbol{\beta}_l+\mathbf{x}'_{ist}\boldsymbol{\gamma}_{il},1\right), & l<k, \end{cases} \tag{3.17}
$$

where TN_A denotes the truncated normal distribution restricted to the set A. We sample α_{jk} for states $j \neq k$ from the full conditional

$$
\begin{aligned}
\alpha_{jk}|\cdot &\sim \text{N}\left(m_{jk},v_{jk}\right) \\
v_{jk} &= \left(\sigma_\alpha^{-2}+n_{jk}\right)^{-1} \\
m_{jk} &= v_{jk}\left[\mu_\alpha\sigma_\alpha^{-2}+\sum_{ist:z_{ist}\geq k,z_{is,t-1}=j}\left(w_{jistk}-\mathbf{x}'_{ist}\boldsymbol{\beta}_k-\mathbf{x}'_{ist}\boldsymbol{\gamma}_{ik}\right)\right],
\end{aligned} \tag{3.18}
$$

where $n_{jk} = \sum_{i=1}^{n} \sum_{s=1}^{S_i} \sum_{t=2}^{T} I(z_{ist} \geq k, z_{is,t-1} = j)$. We sample α_{jj} from the full conditional

$$\begin{aligned} \alpha_{jj}|\cdot &\sim \mathrm{N}\left(m_j, v_j\right) \\ v_j &= (v_\alpha^{-1} + n_{jj})^{-1} \\ m_j &= v_j \left[m_\alpha v_\alpha^{-1} + \sum_{ist: z_{ist} \geq j, z_{is,t-1} = j} \left(w_{jistj} - \mathbf{x}'_{ist}\boldsymbol{\beta}_j - \mathbf{x}'_{ist}\boldsymbol{\gamma}_{ij} \right) \right] \end{aligned} \tag{3.19}$$

where $n_{jj} = \sum_{i=1}^{n} \sum_{s=1}^{S_i} \sum_{t=2}^{T} I(z_{ist} \geq j, z_{is,t-1} = j)$. We sample m_α from

$$\begin{aligned} m_\alpha|\cdot &\sim \mathrm{N}(m^*, v^*) \\ v^* &= \left(\frac{K}{v_\alpha} + v_0^{-1} \right)^{-1} \\ m^* &= v^* \left(\frac{\sum_{j=1}^{K} \alpha_{jj}}{v_\alpha} + m_0 v_0^{-1} \right), \end{aligned} \tag{3.20}$$

where K is the current number of occupied states. We sample σ_α^{-2} and v_α^{-1} from

$$\sigma_\alpha^{-2}|\cdot \sim \mathrm{Gamma}\left[1 + \frac{K(K-1)}{2}, 1 + \frac{\sum_{j \neq k} \alpha_{jk}^2}{2} \right] \tag{3.21}$$

$$v_\alpha^{-1}|\cdot \sim \mathrm{Gamma}\left[1 + \frac{K}{2}, 1 + \frac{\sum_{j=1}^{K} (\alpha_{jj} - m_\alpha)^2}{2} \right]. \tag{3.22}$$

We sample $\boldsymbol{\beta}_k$ from

$$\begin{aligned} \boldsymbol{\beta}_k|\cdot &\sim \mathrm{N}\left(\mathbf{m}_k, \mathbf{V}_k\right) \\ \mathbf{V}_k &= \left(\mathbf{I} + \mathbf{X}'_k\mathbf{X}_k\right)^{-1} \\ \mathbf{m}_k &= \mathbf{V}_k \left[\mathbf{X}'_k(\mathbf{w}_k - \boldsymbol{\alpha}_k - \boldsymbol{\Gamma}_k) \right], \end{aligned} \tag{3.23}$$

where $\mathbf{X}_k$ has $n_k = \sum_{i=1}^{n} \sum_{s=1}^{S_i} \sum_{t=1}^{T} I(z_{ist} \geq k)$ rows and is the matrix of covariates for all i, s, t such that $z_{ist} \geq k$, $\mathbf{w}_k$ is the n_k-dimensional vector of w_{jistk} for all j, i, s, t such that $z_{ist} \geq k$ and $z_{is,t-1} = j$, $\boldsymbol{\alpha}_k$ is the n_k-dimensional vector of $\alpha_{z_{is,t-1},k}$ for all i, s, t such that $z_{ist} \geq k$, and $\boldsymbol{\Gamma}_k$ is the n_k-dimensional vector of $\mathbf{x}'_{ist}\boldsymbol{\gamma}_{ik}$ for all i, s, t such that $z_{ist} \geq k$. The subject-specific effects

$\boldsymbol{\gamma}_{ik}$ are updated similarly. We sample κ^{-2} from

$$\kappa^{-2}|\cdot \sim \text{Gamma}\left[1+\frac{n^*K}{2}, 1+\frac{1}{2}\sum_{i^*=1}^{n^*}\sum_{k=1}^{K}(\boldsymbol{\gamma}_{i^*k}-\boldsymbol{\mu}_\gamma)'\boldsymbol{\Sigma}_\gamma^{-1}(\boldsymbol{\gamma}_{i^*k}-\boldsymbol{\mu}_\gamma)\right], \tag{3.24}$$

where, here, $i^* = 1,\ldots,n^*$ denote the unique subjects.

Conditional on the transition distribution parameters, the updates for the transition probabilities are deterministic. For individuals $i = 1,\ldots,n$, sampling days $s = 1,\ldots,S_i$ and times $t = 1,\ldots,T$, we calculate $\{\pi_{jk}(\mathbf{x}_{ist})\}_{j=1,k=1}^{K,K}$ for the K currently occupied states as

$$\pi_{jk}(\mathbf{x}_{ist}) = \Phi(\alpha_{jk}+\mathbf{x}'_{ist}\boldsymbol{\beta}_k+\mathbf{x}'_{ist}\boldsymbol{\gamma}_{ik})\prod_{l<k}\left[1-\Phi(\alpha_{jl}+\mathbf{x}'_{ist}\boldsymbol{\beta}_l+\mathbf{x}'_{ist}\boldsymbol{\gamma}_{il})\right], \tag{3.25}$$

and we calculate $\{\pi_{j,K+1}(\mathbf{x}_{ist})\}_{j=1}^{K}$ for transitions into a potential new state $K+1$ as

$$\pi_{j,K+1}(\mathbf{x}_{ist}) = 1-\sum_{k=1}^{K}\pi_{jk}(\mathbf{x}_{ist}). \tag{3.26}$$

On data sets with no missing observations, we update $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ with Gibbs sampling. The full conditional for $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is

$$\begin{aligned}
(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)|\cdot &\sim \text{Normal-Inverse-Wishart}\left(\mu_{n_k}, \lambda_{n_k}, \Sigma_{n_k}, \nu_{n_k}\right) \\
\mu_{n_k} &= \frac{\lambda\boldsymbol{\mu}_0+\tilde{n}_k\bar{\mathbf{y}}_k}{\lambda+\tilde{n}_k} \\
\lambda_{n_k} &= \lambda+\tilde{n}_k \\
\nu_{n_k} &= \nu+\tilde{n}_k \\
\Sigma_{n_k} &= \mathbf{I}_p+\sum_{ist:z_{ist}=k}(\mathbf{y}_{ist}-\bar{\mathbf{y}}_k)(\mathbf{y}_{ist}-\bar{\mathbf{y}}_k)'+\frac{\lambda\tilde{n}_k}{\lambda+\tilde{n}_k}(\bar{\mathbf{y}}_k-\boldsymbol{\mu}_0)(\bar{\mathbf{y}}_k-\boldsymbol{\mu}_0)',
\end{aligned} \tag{3.27}$$

where $\tilde{n}_k = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=1}^{T} I(z_{ist}=k)$, $\bar{\mathbf{y}}_k = \frac{1}{\tilde{n}_k}\sum_{ist:z_{ist}=k}\mathbf{y}_{ist}$. First, we sample

$$\boldsymbol{\Sigma}_k|\cdot \sim \text{Inverse Wishart}\left(\nu_{n_k}, \Sigma_{n_k}\right). \tag{3.28}$$

Then, we sample

$$\boldsymbol{\mu}_k|\cdot \quad \sim \quad \mathrm{N}\left(\mu_{n_k}, \frac{1}{\lambda_{n_k}}\boldsymbol{\Sigma}_k\right). \tag{3.29}$$

To improve mixing and empirical performance on data sets with missing observations, we reparameterize $\boldsymbol{\Sigma}_k$ so we can sample each parameter of the covariance matrices separately. We follow Chan and Jeliazkov (2009) and let $\boldsymbol{\Sigma}_k = \mathbf{L}_k^{-1}\mathbf{D}_k\left(\mathbf{L}_k^{-1}\right)'$, where

$$\mathbf{L}_k \equiv \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{k,21} & 1 & 0 & \cdots & 0 \\ a_{k,31} & a_{k,32} & 1 & \cdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_{k,p1} & a_{k,p2} & & \cdots & 1 \end{bmatrix}, \quad \mathbf{D}_k \equiv \begin{bmatrix} \delta_{k,1} & 0 & \cdots & 0 \\ 0 & \delta_{k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_{k,p} \end{bmatrix}. \tag{3.30}$$

We model each of the parameters of $\boldsymbol{\Sigma}_k$ separately as

$$\delta_{k,j} \overset{\text{ind}}{\sim} \text{Inverse Gamma}\left(\frac{\nu+j-p}{2}, \frac{1}{2}\right) \quad j = 1,\ldots,p \tag{3.31}$$

$$a_{k,jl} \overset{\text{ind}}{\sim} \mathrm{N}(0, \delta_{k,j}) \quad j = 2,\ldots,p, \text{ and } l = 1,\ldots,j-1. \tag{3.32}$$

With this formulation, Chan and Jeliazkov (2009) demonstrate that $\boldsymbol{\Sigma}_k$ follows an inverse-Wishart distribution with degrees of freedom ν and scale matrix $\mathbf{I}_p$.

We implement an independence Metropolis-Hastings sampler to update $\delta_{k,j}$ and $a_{k,jl}$ for all j, l, and currently occupied k. Metropolis-Hastings updates help avoid local modes in mixture distributions (Celeux et al., 2000) and were used in a periodic infinite hidden Markov model in Spezia et al. (2011). We found the independence sampler empirically performed better than Gibbs sampling and random walk Metropolis-Hastings, particularly when small clusters form that contain many missing observations.

Let $\theta^{(b)}$ denote the value of the parameter θ at iteration b. Let $p(\cdot)$ denote a prior distribution, $q(\cdot)$ a proposal distribution, and $f(\cdot)$ a likelihood. For $k = 1,\ldots K$ and $j = 1,\ldots,p$, we imple-

ment the following steps to update $\delta_{k,j}$. First, we propose $\delta^*_{k,j} \sim$ Inverse Gamma(a_δ, b_δ). Then, we create $\mathbf{D}^*_k$, which is equivalent to $\mathbf{D}^{(b)}_k$ except $\delta^*_{k,j}$ replaces $\delta^{(b)}_{k,j}$. We calculate the Metropolis-Hastings ratio as

$$r_\delta = \frac{f\left(\mathbf{Y}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}^*_k\right) f\left(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}^*_k\right) p\left(\delta^*_{k,j}\right) q\left(\delta^{(b)}_{k,j}\right)}{f\left(\mathbf{Y}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}^{(b)}_k\right) f\left(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}^{(b)}_k\right) p\left(\delta^{(b)}_{k,j}\right) q\left(\delta^*_{k,j}\right)}, \tag{3.33}$$

where $\mathbf{Y}_k$ is the $n_{kk} \times p$ matrix of exposure data for all time points assigned to hidden state k and $\boldsymbol{\Sigma}^*_k = \mathbf{L}^{(b)^{-1}}_k \mathbf{D}^*_k \mathbf{L}^{(b)^{-1}T}_k$. Finally, we accept $\delta^*_{k,j}$ with probability $\min(1, r_\delta)$.

For $k = 1, \ldots, K$, $j = 2, \ldots, p$, and $l = 1, \ldots, j-1$, we implement the following steps to update $a_{k,jl}$. First we propose $a^*_{k,jl} \sim \mathrm{N}(0, \tau^2)$. Then, we create $\mathbf{L}^*_k$, which is equivalent to $\mathbf{L}^{(b)}_k$ except $a^*_{k,jl}$ replaces $a^{(b)}_{k,jl}$. Finally we calculate the Metropolis-Hastings ratio as

$$r_a = \frac{f\left(\mathbf{y}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}^*_k\right) f\left(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}^*_k\right) p\left(a^*_{k,jl}\right) q\left(a^{(b)}_{k,jl}\right)}{f\left(\mathbf{y}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}^{(b)}_k\right) f\left(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}^{(b)}_k\right) p\left(a^{(b)}_{k,jl}\right) q\left(a^*_{k,jl}\right)}, \tag{3.34}$$

where $\boldsymbol{\Sigma}^*_k = \mathbf{L}^{*^{-1}}_k \mathbf{D}^{(b)}_k \mathbf{L}^{*^{-1}T}_k$. We accept $a^*_{k,jl}$ with probability $\min(1, r_a)$.

Through empirical testing, we found the tuning parameters $a_\delta = 10, b_\delta = 1$, and $\tau^2 = 0.25$ achieve acceptance probabilities between 0.1 and 0.4 for all parameters. A resolvent kernel was used to obtain optimal acceptance ratios for $a^*_{k,jl}$ (Robert and Casella, 2004). At each iteration, we randomly generated the number of times to repeat the Metropolis Hasting steps for updating $a^*_{k,jl}$ from a geometric distribution with probability parameter 1/5.

Algorithm 1: MCMC algorithm for covariate-dependent iHMM for multiple time series with missing data

Result: Posterior samples of estimated hidden states $\mathbf{z}_{is,1:T}$ for individuals $i = 1,\ldots,n$ on sampling days $s = 1,\ldots,S_i$, and state-specific parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for all occupied k. Multiple imputations for observations that are MAR or below the LOD.

Specification of fixed hyperparameters.

Random initialization for all parameters.

for *each iteration* **do**

  for i *in* $1,\ldots,n$ **do**

    for s *in* $1,\ldots,S_i$ **do**

      Draw $u_{ist}|z_{ist}, z_{is,t-1}, \mathbf{x}_{ist}, \boldsymbol{\pi}_{is}$ for each t;

      Define the set of possible states for each t as $\tilde{k}_{ist} = \{z_{ist} : 0 < u_{ist} < \pi_{z_{is,t-1},z_{ist}}(\mathbf{x}_{ist}) \text{ and } 0 < u_{is,t+1} < \pi_{z_{ist},z_{is,t+1}}(\mathbf{x}_{is,t+1})\}$;

      Calculate $p(z_{is1} = k|\mathbf{y}_{is1}, u_{is1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \propto f(\mathbf{y}_{is1}|z_{is1} = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for all $k \in \tilde{k}_{is1}$;

      for t *in* $2,\ldots,T$ **do**

        Calculate $p(z_{ist} = k|\mathbf{y}_{is,1:t}, \mathbf{u}_{is,1:t}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \propto f(\mathbf{y}_{ist}|z_{ist} = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sum_{z_{is,t-1}: u_{ist} < \pi_{z_{is,t-1},k}(\mathbf{x}_{ist})} p(z_{is,t-1}|\mathbf{Y}_{is,1:t-1}, \mathbf{u}_{is,1:t-1}, \boldsymbol{\pi}_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\forall k \in \tilde{k}_{ist}$;

      end

    end

    Sample z_{ist} where $p(z_{ist} = k|-) \propto p(z_{ist} = k|\mathbf{Y}_{is,1:t}, \mathbf{u}_{i,1:t}, \boldsymbol{\pi}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(k \in \tilde{k}_{ist})$ for all t;

  end

  Update K, the number of unique non-empty states;

  Sample $\boldsymbol{\mu}_k|-$ for k in $1,\ldots,K$ using Gibbs sampling;

  Sample $\boldsymbol{\Sigma}_k|-$ for k in $1,\ldots,K$ using Gibbs sampling, or Metropolis-Hastings if there are missing data;

  Sample $\mathbf{w}$.....$|-$ from its full conditional;

  Sample the transition distribution parameters from their full conditionals;

  Update $\boldsymbol{\pi}_{is}|-$ deterministically for each $i = 1,\ldots,n$ and $s = 1,\ldots,S_i$

  Sample missing values from posterior predictive distribution, if applicable

end

Multiple imputation for MAR data and data below the LOD proceeds by sampling from the posterior predictive distributions of the missing data given the observed data, which are described in Section 3.3.3. By definition, $\mathbf{y}_{ist}$ are conditionally independent of all other exposure data given z_{ist}. We can therefore sample from the posterior predictive distribution of the missing data separately for each time point. For data below the LOD, where the posterior predictive distribution is a truncated multivariate normal distribution, we use the hybrid sampler proposed by Li and Ghosh (2015). When $\mathbf{y}_{ist}$ has both data that are MAR and below the LOD, we sample from the posterior predictive distribution for each missing data type conditional on the other missing data type, which is then assumed to be part of the observed data.

3.4 Simulation Studies

We tested the performance of our proposed method in a simulation study. We compared five models. The first two models are variations of our proposed approach, which we term 'joint' models since we fit our model jointly to all time series. The 'joint cyclical' model includes a cyclical harmonic function of time as covariates to reflect cyclical daily patterns. In the simulation study, we do not consider repeated time series for individuals and do not consider subject-specific effects. We therefore drop the subscript s in the notation in this section. To create the cyclical function, we scaled the time of day to the interval $(0, 2\pi)$ and defined $\mathbf{x}'_{it} = [\sin(h_{it}), \cos(h_{it}), \sin(2h_{it}), \cos(2h_{it})]$, where h_{it} denotes the scaled time of day for individual i at time point t. The 'joint no covariates' model does not include any covariates. To evaluate the benefit of our joint approach for all time series over the naive approach of fitting independent models for each time series, we fit the cyclical model and the model without covariates separately to each time series ('independent cyclical' and 'independent no covariates', respectively). Finally, to quantify the importance of temporal structure in the modeling approach, we fit a Dirichlet process mixture model (joint DPMM) that allows shared states among time series but includes no temporal dependency. All models in our simulation study account for missing data using the same missing data model described in 3.3.3. We did not consider

other iHMMs in our simulation study because methods with existing software lack the ability to simultaneously impute both MAR and below LOD data and accommodate multiple asynchronous time series.

Our proposed approach is computationally complex, but mixes quickly due to beam sampling of the hidden state trajectories. The computational time is $O(nTK^2)$. The time to run 1000 iterations of our MCMC sampler on our simulated data is 46 minutes on a personal laptop (Processor: 3.1 GHz Dual-Core Intel Core i5, Memory: 16 GB 2133 MHz LPDDR3) in R version 4.0.3. We assessed convergence with trace plots of imputations and the estimated number of hidden states (Appendix B.1, Figures B.1-B.3). Evidence of convergence appeared within 5000 iterations. Hence, we based inference on 5000 iterations after a burn-in of 5000 iterations.

3.4.1 Data-Generating Process for Simulated Data

We considered two simulation scenarios, one with shared temporal trends among individuals and one with distinct temporal trends for each individual. In both scenarios, we simulated $n = 20$ time series of length $T = 288$ to emulate data recorded every 5 minutes over a 24-hour period, similar to our application. We considered $p = 3$ mixture components. We set the true number of hidden states to $K = 20$.

For individuals $i = 1,\ldots,20$, we first sampled unordered state labels for $t = 1,\ldots,288$ as $z^*_{it}|\boldsymbol{\rho}_i \sim \text{Categorical}(\boldsymbol{\rho}_i)$ and $\boldsymbol{\rho}_i \sim \text{Dirichlet}_{20}(20,19,18,\ldots,3,2,1)$. We then grouped the states by index. In the shared trends scenario, we sorted the states for each individual as 1 to 20 so all individuals traveled through the states in the same order. We set $t = 1$ to be halfway though state 1 so all individuals started and ended in state 1. Due to small state allocation probabilities, some hidden states were not generated for all individuals. In the distinct trends scenario, we randomly permuted the ordering of the states for each individual. Each individual began and ended in the same state, but the hidden state sequence differed for each individual to reflect distinct temporal trends. Our data-generating process induces implicit dependence on both time and previous state, which is well-represented by our model. In addition, the process generates

some highly-frequented hidden states as well as some hidden states that are only visited for a small number time points, mimicking the heterogeneity observed in the FCCS data. Further, we intentionally did not simulate directly from our proposed model to test performance in a more realistic setting where none of the models considered exactly match the true data-generating mechanism.

In both scenarios, we randomly generated the state-specific emission distribution means $\boldsymbol{\mu}_k, k = 1,\ldots,20$, as $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0$ is a diagonal matrix with elements 0.7, 0.4, and -0.2. To create the state-specific covariance matrices $\boldsymbol{\Sigma}_k$, $k = 1,\ldots,20$, we generated lower diagonal matrices $\mathbf{L}_k$ with 1's on the main diagonal and off-diagonal elements simulated from a $\mathrm{N}(0, 0.5)$ distribution. We defined $\boldsymbol{\Sigma}_k \equiv \left(\frac{1}{100}\right)\mathbf{L}_k^{-1}(\mathbf{L}_k^{-1})'$. We simulated data by $\mathbf{y}_{it}|z_{it} = k \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and then scaled the data so each component had mean 0 and variance 1.

We constructed data sets with missing data levels of 0% (i.e. completely observed data), 5%, 10%, and 20%. For each missing data level, we specified half of the missing data as MAR and half as below the LOD. We randomly removed MAR data in chunks of size 1 to 10 time points to reflect the idea that data may be missing in sequences due to instrument failure or participant noncompliance. For missing data below the LOD, we removed all data that fell below the quantiles defined as half the missing data level (e.g., 2.5%, 5%, and 10%). We simulated 100 data sets for each scenario and missing data level.

3.4.2 Evaluation Criteria

To evaluate hidden state estimation, we reported the mean estimated number of hidden states ($\hat{K}$) for each method. We calculated the estimated number of hidden states as the average number of occupied hidden states in each MCMC iteration post burn-in. For the independently fit iHMMs, we reported the total mean estimated number of hidden states for all 20 time series since these methods estimate unique hidden states for each individual. We assigned estimated states to true states to maximize overlap, and calculated the resulting Hamming distance (Van Gael et al., 2008). Hamming distance is the number of time points at which the true

states and estimated states do not align. Our final evaluation metric was the proportion of incorrectly classified time points. For the independently fit iHMMs, we calculated Hamming distance separately for each time series and reported the mean proportion of incorrectly classified time points across the 20 time series. We evaluated state-specific mean estimation via mean squared error (MSE) (see Appendix B.2 for details). On data sets with missing observations, we calculated MSE and bias for MAR data and data below the LOD averaged over 400 imputations. We reported the mean for each measure across 100 simulated data sets.

3.4.3 Simulation Results

Results from the shared trends scenario simulation study are shown in Table 3.1. At all levels of missingness, the joint cyclical model was best able to estimate hidden states. By fitting a single model to all time series instead of fitting a separate model to each time series, the joint cyclical model estimated fewer, larger states. In most cases, this translated into better estimation of the state-specific means and better imputation of missing data.

On completely observed data, the joint cyclical model had an average estimated number of hidden states (14.25) closest to the truth, smallest Hamming distance (0.23), and MSE for state-specific means of 0.06. The next best method was the joint no covariates model, which estimated 12.78 hidden states on average, and had mean Hamming distance of 0.31 and MSE for state-specific means of 0.08. The joint DPMM followed, with an estimated 30.13 hidden states, Hamming distance of 0.33, and MSE for state-specific means of 0.05. The independently fit iHMMs performed worst in both measures, and substantially over-estimated the number of hidden states ($\hat{K}$ = 125.97 for independent cyclical model and $\hat{K}$ = 100.80 for independent no covariates model) since they estimate unique states for each time series.

At 5% missing data, the joint cyclical model estimated 13.50 hidden states on average, with Hamming distance of 0.30 and MSE for state-specific means of 0.08. The joint no covariates model was the next best method, estimating an average of 11.06 hidden states with Hamming distance of 0.39 and MSE for state-specific means of 0.10. The independently fit iHMMs ($\hat{K}$ =

Table 3.1: Results from the shared trends scenario simulation study. The two variations of our proposed joint iHMM approach are the model with cyclical trends (joint cyclical) and the model with no covariates (joint no covariates). We include the model with cyclical trends fit independently to each time series (indep. cyclical) and the model with no covariates fit independently to each time series (indep. no covariates). Last is the Dirichlet process mixture model (joint DPMM) fit jointly to all time series. The table shows the following measures: mean estimated number of hidden states ($\hat{K}$); mean Hamming distance, which is a measure of the distance between the estimated hidden state trajectories and the true hidden state trajectories; mean MSE for the state-specific means ($\boldsymbol{\mu}_{\text{MSE}}$); mean MSE and bias for the MAR and below LOD data imputations. Results are shown for four levels of missing data: 0%, 5%, 10%, and 20%. Standard errors are shown in Table B.1.

	Method	$\hat{K}$	Hamming	$\boldsymbol{\mu}_{\text{MSE}}$	MAR MSE	LOD MSE	MAR bias	LOD bias
0%	joint cyclical	14.25	0.23	0.07	–	–	–	–
	joint no covariates	12.78	0.31	0.08	–	–	–	–
	indep. cyclical	125.97	0.52	0.38	–	–	–	–
	indep. no covariates	100.80	0.61	0.48	–	–	–	–
	joint DPMM	30.13	0.33	0.05	–	–	–	–
5%	joint cyclical	13.50	0.30	0.08	0.46	3.01	-0.06	-0.77
	joint no covariates	11.06	0.39	0.10	0.62	2.24	-0.06	-0.60
	indep. cyclical	122.62	0.52	0.28	1.02	4.26	-0.06	-1.08
	indep. no covariates	96.82	0.61	0.35	1.11	3.49	-0.05	-0.93
	joint DPMM	47.84	0.49	8.85	109.05	415.44	-2.83	-8.68
10%	joint cyclical	12.73	0.35	0.20	0.71	4.52	-0.05	-0.83
	joint no covariates	11.69	0.44	0.21	0.82	3.60	-0.07	-0.86
	indep. cyclical	117.80	0.53	0.32	1.12	4.29	-0.07	-1.00
	indep. no covariates	93.32	0.62	0.40	1.29	3.67	-0.08	-0.92
	joint DPMM	54.78	0.51	16.84	110.77	343.35	-2.69	-6.98
20%	joint cyclical	13.55	0.33	0.38	0.96	4.33	-0.14	-1.00
	joint no covariates	11.14	0.46	0.30	0.90	3.12	-0.10	-0.79
	indep. cyclical	109.90	0.57	0.49	1.60	7.19	-0.14	-1.37
	indep. no covariates	83.84	0.67	0.54	1.53	5.36	-0.12	-1.12
	joint DPMM	62.84	0.56	72.22	308.17	618.32	-7.72	-11.44

122.62 for independent cyclical model and $\hat{K} = 96.82$ for independent no covariates model) and the joint DPMM ($\hat{K} = 47.84$) over-estimated the number of hidden states, with Hamming distances ranging from 0.49 for the joint DPMM to 0.61 for the independent no covariates model. In state-specific mean estimation, the independently fit iHMMs outperformed the joint DPMM. The same relative performance of all models existed at 10% missing data. At 20% missing data, the joint cyclical model continued to most accurately estimate the hidden states (Hamming distance = 0.33), followed by the joint no covariates model (Hamming distance = 0.46). In state-specific mean estimation, both joint iHMMs performed similarly and outperformed the independently fit iHMMs and, by far, the joint DPMM.

The slight under-estimation of the number of states using the joint iHMMs is a result of a tendency to merge small states, which may contain only one or two time points, with other states. It is clear from the results that under-estimating the number of hidden states is preferred to over-estimating since our proposed joint approaches had lower Hamming distances and lower MSE for estimated state-specific means than the independently fit iHMMs. The poor estimation performance of the joint DPMM in the presence of missing data demonstrates the importance of including temporal dependency in the modeling framework. The relative improvement of the models with covariates compared to those without covariates demonstrates the value of including covariates in the transition dynamics.

The improved hidden state and state-specific mean estimation in the proposed joint models resulted in more accurate imputations for missing data. At 5% missing data, the joint iHMMs had smallest MSE for both types of imputations (joint cyclical: MAR MSE = 0.46, LOD MSE = 3.01; joint no covariates: MAR MSE = 0.62, LOD MSE = 2.24). The independently fit iHMMs followed. At 10% and 20% missing data, the proposed joint iHMMs had smaller MSE for MAR imputations than the independently fit iHMMs. For below LOD imputations, all iHMMs performed similarly when considering the size of Monte Carlo standard errors (Appendix B.3, Table B.1). The joint DPMM performed worst at all levels of missingness with high MSE for both types of imputations.

In the distinct trends scenario, both our proposed joint cyclical model and the joint no covariates model performed similarly regarding hidden state and state-specific mean estimation (Appendix B.3, Table B.2). Hence, there are minimal drawbacks of including cyclical trends in the model when they are not present in the data. Relative performance of the other models was similar in both scenarios.

3.5 Application to FCCS Data

We applied our proposed method to the FCCS data described in Section 3.2. First, we conducted a validation study to test our multiple imputation approach using holdout observations. We compared variations of our proposed model with different covariates in a situation with an unknown latent structure. Second, we used our proposed method to estimate a hidden state structure in the FCCS data and impute missing observations.

3.5.1 Validation Study

We created 20 data sets for validation. In each data set, we removed an additional 5% of the observed data, which amounted to 2160 observations, split evenly between MAR and below the LOD. We used the same method for removing data as in our simulation study, and specified new LODs at the 0.025 quantiles of the observed data for each exposure. Hence, the additional MAR data was different for each data set, but the data below the LOD was the same for each data set in the validation.

We fit our proposed model with five different specifications for covariates. We fit the joint cyclical model and the joint no covariates model as described in Section 3.4. To account for repeated sampling days, we fit a joint subject-specific cyclical model, which uses the same harmonic function calculated as in Section 3.4 as covariates, as well as subject-specific effects of the harmonic function, as described in (3.4). We also fit a model with the five manually defined microenvironments (home, work, eateries, transit, and other) as categorical predictors, as well as a model with subject-specific effects of the microenvironments, as described in (3.4). We

considered three comparison models: the joint DPMM, a pooled approach, and a stratified approach. In the pooled approach, we fit a single multivariate normal distribution to the entire data set. In the stratified approach, we fit separate multivariate normal distributions to the data within each of the five manually assigned microenvironments. We imputed missing data for the pooled and stratified approaches by sampling from the posterior predictive distributions of the grouped data. To evaluate imputations, we calculated mean MSE and bias over 400 imputations.

Results from our validation study are shown in Table 3.2. All five variations of our proposed model performed similarly and were the best methods for MAR imputations, with mean MSE ranging from 1.20 to 1.39. For the pooled and stratified approaches, mean MSE for MAR imputations was 2.21 and 2.05, respectively. For below LOD imputations, the pooled and stratified approaches had lowest MSE on the majority of the data sets. For our proposed approaches, the minimum MSE for imputations below the LOD ranged from 0.70 to 1.08, with means ranging from 1.94 to 2.58. Meanwhile, the minimum MSE for the pooled and stratified approaches was 1.10 and 1.09, with mean MSE of 1.13 and 1.12, respectively. The joint DPMM had very poor imputations for both types of missing data. For all methods, imputations tended to be negatively biased and more so for below LOD imputations.

Table 3.2: Results from the imputation validation study using FCCS data. The table shows the minimum (min), median, mean, and maximum (max) mean squared error (MSE) for imputations of MAR and below LOD data. The five variations of our proposed joint iHMM approach include the model with no covariates (joint no covariates), the model with cyclical trends (joint cyclical), the model with subject-specific cyclical trends (joint s.s. cyclical), the model with microenvironments as categorical predictors (joint microenv.), and the model with subject-specific microenvironment effects (joint s.s. microenv.) In the pooled approach, a single multivariate normal distribution was fit to all data. In the stratified approach, multivariate normal distributions were fit to all data within each FCCS assigned microenvironment. Last is the Dirichlet process mixture model (joint DPMM) fit jointly to all time series.

	MSE				bias			
	min	median	mean	max	min	median	mean	max
MAR								
joint no covariates	0.93	1.19	1.23	1.70	-0.26	-0.15	-0.15	-0.04
joint cyclical	0.99	1.19	1.26	1.81	-0.20	-0.15	-0.13	-0.08
joint s.s. cyclical	0.90	1.10	1.20	2.34	-0.26	-0.15	-0.14	-0.06
joint microenv.	1.03	1.23	1.28	1.67	-0.23	-0.15	-0.14	-0.06
joint s.s. microenv.	1.00	1.20	1.39	4.07	-0.29	-0.19	-0.17	-0.11
pooled	2.13	2.19	2.21	2.31	-0.23	-0.16	-0.15	-0.01
stratified	1.96	2.04	2.05	2.17	-0.23	-0.16	-0.15	-0.01
joint DPMM	302.83	568.48	613.70	1109.12	-18.05	-12.54	-12.65	-8.17
Below LOD								
joint no covariates	0.71	1.82	1.94	5.11	-1.07	-0.24	-0.30	0.14
joint cyclical	0.84	1.92	2.09	5.26	-0.59	-0.25	-0.23	0.17
joint s.s. cyclical	0.70	1.95	2.14	5.53	-0.97	-0.29	-0.32	0.21
joint microenv.	0.77	1.74	1.96	4.06	-0.98	-0.21	-0.29	-0.01
joint s.s microenv.	1.08	2.03	2.58	10.23	-1.17	-0.53	-0.45	-0.04
pooled	1.10	1.13	1.13	1.19	-0.27	-0.26	-0.26	-0.25
stratified	1.09	1.12	1.12	1.18	-0.27	-0.26	-0.26	-0.25
joint DPMM	663.08	1291.68	1498.04	2970.76	-44.84	-26.72	-27.73	-19.38

These results demonstrate that the hidden state estimation offered by our proposed approach improves imputations for MAR data over naive fixed-state approaches and a DPMM with no temporal structure. For data below the LOD, our results must be interpreted with caution and only in the context of this data set and validation design, since data below the LOD was not randomly generated as in our simulation study. Much of the data below the LOD was clustered within a few sampling days for a long period of time. Imputations of the FCCS data were not sensitive to the covariates specified in our proposed approach. In particular, our models

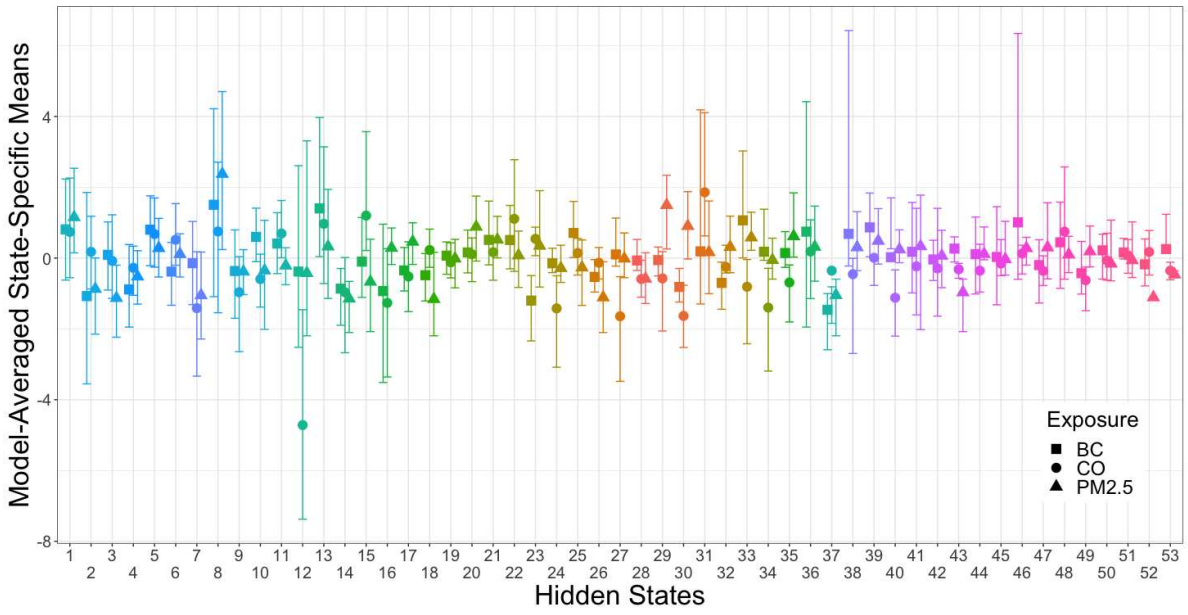
with cyclical trends performed just as well as the models using microenvironments as predictors, suggesting that the microenvironment data, which may be costly to obtain and subject to error, were not necessary for accurate imputations of multivariate exposures.

Our imputation approach was sensitive to the specification of the emission distribution parameter λ. We conducted a sensitivity analysis of the parameter λ included in Appendix B.4. Smaller values of λ led to higher MSE for imputations in our proposed approaches due to larger a priori variation in the data and state-specific means (Tables B.3 and B.4).

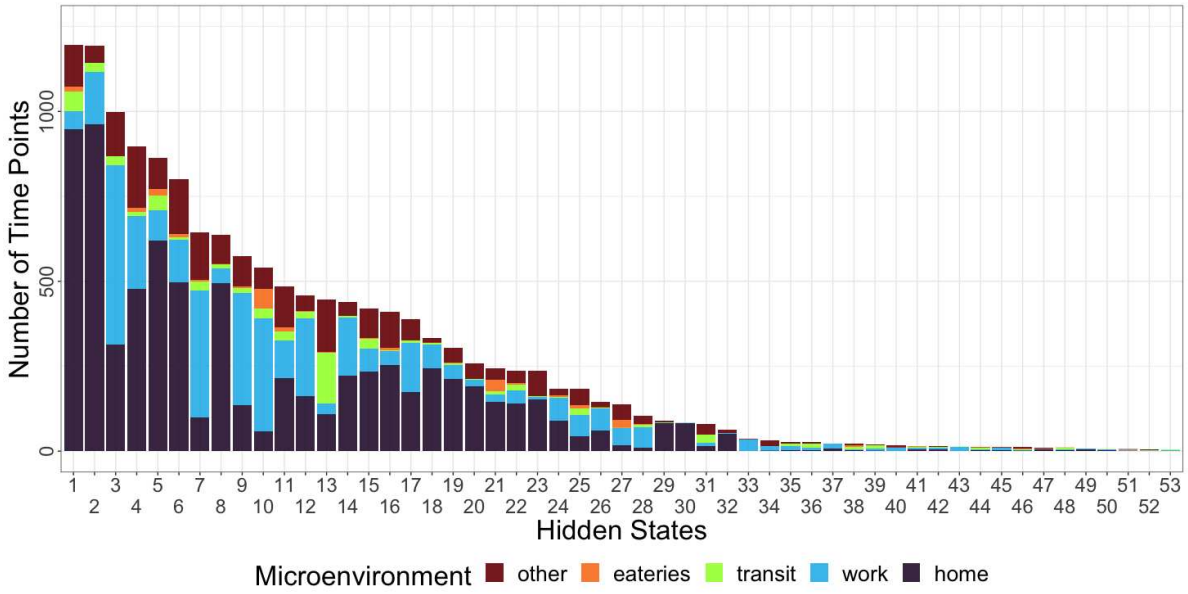
3.5.2 Case Study

We applied the joint subject-specific cyclical model to the FCCS data set described in Section 3.2. Although all variations of covariate structure that we considered in the validation study performed similarly, the joint subject-specific cyclical model best represents our prior belief in the underlying data structure. We based inference on 5000 iterations after discarding 5000 iterations as burn-in. Figures B.4-B.6 in Appendix B.5 show trace plots of the imputations and estimated number of hidden states, demonstrating evidence of convergence within 5000 iterations. The computational time to run our MCMC sampler for 1000 iterations in our application to the FCCS data set was 3.2 hours on a personal laptop (Processor: 3.1 GHz Dual-Core Intel Core i5, Memory: 16 GB 2133 MHz LPDDR3) in R version 4.0.3.

We classified hidden states using the draws-based latent structure optimization method described by Dahl (2006) with the variation of information loss function (Wade and Ghahramani, 2018). Using this method, we estimated 53 hidden states shared across the 50 sampling days. Figure 3.1a shows the model-averaged exposure means for each state, with error bars representing the empirical minimum and maximum exposures within each state. In Figure 3.1b, we show the number of time points assigned to each state and the proportion that lie within each of the five manually assigned microenvironments from the FCCS. Some hidden states were frequently visited and others were relatively rare. Most hidden states encompassed several microenvironments, of which home and work were most frequently visited.



(a) Model averaged state-specific exposure means



(b) Hidden states and microenvironments

Figure 3.1: Results from analysis of FCCS data using joint subject-specific cyclical model. Panel (a) shows model averaged log-transformed exposure means for each of the 53 hidden states estimated in the most optimal partitioning of the FCCS data. Exposures include black carbon (BC), carbon monoxide (CO), and fine particulate matter ($PM_{2.5}$). Error bars depict the minimum and maximum empirical exposures within each state. Panel (b) shows the number of time points in each hidden state and the proportion of time points that intersected with each of the manually assigned microenvironments from FCCS. The total number of time points in this analysis was 14,400.

The hidden states provide opportunity for further investigation of time-activity patterns associated with the exposures. To illustrate this, we investigated in detail hidden state 8 and hidden state 12. Hidden state 8 had higher than average mean exposure for each of three pollutants. By far the most common microenvironment in state 8 was home, followed by other, and then work. In this analysis, 637 time points were assigned to state 8 across 38 sampling days and 9 unique people (Appendix B.6, Table B.5). The presence of this state among many people and days suggests that people frequently experience periods of time when their home microenvironments are subject to higher than average levels of exposure. This state tends to occur around typical breakfast and dinner times and likely corresponds to cooking events. Hidden state 12, on the other hand, had markedly lower than average mean exposure to CO. The 458 time points assigned to hidden state 12 spanned 31 days and 9 unique people (Table B.5), with approximately half of the time points occurring at work and half at home. The very low CO exposure mean for this state suggests that many of the time points assigned to this state may have CO levels below the LOD.

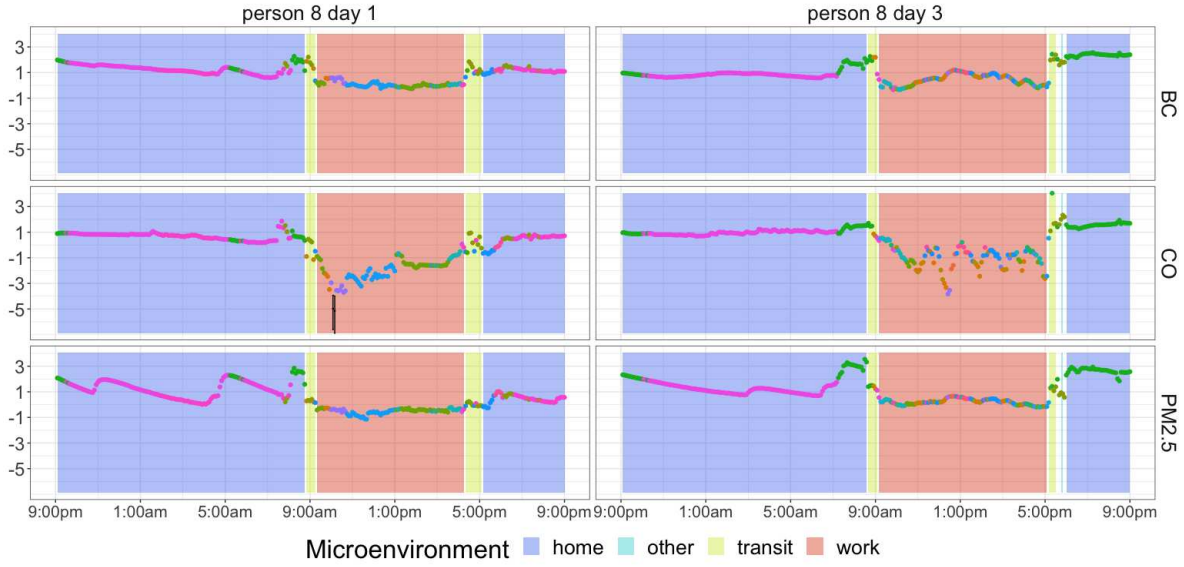
Next we discuss the hidden state trajectories for two sampling days for two separate individuals. We selected two individuals that represent two different patterns in the data: one with very similar exposure patterns and one with different exposure patterns over repeated sampling days. In Figure 3.2a we show the estimated hidden states, reported microenvironments, and imputations for person 8 on sampling days 1 and 3. The left column of panels shows the observed exposures for BC, CO, and $PM_{2.5}$ for person 8 on sampling day 1, and the right column shows the observed exposures for person 8 on sampling day 3. In Figure 3.2b, we show the same for person 37 on sampling days 1 and 2. Microenvironment patterns were similar across all four sampling days shown in Figure 3.2, with individuals generally first spending a large portion of the day at home, followed by a short time in transit, a chunk of time at work, then transit again and ending the day at home. Hidden state change-points generally aligned with microenvironment change-points, showing that our model is able to pick up on changes in activity that

coincide with differences in the distribution of exposures. Our model also subdivides the microenvironments to reflect changing conditions over time.

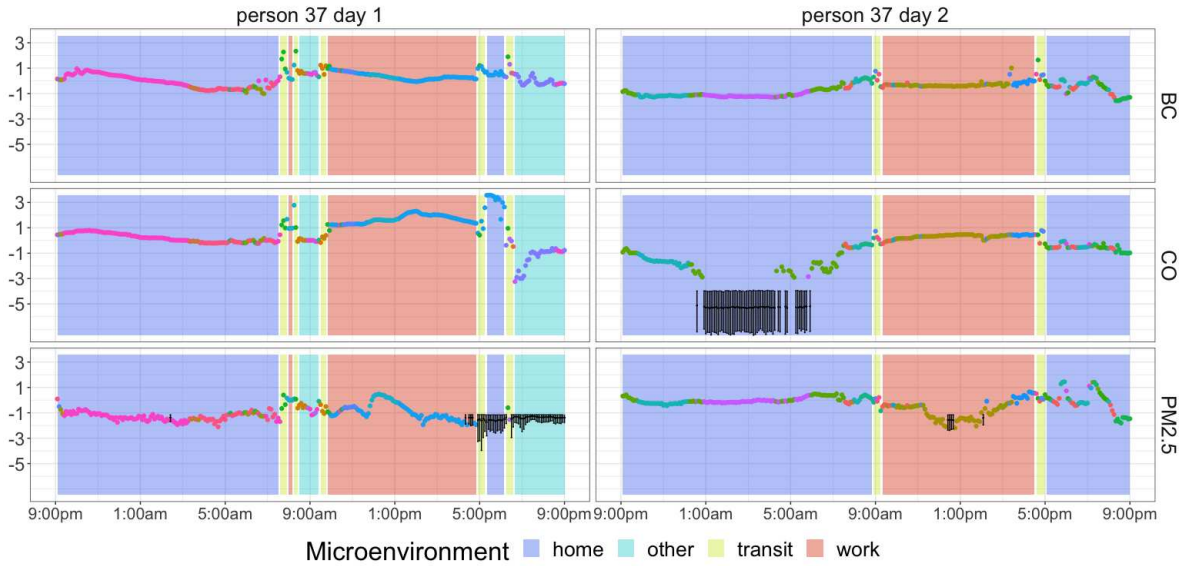
For person 8, similar microenvironment patterns over the two sampling days elicited similar levels of exposure, which our model identified via shared hidden states. On both sampling days shown, person 8 traversed through hidden states 1 and 8 during their time at home. State 8, which had higher than average exposure means for all pollutants, was mostly visited during the evening hours (5-9pm) and mid-morning hours (7-9am). State 1 occurred during the overnight hours. All three exposure means were higher in state 8 than in state 1. It appears that our model is identifying shared activity patterns related to cooking (state 8) and sleeping (state 1), which produce different exposure levels within the same location.

On the contrary, person 37 exhibited differences in exposures between the two sampling days, even within the same microenvironment. Our model captured these differences by estimating different hidden states on the two sampling days for this individual. Person 37 also had a substantial amount of missing data on these two days. All of the imputations shown in Figure 3.2b represent data below the LOD. The time points with CO below the LOD were all assigned to hidden state 12. At these times, $PM_{2.5}$ and BC exposures remained relatively constant. Hence, through the estimation of hidden state 12, our model used the observed data within the state to inform imputations for the long stretch of missing data seen for person 37 on sampling day 2.

Our approach produces a rich output regarding the hidden states, providing plenty of opportunity for further investigation. For example, we estimated several rare hidden states that were present in only one or two subjects. In particular, state 43 was present in only one subject for a total of 13 time points across three days (Table B.5). Figure 3.1 shows that hidden state 43 was defined by slightly lower than average exposure to CO and $PM_{2.5}$, and mainly appeared in the work microenvironment. An uncommon feature of this person's work may have produced this unique distribution of exposures. On the other hand, some hidden states were common among many sampling days, but were only visited for a short period of time each day. One example is hidden state 31, which occurred in 16 sampling days for a total of 81 time points (Table



(a) Person 8 sampling days 1 and 3



(b) Person 37 sampling days 1 and 2

Figure 3.2: Multivariate exposures for two sampling days for each of two individuals in the FCCS data. Panel (a) shows person 8 on sampling days 1 (left panel) and 3 (right panel). Panel (b) shows person 37 on sampling days 1 (left panel) and 2 (right panel). Points represent exposure data with colors determined by the hidden state to which each time point was assigned in the most optimal partitioning of the data. Background colors represent the microenvironments as assigned by the FCCS based on time-activity diaries and GPS data. Black points and associated error bars show the mean imputed values and 95% credible intervals. Time is the local Mountain Daylight Time.

B.5). Hidden state 31 had the highest mean exposure to CO and occurred in the microenvironments home, transit, and other (Figure 3.1). This hidden state suggests that multiple people experience high exposure to CO for a short period of time, which may disproportionately influence daily cumulative exposures. Through visualization of the time series containing hidden states 43 and 31, we can use our model's output to shed light on possible activities associated with rare combinations of exposures as well as short periods of high exposure.

3.6 Discussion

In this paper, we proposed a coherent modeling framework to identify shared exposure patterns and impute missing data in time-resolved ambient pollutant exposure data collected with personal monitors. Our model is a covariate-dependent iHMM for multiple multivariate time series with missing data. We model hidden state transitions with a PSBP, which flexibly allows time-varying covariates and subject-specific effects to inform hidden state transitions and improve imputations. Our model imputes data that are MAR or below the LOD.

In simulation, our approach offers improvements in hidden state estimation and imputation over models fit independently to each time series or a DPMM with no temporal structure. On the FCCS data, our approach best imputes MAR data compared to competing methods.

In our analysis of the FCCS data, we investigated the utility of our proposed approach. In particular, our model can impute missing data for multiple multivariate exposure assessments. To our knowledge, this is the first iHMM developed that can impute data that are both MAR and below the LOD for multiple time series. Accurate imputations are critical in exposure assessments so the data can be reliably used for health effects studies. Additionally, through estimation of the hidden state trajectories, our proposed model can identify both shared and unique states among multiple individuals that correspond to high or low exposures. The estimated hidden states allow us to make inference on time-activity patterns for the individuals in the data set, which can subsequently inform possible interventions.

A limitation of our approach is the challenge of interpreting covariate effects in the PSBP due to the probit transformation, stick-breaking formulation, and possible label switching in the MCMC. While our primary interest was to use covariates to inform hidden state transitions, other methods, such as the mixed HMM (Altman, 2007) or the stick-breaking Pólya-gamma approach (Linderman et al., 2015), could be considered if interest focuses on interpreting covariate effects.

Our work offers a number of promising future directions. First, uncertainty in the LOD and the missing data classification could be accommodated by estimating the LOD and modeling the missing data type with a binary variable, respectively. Second, while our method was developed to cluster time points, extensions may consider hierarchical clustering of sampling days or subjects. Clustering sampling days would provide insights into weekly or seasonal patterns in exposures, while clustering subjects may elucidate individual- or group-level activities related to exposures. Third, the method could be extended to accommodate continuous time series and non-Gaussian emissions. With the rapid increase in the use of personal monitors in studies of air pollution exposure and health, methods such as we proposed in this paper, as well as these potential extensions, are essential to maximize the information researchers can obtain from these data.

Chapter 4

Association Between Air Pollution and COVID-19 Disease Severity via Bayesian Multinomial Logistic Regression with Partially Missing Outcomes

4.1 Introduction

Ambient air pollution exposure is a major global environmental health concern (Global Burden of Diseases 2019 Risk Factors Collaborators, 2020; Health Effects Institute, 2018). Long-term exposure to air pollution is associated with increased rates and severity of chronic diseases including cardiovascular disease, diabetes, asthma, chronic obstructive pulmonary disease, and mortality (Di et al., 2017a,b; Dockery and Pope, 1994; Dockery et al., 1993; Pan et al., 2018). In addition, poor air quality has a negative impact on infectious diseases, and has been linked to increased rates of influenza (Landguth et al., 2020) and increased fatalities from sudden acute respiratory syndrome (SARS) (Cui et al., 2003). Previous evidence indicates long-term exposure to air pollution increases susceptibility to viral disease, leading to more severe outcomes (Ciencewicki and Jaspers, 2007). It is hypothesized that air pollution exposure may be linked to increased severity in the ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by the novel coronavirus SARS-CoV-2 (Comunian et al., 2020; Domingo and Rovira, 2020; Frontera et al., 2020; Setti et al., 2020b). Similar biological pathways that have been observed with influenza and other respiratory viral infections may exist between exposure to particulate matter and SARS-CoV-2 infection, highlighting the possibility of increased COVID-19 severity among individuals with higher exposure to air pollution (Frontera et al., 2020).

The study of the effects of air pollution on COVID-19 health endpoints has been identified as a critically important area of research for developing solutions to the global COVID-19 pandemic (Bhaskar et al., 2020). Studies investigating this relationship have considered exposures

such as air quality index, fine and coarse particulate matter, nitrogen oxides, ozone, carbon monoxide, and sulfur dioxide, as well as meteorological parameters including temperature and relative humidity. In two literature reviews of studies taking place world-wide, a majority of articles identified significant associations between short- and long-term exposure to air pollution and negative COVID-19 endpoints (Bhaskar et al., 2020; Copat et al., 2020). The COVID-19 endpoints varied among studies and included number of cases, number of deaths, case-hospitalization rate, case-fatality rate, percent of severe infection, basic reproduction number, intensive care unit (ICU) admissions, and epidemic escalation. In addition, emerging cohort studies suggest long-term exposure to air pollution prior to the pandemic is associated with a higher risk of severe COVID-19 in those infected with SARS-CoV-2 (Bozack et al., 2021; Kogevinas et al., 2021).

The vast majority of existing studies used ecological designs with aggregated, most commonly county-level, data. Ecological studies suffer from ecological fallacy; that is, characteristics of the group cannot be attributed to individuals. In their review, Brandt and Mersha (2021) emphasized the need for individual-level air pollution exposure data and detailed clinical data to establish a causal relationship between air pollution exposure and COVID-19 outcomes. Individual-level exposure and risk factor data are needed to minimize bias and potential confounding that can occur at larger spatial resolutions. In addition, health endpoints for COVID-19 that are measured at the individual level are more accurate than regional endpoints, which may be subject to variations among regions or error due to unmeasured asymptomatic cases and under-reporting of cases and deaths. The current literature is sparse with regards to individual-level studies on the association between air pollution exposure and COVID-19 outcomes.

We conduct an individual-level analysis of the association between long-term exposure to air pollution and weather and peak COVID-19 severity in a Denver, Colorado, USA administrative cohort. We consider all cases of COVID-19 that were reported to the Colorado Department of Public Health and Environment (CDPHE) between March 6, 2020 and February 28, 2021, re-

sulting in a cohort size of 57,027 verified COVID-19 infections. As the primary health outcome, we consider peak severity. Our peak severity outcome takes on one of six mutually exclusive categorical values: asymptomatic, symptomatic, hospitalized, admitted to the ICU, placed on a mechanical ventilator, or death. Our primary interest is estimating the association between long-term exposure to ambient air pollution and weather and peak COVID-19 severity.

A key challenge when using individual-level administrative data, especially in the rapidly evolving COVID-19 pandemic, is the presence of missing health outcomes. Individual health outcomes may be missing due to non-response or logistical problems with data attainment. In the Denver, Colorado cohort, health outcomes are either observed or partially missing. For example, it may be known that an individual was not hospitalized or worse, but it is unknown whether the individual was symptomatic or asymptomatic. Observations with partially missing outcomes are often discarded prior to a complete case analysis; however, there is valuable information to gain from the partially missing observations. Hence, there is a need for statistical methods for regression analysis of data with partially missing categorical outcomes.

In classical statistics, multiple imputation approaches for categorical outcome data include nearest-neighbor based methods (Zhou et al., 2017), bootstrap hotdeck multiple imputation (Wang and Hsu, 2020), inverse probability weighting, and expected estimating equations. These methods generally require discrete covariates, though continuous covariates can be incorporated through discretization. In Bayesian statistics, missing data are handled naturally by sampling from the posterior predictive distribution of the missing data given the observed data. Currently, however, there are no fully Bayesian approaches for multinomial logistic regression with missing outcome data.

In this paper, we propose a Bayesian multinomial logistic regression model for data that contain observations with partially missing categorical outcomes. Fully Bayesian inference in categorical and multinomial regression has been historically challenging due to non-conjugate priors for the model's likelihood. In our analysis, we base inference on the odds ratio for each peak severity category; hence, we require a logit link function. Polson et al. (2013) proposed a

Pólya-gamma data augmentation approach for Bayesian logit models, and extended the Pólya-gamma approach to multinomial models by combining it with the data augmentation approach from Holmes and Held (2006). This approach requires sampling the category-specific regression coefficients one at a time, which can cause slow mixing and convergence in correlated models. To address this issue, Linderman et al. (2015) proposed modeling the multinomial distribution recursively with binomial distributions via the stick-breaking representation. The stick-breaking approach permits parallelized updates of the regression parameters, leading to more efficient mixing. Though the stick-breaking approach offers computational improvements, it presents an inferential challenge because the odds ratio ceases to be a linear function of the exposures.

We develop the first fully Bayesian multinomial logistic regression model for partially missing outcome data in which the primary goal is inference on the odds ratios. Our method builds on the approach of Linderman et al. (2015), and we address the inferential challenges induced by the stick-breaking approach through post-processing and visualization of the posterior distribution. Our model imputes partially missing health outcome data, where the number of missing outcome categories can vary by individual. Using the proposed model, we estimate the association between long-term exposure to fine particulate matter, ozone, and temperature and peak COVID-19 severity in the presence of missing outcome data, while controlling for individual- and neighborhood-level risk factors. We find evidence of a positive association between exposure to fine particulate matter and increased risk of severe COVID-19, as well as a possible interaction effect between fine particulate matter and ozone. Our individual-level analysis supports existing research on air pollution and COVID-19, and provides the additional contribution of beginning to draw a causal link.

4.2 Data

4.2.1 Health Data

We obtained health outcome data from Denver Public Health, a department of Denver Health and Hospital Authority (DHHA). Our study population includes 57,027 laboratory-confirmed cases of COVID-19 in the City and County of Denver, Colorado reported between March 6, 2020 and February 28, 2021. The data include information about the case status including if an individual was symptomatic, hospitalized, admitted to an ICU, placed on a mechanical ventilator, or died. The case outcome data had missing observations, primarily due to lack of staff capacity to follow-up with cases regarding disease outcomes. Hence, the missing mechanism was assumed to be missing at random. We made two assumptions to deterministically fill in some of the missing outcome data. First, since deaths were accurately reported to the City and County of Denver, we assumed that a case with missing death status did not die. Second, we assumed that a case that was not symptomatic was not hospitalized, a case that was not hospitalized was not admitted to the ICU, and a case that was not admitted to the ICU was not placed on a mechanical ventilator. After deterministically imputing missing outcome data using these basic assumptions, we assigned each case to its most severe outcome. When peak severity could not be determined for an individual due to missing data, all possible peak severity outcome categories were left as missing and imputed by our model. Table 4.1 shows the resulting pattern of missingness in the data.

Table 4.1: Missing data pattern for the peak severity outcome categories in our analysis of the Denver, Colorado cohort (n = 55273). Cases with partially missing outcomes were missing between 2 and 5 outcome categories. The table shows the number and percent of cases with each missing outcome category pattern.

# missing outcomes	missing categories	# cases	% cases
0	–	20872	37.8
2	(Asymptomatic, Symptomatic)	2916	5.3
3	(Symptomatic, Hospitalized, ICU)	59	0.1
4	(Symptomatic, Hospitalized, ICU, Ventilator)	8725	15.8
5	(Asymptomatic, Symptomatic, Hospitalized, ICU, Ventilator)	22701	41.1

4.2.2 Exposure Data

We obtained air pollutant and meteorological exposure data from the Colorado Department of Public Health and Environment (CDPHE) website (Department of Public Health and Environment, 2021). The exposure metric of interest was annual average exposure to fine particulate matter with an aerodynamic diameter less than 2.5 μm ($PM_{2.5}$; $\mu g/m^3$), ozone (ppb), and temperature (degrees Fahrenheit) in the year prior to the COVID-19 pandemic in Denver, Colorado. The first officially documented case of COVID-19 in Denver was March 6, 2020. We therefore define the year prior to the pandemic, our exposure period, as March 1, 2019 through February 29, 2020. We calculated daily average exposure during the exposure period for $PM_{2.5}$ and temperature, and 1-hour maximum daily average for ozone from hourly measurements recorded at ground monitoring stations. We excluded daily variables if more than 25% of hourly observations recorded at that monitoring site for that day were missing. We excluded monitors that were located in the Rocky Mountains west of Denver because that area experiences unique meteorological conditions not representative of the study area (Vedal et al., 2009). Using inverse-distance weighting, we assigned daily exposures to individual residential locations using data from all monitors within 50km of the individual's address. Finally, we calculated each individual's annual average exposure during the year prior to the pandemic by averaging the daily

exposure values. For our analyses, exposure data were centered and divided by the interquartile range (IQR) prior to model fitting.

4.2.3 Covariate Data

We included individual- and census tract-level variables. We obtained individual-level variables from Denver Public Health's COVID-19 case investigation database. These variables included the continuous covariate age and the categorical covariates gender, race/ethnicity, and pregnancy status. We also included case report date, defined as the date the case was first reported to CDPHE. The individual-level covariate data contained a small number of missing observations. To impute missing categorical covariate data, we first assumed that if the case was listed as male then the case was not pregnant. We then singly imputed the missing values for categorical covariates with 0 and added a dummy variable for each covariate with missing data that indicated which values of the covariate were missing. For gender, a value of 'other' was combined with the missing group due to the small number in the 'other' group ($n = 3$).

We obtained census-tract variables summarizing socioeconomic status from the 2015-2019 American Community Survey (United States Census Bureau, 2020) using the `tidycensus` package in R (Walker et al., 2021). These variables included median income, percent of the civilian workforce aged 16 and older that is unemployed (unemployed), percent of the population aged 25 and older with less than a high school diploma or equivalent education (low education), and percent of individuals in the census tract with past year's income below the poverty level (poverty).

We obtained a final sample size of 55,273 individuals for which we were able to link the health outcome data with complete covariate and exposure data. We provide a summary of the demographic characteristics of the sample in Appendix C.1 (Table C.1). This study was approved by the Institutional Review Board of Colorado State University.

4.3 Statistical Model

We first present the model for complete data. Then we describe our Markov chain Monte Carlo (MCMC) sampler, our multiple imputation approach, and our inferential approach. Software to fit our proposed method is in the `R` package `pgmultinomr` (Hoskovec, 2021a) available at github.com/lvhoskovec/pgmultinomr.

4.3.1 Complete Data Model

For a sample $i = 1,\ldots,n$, let $\mathbf{y}_i$ denote the K-dimensional vector indicating to which of K possible outcome categories individual i belongs. Hence, $\mathbf{y}_i = (y_{i1},\ldots,y_{iK})$ contains one 1 and the remaining $K-1$ observations are all 0. Let $\mathbf{x}_i$ denote the vector of exposures for individual i. In our analysis, $\mathbf{x}_i$ contains three exposures and all pairwise interactions. Let $\mathbf{w}_i$ denote the vector of covariates measured for individual i, including an intercept term.

We model $\mathbf{y}_i$ with a multinomial distribution where the number of trials is 1. Using the stick-breaking representation of the multinomial distribution, we model the complete data for individuals $i = 1,\ldots,n$ by

$$\mathbf{y}_i \sim \prod_{k=1}^{K-1} \text{binom}(y_{ik}|N_{ik},\tilde{\pi}_{ik}), \tag{4.1}$$

where $N_{i1} = 1$ and $N_{ik} = 1 - \sum_{j<k} y_{ij}$. In (4.1), $\tilde{\pi}_{ik}$ for $k = 1,\ldots,K-1$ are the stick-specific weights for individual i, denoting the proportion of the remaining probability mass assigned to the k^{th} category. The parameter N_{ik} denotes the number of remaining trials for the k^{th} category, which, in our case, will always be either 0 or 1. We model each $\tilde{\pi}_{ik}$ for $i = 1,\ldots,n$ and $k = 1,\ldots,K-1$ using a logit link function of exposures and covariates. The logit link for the stick-specific weights is given by

$$\tilde{\pi}_{ik} = \frac{\exp(\psi_{ik})}{1+\exp(\psi_{ik})} \tag{4.2}$$

$$\psi_{ik} = \mathbf{x}_i^T\boldsymbol{\beta}_k + \mathbf{w}_i^T\boldsymbol{\gamma}_k, \tag{4.3}$$

where $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ are category-specific regression coefficients for the exposures and covariates, respectively. The model in (4.1) is equivalent to the standard multinomial model

$$\mathbf{y}_i \sim \text{multinom}_K(1, \boldsymbol{\pi}_i), \tag{4.4}$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$ and

$$\begin{aligned} \pi_{i1} &= \tilde{\pi}_{i1} \\ \pi_{ik} &= \tilde{\pi}_{ik}\left(1 - \sum_{j<k} \pi_{ij}\right) \quad \text{for } k = 2, \ldots, K-1 \\ \pi_{iK} &= 1 - \sum_{k=1}^{K-1} \pi_{ik}. \end{aligned} \tag{4.5}$$

To achieve efficient Gibbs sampling of the posterior distribution, we implement a Pólya-gamma data augmentaton scheme (Linderman et al., 2015; Polson et al., 2013). We introduce latent Pólya-gamma random variables ω_{ik} for $i = 1, \ldots, n$ and $k = 1, \ldots, K-1$ such that $\omega_{ik} \sim \text{PG}(N_{ik}, 0)$, where $\text{PG}(\cdot, \cdot)$ denotes the Pólya-gamma distribution. Using the stick-breaking representation and Pólya-gamma augmentation, the multinomial likelihood for individual i can be written as

$$\begin{aligned} f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \prod_{k=1}^{K-1} \binom{N_{ik}}{y_{ik}} \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_k + \mathbf{w}_i^T \boldsymbol{\gamma}_k)^{y_{ik}}}{\left[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}_k + \mathbf{w}_i^T \boldsymbol{\gamma}_k)\right]^{N_{ik}}} \\ &= \prod_{k=1}^{K-1} \binom{N_{ik}}{y_{ik}} \frac{\exp\left[\kappa_{ik}\left(\mathbf{x}_i^T \boldsymbol{\beta}_k + \mathbf{w}_i^T \boldsymbol{\gamma}_k\right)\right]}{2^{N_{ik}}} \text{E}_{\omega_{ik}}\left[\exp\left\{-\frac{1}{2}\omega_{ik}\left(\mathbf{x}_i^T \boldsymbol{\beta}_k + \mathbf{w}_i^T \boldsymbol{\gamma}_k\right)^2\right\}\right], \end{aligned} \tag{4.6}$$

where $\kappa_{ik} = y_{ik} - N_{ik}/2$ and $\text{E}_{\omega_{ik}}(\cdot)$ denotes the expectation taken with respect to the Pólya-gamma random variable ω_{ik}. By conditioning (4.6) on $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iK-1})$, we obtain

$$f(\mathbf{y}_i | \boldsymbol{\omega}_i, \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \prod_{k=1}^{K-1} \exp\left[-\frac{1}{2}\omega_{ik}\left\{\mathbf{x}_i^T \boldsymbol{\beta}_k + \mathbf{w}_i^T \boldsymbol{\gamma}_k - \left(\frac{\kappa_{ik}}{\omega_{ik}}\right)\right\}^2\right], \tag{4.7}$$

which is a Gaussian kernel with respect to the regression coefficients. The prior distributions on the regression coefficients are $\boldsymbol{\beta}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\gamma}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ for $k = 1, \ldots, K-1$, which allow for efficient Gibbs sampling of the posterior distribution.

4.3.2 Posterior Computation

In our model, the number of trials is 1. We sample ω_{ik} only for those individuals i and categories k such that $N_{ik} = 1$. The full conditional for ω_{ik} is

$$\omega_{ik} | N_{ik} = 1, \cdot \quad \sim \quad \mathrm{PG}(1, \mathbf{x}_i^T \boldsymbol{\beta}_k + \mathbf{w}_i^T \boldsymbol{\gamma}_k). \tag{4.8}$$

The regression coefficients for the k^{th} category only depend on data from individuals where $N_{ik} = 1$. For $k = 1, \ldots, K-1$, let $\mathbf{X}_k$ be a matrix with rows $\mathbf{x}_i$ and $\mathbf{W}_k$ be a matrix with rows $\mathbf{w}_i$ for each i such that $N_{ik} = 1$. We sample the exposure regression coefficients from

$$\begin{aligned} \boldsymbol{\beta}_k | \cdot \quad &\sim \quad \mathrm{N}(\mathbf{m}_k, \mathbf{V}_k) \\ \mathbf{V}_k \quad &= \quad \left(\mathbf{I} + \mathbf{X}_k^T \boldsymbol{\Omega}_k \mathbf{X}_k \right)^{-1} \\ \mathbf{m}_k \quad &= \quad \mathbf{V}_k \left[\mathbf{X}_k^T \boldsymbol{\Omega}_k \left(\mathbf{z}_k - \mathbf{W}_k^T \boldsymbol{\gamma}_k \right) \right], \end{aligned} \tag{4.9}$$

where $\boldsymbol{\Omega}_k$ is a diagonal matrix with elements ω_{ik} for each i such that $N_{ik} = 1$ and $\mathbf{z}_k$ is a vector of κ_{ik}/ω_{ik} for each i such that $N_{ik} = 1$. The regression coefficients for the covariates are similarly updated.

4.3.3 Multiple Imputation

We assume missing outcome data are missing at random (Little and Rubin, 2019), but may be conditional on partial outcome information. The complete data vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$ may contain any combination of observed and missing values for the K categories. If the vector contains missing data, then any observed values must be 0, since the total number of trials is 1.

To impute missing outcome data, we sample from the posterior predictive distribution of the missing data given the observed data.

Consider an individual i with missing outcome data. Let $\mathbf{y}_{i,\text{miss}}$ denote the set of outcome categories with missing data and $\mathbf{y}_{i,\text{obs}}$ the set of outcomes categories that are observed for individual i. If $\mathbf{y}_{i,\text{obs}} = \{\emptyset\}$ (i.e. the individual is missing outcome data for all K categories), then the posterior predictive distribution is

$$\mathbf{y}_{i,\text{miss}}|\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \text{multinom}_K(1, \boldsymbol{\pi}_i), \tag{4.10}$$

where $\boldsymbol{\pi}_i$ is calculated by (4.2), (4.3), and (4.5).

If some outcome categories for individual i are observed and some are missing, then we leverage the partial information to improve imputations. Let $\mathcal{K}_{i,\text{miss}}$ denote the set of categories with missing data for individual i and let $\mathcal{K}_{i,\text{obs}}$ denote the set of categories that are observed. Note that $\mathcal{K}_{i,\text{miss}} \cup \mathcal{K}_{i,\text{obs}} = \{1, \ldots, K\}$. For partially missing outcomes in our analysis, the number of missing outcome categories may range from 2 to 5. Since we know the observed outcome categories must all be 0, we can sample the missing outcome categories from a reduced dimensional multinomial distribution. First, we calculate the entire probability vector $\boldsymbol{\pi}_i$ by (4.2), (4.3), and (4.5). Then we calculate $\boldsymbol{\pi}_{i,\text{miss}} = \{\pi_{ik,\text{miss}} : k \in \mathcal{K}_{i,\text{miss}}\}$ where

$$\pi_{ik,\text{miss}} = \frac{\pi_{ik}}{\sum_{k' \in \mathcal{K}_{i,\text{miss}}} \pi_{ik'}}. \tag{4.11}$$

Finally, we sample the missing outcome categories from

$$\mathbf{y}_{i,\text{miss}}|\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \text{multinom}_{|\mathcal{K}_{i,\text{miss}}|}(1, \boldsymbol{\pi}_{i,\text{miss}}), \tag{4.12}$$

where $|\mathcal{K}_{i,\text{miss}}|$ is the number of missing outcome categories for individual i.

4.3.4 Inference

The traditional multinomial model (4.4) exhibits straightforward inference on the regression coefficients, assuming the model only includes main effects. The exponentiated regression coefficient $\exp(\beta_{jk})$ is the odds ratio for category k relative to the reference category that is associated with a one-unit increase in exposure j, holding all other exposures constant. When interactions are included, the interpretation of regression coefficients in the traditional multinomial model is complicated. Unless all co-exposures are set to zero, in which case inference simply ignores interactions, it is impossible to increase an exposure while holding constant an interaction term containing that exposure.

The stick-breaking representation of the multinomial distribution also presents challenges in interpreting the regression coefficients because $\text{logit}(\psi_{ik})$ for each k is conditional on not being in any category $k' < k$. In the stick-breaking model, $\exp(\beta_{jk})$ is the odds ratio for category k relative to a category greater than k, conditional on not being in a category less than k, that is associated with a one-unit increase in exposure j, holding all other exposures constant. Not only is the stick-breaking interpretation difficult to comprehend, it also heavily depends on the ordering of the k categories since the reference is to a category greater than k. The stick-breaking model presents the same problem with interpreting interactions as the traditional multinomial model. We propose a visualization approach for inference on the stick-breaking multinomial model to address these problems.

Due to the fully Bayesian nature of our model, we use the posterior distribution of the regression coefficients to recover the traditional odds ratio inference that is common in logistic regression. We consider the odds ratio as a function of exposures, and set all covariates to 0 in our calculations.

Let $\theta^{(s)}$ denote the sampled value for the parameter θ at iteration s of the MCMC sampler. For iterations $s = 1, \ldots, S$ post burn-in, we calculate the posterior distribution of the assignment

probabilities, $P(y=k|\mathbf{x})$, for each category, by

$$\begin{aligned}
\dot{\pi}_k^{(s)} &= \frac{\exp\left(\mathbf{x}^T\boldsymbol{\beta}_k^{(s)}\right)}{1+\exp\left(\mathbf{x}^T\boldsymbol{\beta}_k^{(s)}\right)}, \quad \text{for } k=1,\ldots,K-1 \qquad (4.13)\\
P(y=1|\mathbf{x})^{(s)} &= \dot{\pi}_1^{(s)}\\
P(y=k|\mathbf{x})^{(s)} &= \dot{\pi}_k^{(s)}\left(1-\sum_{j<k}P(y=j|\mathbf{x})^{(s)}\right), \quad \text{for } k=2,\ldots,K-1\\
P(y=K|\mathbf{x})^{(s)} &= 1-\sum_{k=1}^{K-1}P(y=k|\mathbf{x})^{(s)}.
\end{aligned}$$

With this method, any category may be selected as the reference category. Let k^* denote the selected reference category. We calculate the posterior distribution of the odds ratio (OR) for specified exposure values $\mathbf{x}^*$ relative to baseline exposure values $\mathbf{x}_0$ by

$$\widehat{\mathrm{OR}}^{(s)}(\mathbf{x}^*,\mathbf{x}_0)=\left[\frac{P(y=k|\mathbf{x}=\mathbf{x}^*)^{(s)}}{P(y=k^*|\mathbf{x}=\mathbf{x}^*)^{(s)}}\right]\Bigg/\left[\frac{P(y=k|\mathbf{x}=\mathbf{x}_0)^{(s)}}{P(y=k^*|\mathbf{x}=\mathbf{x}_0)^{(s)}}\right], \qquad (4.14)$$

for all $k\neq k^*$. In our analysis, we consider three exposures and their pairwise interactions. To create the matrix $\mathbf{x}^*$ in (4.14), we generate a sequence of evenly-spaced exposure values within the mean plus or minus two IQR for a primary exposure of interest, and set the two secondary exposures to a fixed percentile. The pairwise interactions are then calculated and all six exposure variables (3 main effects and 3 interactions) are included in $\mathbf{x}^*$. The baseline exposure matrix, $\mathbf{x}_0$, includes the primary exposure set to its mean value, the other two secondary exposures set to the specified percentiles, and the pairwise interactions. Hence, the OR will always be 1 at the mean value of the primary exposure. To visualize the posterior distribution of $\widehat{\mathrm{OR}}(\mathbf{x}^*,\mathbf{x}_0)$, we plot the posterior mean and 95% credible intervals as a function of the primary exposure, holding the secondary exposures at the same fixed percentile. We visualize interaction effects by plotting $\widehat{\mathrm{OR}}(\mathbf{x}^*,\mathbf{x}_0)$ for different percentiles of the secondary exposures. We repeat this procedure three times so each of the three exposures included in our analysis is used as the primary exposure.

In some situations, the peak severity outcomes follow a logical order. Such is the case in our analysis with outcomes: asymptomatic, symptomatic, hospitalized, admitted to the ICU, placed on a mechanical ventilator, and death. In these situations, the ordinal regression model may seem appropriate. However, ordinal regression requires the strong assumption that the odds ratio for being in a category less than or equal to k, relative to being in a category greater than k, is the same for all categories. Ordinal regression also presents the same interpretation problems for interaction effects as discussed previously for traditional multinomial regression. Hence, we utilize the flexibility and Bayesian nature of our model to estimate the incremental odds ratio (IOR), which is interpreted as the odds ratio of being in category k relative to being in any of the less severe categories. Following a similar approach as we did for the OR, we calculate the IOR by

$$\widehat{\text{IOR}}^{(s)}(\mathbf{x}^*, \mathbf{x}_0) = \left[\frac{P(y=k|\mathbf{x}=\mathbf{x}^*)^{(s)}}{P(y<k|\mathbf{x}=\mathbf{x}^*)^{(s)}}\right] \Bigg/ \left[\frac{P(y=k|\mathbf{x}=\mathbf{x}_0)^{(s)}}{P(y<k|\mathbf{x}=\mathbf{x}_0)^{(s)}}\right], \tag{4.15}$$

for the ordered outcome categories $k = 1, \ldots, K$.

4.4 Simulation Study

4.4.1 Simulation Study Design and Evaluation Metrics

We conducted a simulation study to evaluate the proposed method's performance at imputing missing outcome data and estimating regression coefficients. We considered eight simulation scenarios that vary 1) the proportion of observations in each outcome category, 2) whether or not exposures and covariates are predictive of the outcome categories, and 3) whether there are partially missing outcomes or fully missing outcomes. For each scenario, we compared the proposed model to a complete case analysis estimated with a similar Pólya-gamma augmented stick-breaking model using the same priors on the regression coefficients.

For all eight scenarios, we generated exposure data $\mathbf{x}_i$ for $i = 1, \ldots, n$ with $p = 3$ components from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{\Sigma}_X = \begin{bmatrix} 1 & -0.13 & 0.02 \\ -0.13 & 1 & -0.86 \\ 0.02 & -0.86 & 1 \end{bmatrix}, \tag{4.16}$$

where $\mathbf{\Sigma}_X$ is the correlation matrix of the real exposure data. We generated covariate data $\mathbf{w}_i$ for $i = 1 \ldots, n$ with $q = 5$ components from independent standard normal distributions. We simulated outcome data with $K = 6$ categories and used a sample size of $n = 5000$.

Scenarios 1-4 encompassed the “data probabilities" setting, in which we set the category-specific intercepts so the outcome category sizes mimicked the complete cases of the real data as much as possible. In scenarios 5-8, termed the “equal probabilities" setting, we set the intercepts so outcome categories were approximately equal-sized. Outcome categories in the equal probabilities setting were not exactly equal-sized due to the randomness in the data generating process. Rather, the equal probabilities setting provides a setting in which all six categories have a substantial amount of data, on average roughly equal amounts, and there are no very small or very large categories as is the case in the data probabilities setting. The intercepts were appended to the covariate matrix $\mathbf{w}$. The outcome category assignment proportions for each scenario, as well as the true proportions for the complete case data, are shown in Table 4.2.

Table 4.2: Classification probabilities into each of the 6 outcome categories. The table shows the outcome probabilities for the complete cases of the real data ("real data") and for the complete data in our simulation scenarios. Measures for the simulated data were taken from 500 simulated data sets. Table shows the mean (minimun, maximum) classification probabilities for scenarios with a signal, and the fixed classification probabilities for null scenarios. Classification probabilities for null scenarios did not differ among the simulated data sets. Probabilities are shown for both the "data probabilities" and "equal probabilities" simulation design settings.

		data probabilities		equal probabilities	
	real data	signal	null	signal	null
Symptomatic	0.76	0.71 (0.64, 0.81)	0.77	0.19 (0.09, 0.29)	0.14
Asymptomatic	0.15	0.16 (0.08, 0.25)	0.16	0.18 (0.09, 0.27)	0.16
Hospitalized	0.06	0.06 (0.01, 0.14)	0.05	0.15 (0.06, 0.24)	0.19
ICU	0.01	0.03 (0.01, 0.09)	0.01	0.16 (0.07, 0.26)	0.19
Ventilator	0.01	0.02 (0.01, 0.05)	0.01	0.14 (0.07, 0.25)	0.14
Death	0.01	0.01 (0.01, 0.07)	0.01	0.18 (0.07, 0.28)	0.18

We designed scenarios both with and without a signal from the data to determine the effect of a signal on imputations. In scenarios with a signal, exposure and covariate regression coefficients ($\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$, respectively) for categories $k = 1, \ldots, K-1$ were simulated from independent standard normal distributions. In scenarios without a signal (null scenarios), all exposure and regression coefficients were set to 0, with the exception of the intercepts, which were specified to dictate outcome category sizes. In all scenarios, we let $\psi_{ik} = \mathbf{x}_i^T \boldsymbol{\beta_k} + \mathbf{w}_i^T \boldsymbol{\gamma}_k$ and then generated outcome data according to the stick-breaking representation of the multinomial distribution.

We considered missing data levels of 0%, 20%, 50%, and 80%. Each missing data level reflects the percent of cases that have some level of uncertainty in the outcome. The cases with missing outcome data were randomly selected in each simulated data set. We considered "partially missing" outcomes and "fully missing" outcomes. Under partially missing outcomes, a case with missing outcome data was missing anywhere between 2 and 5 outcome categories. The true outcome was always included as one of the missing outcome categories. We randomly selected the additional outcome categories, drawing the number of additional missing outcome

categories (1 to 4) uniformly. Under fully missing outcomes, all cases with missing outcome data were missing data for all 6 outcome categories.

Performance was based on 500 simulated data sets for each scenario and missing data level. We evaluated estimation of the exposure and covariate regression coefficients through root mean squared error (RMSE), bias, coverage of the 95% posterior credible intervals (CI), and CI width, averaged over all regression coefficients. To evaluate imputations, we calculated precision (the proportion of outcomes assigned to a category that truly belong in that category) and recall (the proportion of outcomes that truly belong in a category that were assigned to that category) for each outcome category. We compared our method's estimation performance to a complete case analysis in each of the eight simulation scenarios.

4.4.2 Simulation Study Results

We summarized simulation results for estimation of the exposure regression coefficients in the data probabilities setting in Table 4.3 and in the equal probabilities setting in Table 4.4. Results for covariate regression coefficients are available in Appendix C.2 (Tables C.2 and C.3). We presented precision and recall for 80% missing data in the data probabilities setting in Table 4.5 and in the equal probabilities setting in Table 4.6. Precision and recall for 20% and 50% missing data were similar and are available in Appendix C.2 (Tables C.4 and C.5). For each of Tables 4.3 - 4.6, the four scenarios within each of the two simulation settings reflect the four combinations of the data (providing a signal or being null) and the missing mechanism of the outcomes (partially or fully missing). Hence, the scenarios were termed "partially missing, signal," "fully missing, signal," "partially missing, null," and "fully missing, null."

Both our proposed method and the complete case analysis produced unbiased estimates for the regression coefficients. However, our proposed method resulted in lower variance estimates of the regression coefficients, exhibited by lower RMSE, smaller CI width, and maintenance of the nominal coverage level (0.95). Hence, by retaining the full data set and imputing missing outcomes, we obtained more efficient inference over a complete case only analysis. Further es-

timation gains were achieved through improvements in the imputations, which occurred when there was partial outcome information or larger category sizes.

When outcomes were partially missing, as opposed to fully missing, our method leveraged the information available to improve imputations. In each case, partially missing outcomes resulted in higher precision and recall over fully missing outcomes, controlling for other scenario factors (Tables 4.5 and 4.6). For example, looking at the scenarios with a signal in the data probabilities setting (Table 4.5), precision and recall in category 1, the largest category, were 0.92 when outcomes were partially missing versus 0.85 when outcomes were fully missing. For category 6, the smallest category, precision and recall were 0.31 when outcomes were partially missing versus 0.13 when outcomes were fully missing. Hence, the partial outcome information was particularly valuable for small categories where little observed data were available.

With improved imputations from partially missing outcomes, our proposed method resulted in even more efficient estimation of the regression coefficients compared to a complete case analysis. We saw the greatest estimation gains from the partial information at 80% missing data. At 80% missing data in the data probabilities setting, the partially missing scenario with a signal resulted in RMSE of 0.51, CI width of 1.20, and coverage of 0.93, compared to RMSE of 0.59, CI width of 1.41, and coverage of 0.91 for the fully missing scenario with a signal (Table 4.3). In the respective complete case analysis, RMSE was 0.63, CI width was 1.62, and coverage was 0.96. Similar patterns for partially and fully missing outcomes existed in the null scenarios in the data probabilities setting (Table 4.3), and in all scenarios in the equal probabilities setting (Table 4.4). Hence, our imputation approach offers estimation gains over the complete case analysis for both partially and fully missing outcomes, and these gains are increased further by leveraging the information from partially missing outcomes.

Keeping the scenario constant, regression coefficients were more efficiently estimated in the equal probabilities setting than in the data probabilities setting. This is because the data probabilities setting results in some large categories and some very small categories. The small categories have higher estimation uncertainty and worse imputation performance, as evidenced by

lower precision and recall for categories 3-6, which contained less data in the data probabilities setting than in the equal probabilities setting. On the other hand, the largest category in the data probabilities setting (category 1) had higher precision and recall than any category in the equal probabilities setting. Hence, the differences in estimation and imputation performance between the data probabilities and equal probabilities settings are purely a result of differences in category size. In both settings, there remained substantial gains in estimation performance from our proposed method over the complete case analysis.

A signal in the data also improved imputations. Controlling for other scenario factors, precision and recall were higher in scenarios with a signal than in scenarios with null effects. For small categories containing approximately 1/6 of the data or less (all categories in the equal probabilities setting and categories 3-6 in the data probabilities setting), a signal in the data improved imputations to a greater extent than did the partial outcome information (Tables 4.5 and 4.6). Generally, regression coefficient estimation was at least as efficient in the null scenarios as in the scenarios with a signal. This is likely due to the prior distribution for the regression coefficients being centered on zero. Hence, even though the signal aided imputations, the prior distribution provided more information for estimating the null regression coefficients.

Our simulation study demonstrates that our proposed method is able to impute missing outcomes and offers more efficient inference over a complete case analysis under a wide variety of scenarios. Imputation and estimation performance improved as more information became available, whether in the form of partially missing outcomes, larger categories, or a signal to inform outcomes.

Table 4.3: Simulation study results for the data probabilities setting. The table shows mean across 500 data sets for each measure in four simulation scenarios ("partially missing, signal," "fully missing, signal," "partially missing, null," and "fully missing, null"). The measures are root mean squared error (RMSE), bias, 95% credible interval width (width), and coverage (cov) for exposure regression coefficients. Table shows results from our proposed method and the complete case analysis for missing data levels of 0%, 20%, 50%, and 80%.

		proposed method				complete case analysis			
		RMSE	bias	width	cov	RMSE	bias	width	cov
partially missing, signal	0%	0.35	0.00	0.87	0.95	0.35	0.00	0.87	0.95
	20%	0.38	0.00	0.92	0.94	0.39	0.00	0.96	0.95
	50%	0.43	0.00	1.02	0.94	0.46	0.00	1.16	0.95
	80%	0.51	0.00	1.20	0.93	0.63	0.00	1.62	0.96
fully missing, signal	0%	0.35	0.00	0.87	0.95	0.35	0.00	0.87	0.95
	20%	0.38	0.00	0.93	0.94	0.39	0.00	0.96	0.95
	50%	0.45	0.00	1.08	0.93	0.46	0.00	1.16	0.95
	80%	0.59	0.00	1.41	0.91	0.63	0.00	1.62	0.96
partially missing, null	0%	0.34	0.00	0.83	0.95	0.34	0.00	0.83	0.95
	20%	0.38	0.00	0.90	0.95	0.38	0.00	0.93	0.95
	50%	0.44	0.00	1.07	0.94	0.47	0.00	1.16	0.95
	80%	0.53	0.00	1.35	0.95	0.62	-0.01	1.68	0.96
fully missing, null	0%	0.34	0.00	0.83	0.95	0.34	0.00	0.83	0.95
	20%	0.38	0.00	0.91	0.95	0.38	0.00	0.93	0.95
	50%	0.45	0.00	1.11	0.94	0.47	0.00	1.16	0.95
	80%	0.56	-0.01	1.49	0.94	0.62	-0.01	1.68	0.96

Table 4.4: Simulation study results for the equal probabilities setting. The table shows mean across 500 data sets for each measure in four simulation scenarios (“partially missing, signal," “fully missing, signal," “partially missing, null," and “fully missing, null"). The measures are root mean squared error (RMSE), bias, 95% credible interval width (width), and coverage (cov) for exposure regression coefficients. Table shows results from our proposed method and the complete case analysis for missing data levels of 0%, 20%, 50%, and 80%.

		proposed method				complete case analysis			
		RMSE	bias	width	cov	RMSE	bias	width	cov
partially missing, signal	0%	0.15	0.00	0.40	0.94	0.15	0.00	0.40	0.94
	20%	0.16	0.00	0.42	0.94	0.17	0.00	0.45	0.95
	50%	0.18	0.00	0.47	0.93	0.21	0.00	0.56	0.95
	80%	0.21	0.00	0.55	0.93	0.32	0.00	0.86	0.95
fully missing, signal	0%	0.15	0.00	0.40	0.94	0.15	0.00	0.40	0.94
	20%	0.16	0.00	0.43	0.94	0.17	0.00	0.45	0.95
	50%	0.20	0.00	0.52	0.93	0.21	0.00	0.56	0.95
	80%	0.30	0.00	0.74	0.90	0.32	0.00	0.86	0.95
partially missing, null	0%	0.10	0.00	0.26	0.95	0.10	0.00	0.26	0.95
	20%	0.11	0.00	0.28	0.94	0.11	0.00	0.29	0.95
	50%	0.12	0.00	0.33	0.94	0.14	0.00	0.37	0.95
	80%	0.16	0.00	0.41	0.92	0.22	0.00	0.59	0.95
fully missing, null	0%	0.10	0.00	0.26	0.95	0.10	0.00	0.26	0.95
	20%	0.11	0.00	0.29	0.94	0.11	0.00	0.29	0.95
	50%	0.13	0.00	0.35	0.93	0.14	0.00	0.37	0.95
	80%	0.20	0.00	0.52	0.92	0.22	0.00	0.59	0.95

Table 4.5: Summary of imputation performance in the data probabilities setting. Results are shown for 80% missing data and four simulation scenarios ("partially missing, signal," "fully missing, signal," "partially missing, null," and "fully missing, null"). The table shows mean across 500 data sets for precision and recall for each outcome category. Results for the other missing data levels (20% and 50%) were similar and are shown in Appendix C.2.

		outcome category					
		1	2	3	4	5	6
partially missing, signal	precision	0.92	0.69	0.54	0.43	0.31	0.31
	recall	0.92	0.69	0.53	0.43	0.31	0.31
fully missing, signal	precision	0.85	0.47	0.30	0.21	0.13	0.13
	recall	0.85	0.47	0.30	0.21	0.13	0.13
partially missing, null	precision	0.88	0.50	0.28	0.13	0.07	0.06
	recall	0.88	0.49	0.28	0.13	0.07	0.07
fully missing, null	precision	0.77	0.16	0.05	0.01	0.01	0.01
	recall	0.76	0.16	0.05	0.02	0.01	0.01

Table 4.6: Summary of imputation performance in the equal probabilities setting. Results are shown for 80% missing data and four simulation scenarios ("partially missing, signal," "fully missing, signal," "partially missing, null," and "fully missing, null"). The table shows mean across 500 data sets for precision and recall for each outcome category. Results for the other missing data levels (20% and 50%) were similar and are shown in Appendix C.2.

		outcome category					
		1	2	3	4	5	6
partially missing, signal	precision	0.71	0.67	0.63	0.63	0.59	0.62
	recall	0.71	0.68	0.63	0.63	0.58	0.61
fully missing, signal	precision	0.55	0.50	0.44	0.44	0.38	0.42
	recall	0.56	0.50	0.45	0.43	0.38	0.41
partially missing, null	precision	0.28	0.31	0.35	0.36	0.27	0.35
	recall	0.29	0.31	0.35	0.36	0.27	0.34
fully missing, null	precision	0.14	0.16	0.19	0.19	0.14	0.18
	recall	0.14	0.16	0.19	0.19	0.13	0.18

4.5 Data Analysis

We applied our proposed method to an analysis of the Denver, Colorado cohort data. The data set contained 55,273 cases, of which 62.2% ($n = 34,401$) had partially missing health outcomes. Cases with incomplete health outcomes were missing data for between 2 and 5 outcome categories (Table 4.1). We fit our proposed method to the full data set and imputed missing outcomes. Due to constraints of the stick-breaking representation of the multinomial distribution, the largest probability mass is most often assigned to the first outcome category (Zhang and Zhou, 2018). When fitting the model, we ordered the outcome categories so the largest category was first. Hence, the order was: symptomatic, asymptomatic, hospitalized, admitted to the ICU (ICU), placed on a mechanical ventilator (ventilator), and then death (Table 4.2). For comparison, we also conducted an analysis of the subset of complete cases ($n = 20,872$).

We conducted a sensitivity analysis using logistic regression. In the logistic analysis, we collapsed the multinomial categories to severe (hospitalized, ICU, ventilator, or death) and not severe (asymptomatic or symptomatic). We considered only complete cases.

In all models, we included main effects for prior year exposure to $PM_{2.5}$, ozone, and temperature as well as all pairwise interactions. To control for temporal changes in the pandemic, we included a natural cubic spline function of the case report date with 3 degrees of freedom. To account for potential non-linearities in the effect of age, we included a natural cubic spline function of age with 3 degrees of freedom. We included all covariates described in the data section. We based inference on 5,000 MCMC iterations after a burn-in of 5,000 iterations.

4.5.1 Results

The estimated exponentiated regression coefficients from our proposed method and the complete case analysis are shown in Figure 4.1. The posterior means for the regression coefficients were similar between the two methods. On average, the 95% CI's in our proposed method were 8.2% smaller than those in the complete case analysis, demonstrating the estimation gains from using our proposed method. Exponentiated regression coefficients with 95% credible in-

tervals that do not cross 1.0 existed primarily for the main effect of $PM_{2.5}$ and the interaction effect between $PM_{2.5}$ and ozone.

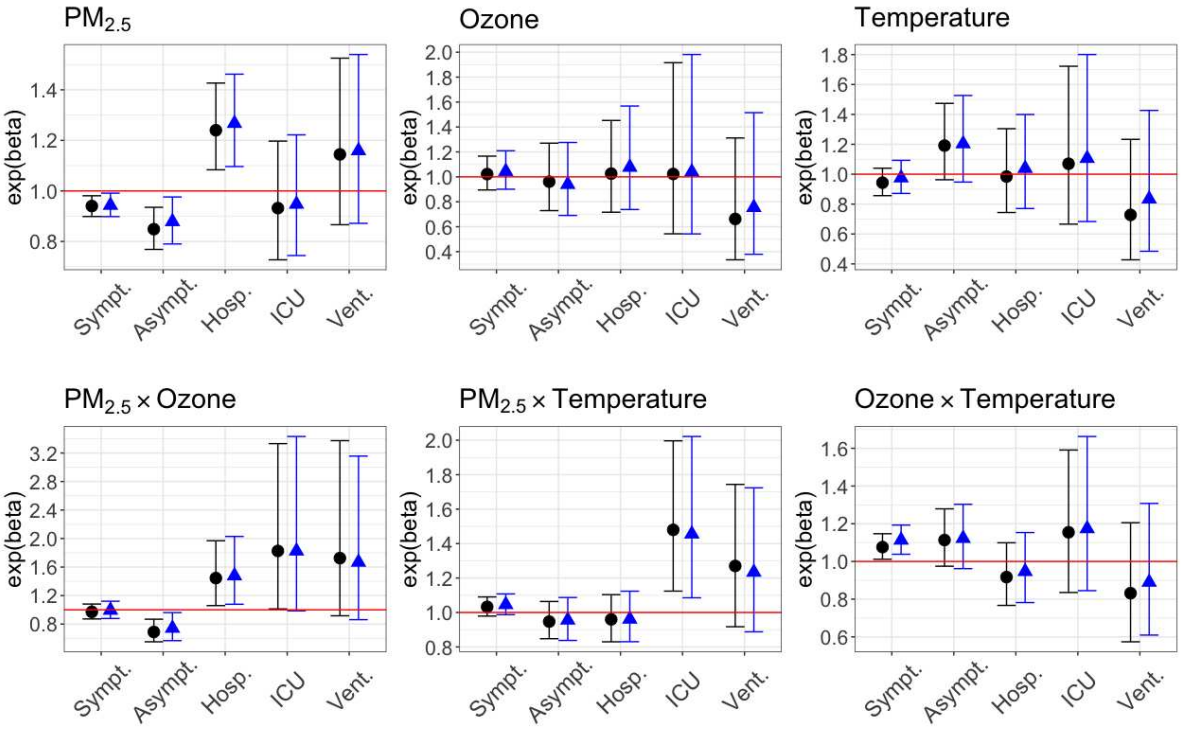
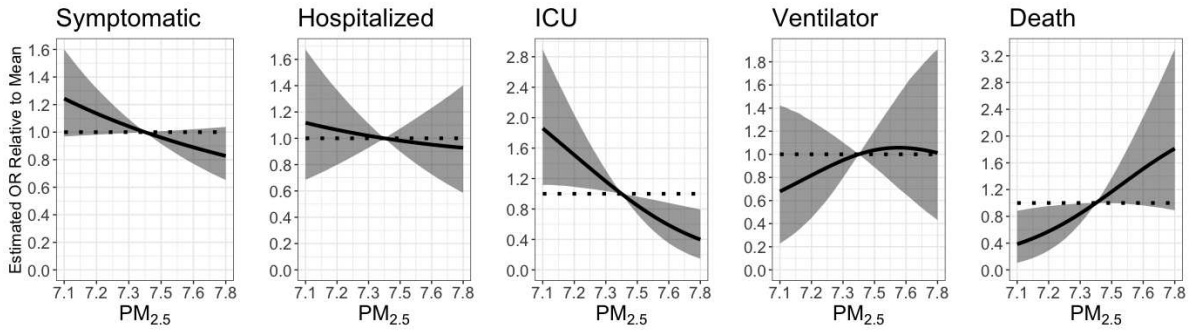


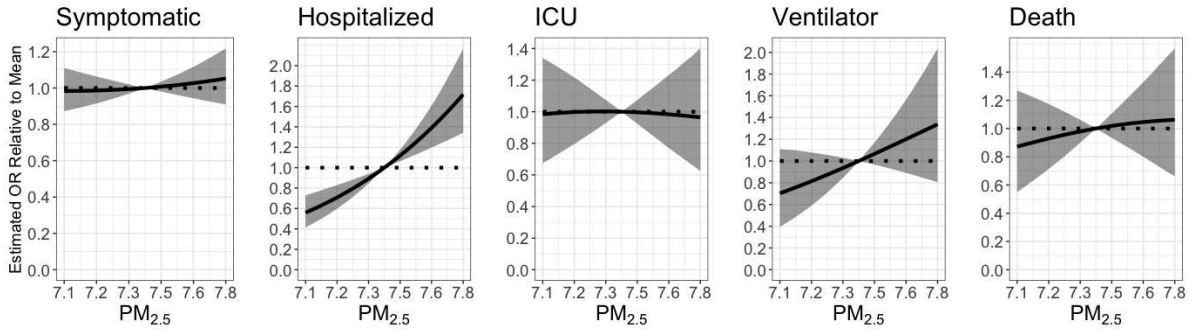
Figure 4.1: Results of the analysis of the Denver, Colorado COVID-19 cohort from our proposed method (black circles) and the complete case analysis (blue triangles). Figure shows the posterior mean and 95% credible intervals for the estimated exponentiated category-specific regression coefficients associated with main effects (top row) and pairwise interactions (bottom row). Exposures are $PM_{2.5}$, ozone, and temperature. Categories are symptomatic (sympt.), asymptomatic (asympt.), hospitalized (hosp.), admitted to the ICU (ICU), and placed on a mechanical ventilator (vent). There are no regression coefficients for the death category because it was the last category, and thus contains the remaining probability mass in the stick-breaking representation.

As described in Section 4.3.4, interpreting the regression coefficients in the stick-breaking multinomial approach is challenging. Instead, we made inference on the results using OR and IOR, as described in Section 4.3.4. We selected asymptomatic as the reference category for inference. We visualized the posterior distribution of the OR and IOR for each severity category as a function of a single exposure, holding the other two exposures at their 25^{th}, 50^{th}, and 75^{th} percentiles.

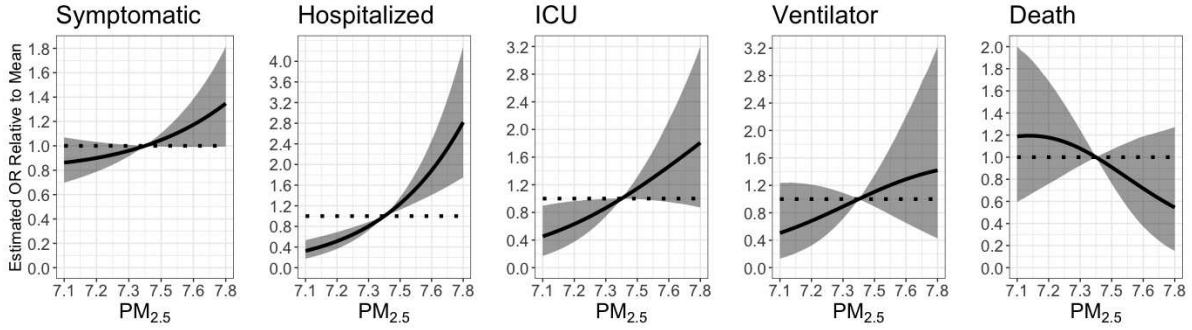
Figure 4.2 shows the posterior distribution of the OR for each peak severity category (symptomatic, hospitalized, ICU, ventilator, and death) relative to asymptomatic, as a function of average year-prior $PM_{2.5}$ exposure, holding ozone and temperature at their 25^{th}, 50^{th}, and 75^{th} percentiles. At the 25^{th} percentiles of ozone and temperature (Figure 4.2a), increased exposure to $PM_{2.5}$ was associated with a decreased risk of being admitted to the ICU, relative to being asymptomatic. There was a suggestive positive effect of $PM_{2.5}$ exposure associated with risk of death, relative to being asymptomatic. When ozone and temperature were at their 50^{th} percentiles (Figure 4.2b), increased annual $PM_{2.5}$ exposure was associated with a starkly increased risk of being hospitalized, relative to being asymptomatic. At these levels of ozone and temperature, exposure to $PM_{2.5}$ was no longer associated with risk of death. A similar pattern continued at the 75^{th} percentiles of ozone and temperature (Figure 4.2c). At these high levels of ozone and temperature, $PM_{2.5}$ exposure was associated with an increased risk of being hospitalized and, to a lesser extent, being symptomatic and admitted to the ICU, relative to being asymptomatic.



(a) 25th percentiles of ozone and temperature



(b) 50th percentiles of ozone and temperature



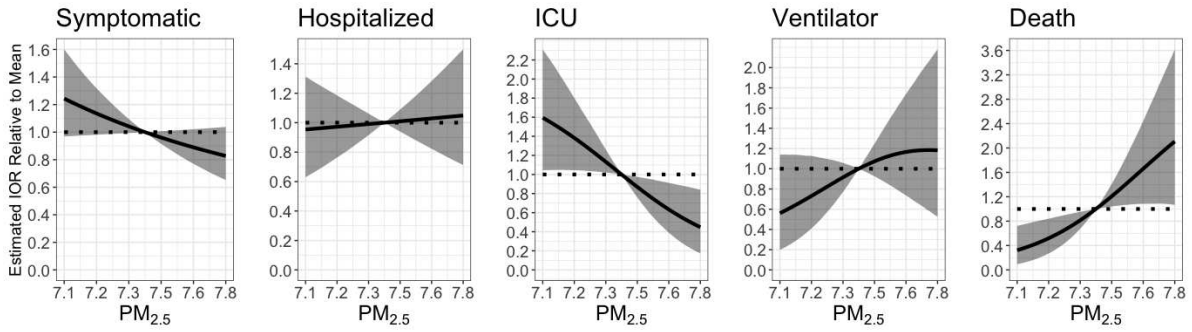
(c) 75th percentiles of ozone and temperature

Figure 4.2: Results from the analysis of the Denver, Colorado COVID-19 cohort using our proposed method. The figure shows the posterior mean (black line) and 95% credible interval (gray shaded area) of the estimated odds ratio (OR) for categories symptomatic, hospitalized, admitted to the ICU (ICU), placed on a mechanical ventilator (ventilator) and death, relative to asymptomatic. The OR was calculated as a function of annual average $PM_{2.5}$ exposure ($\mu g/m^3$) relative to the mean exposure, holding ozone and temperature at their 25^{th} (a), 50^{th} (b), and 75^{th} (c) percentiles.

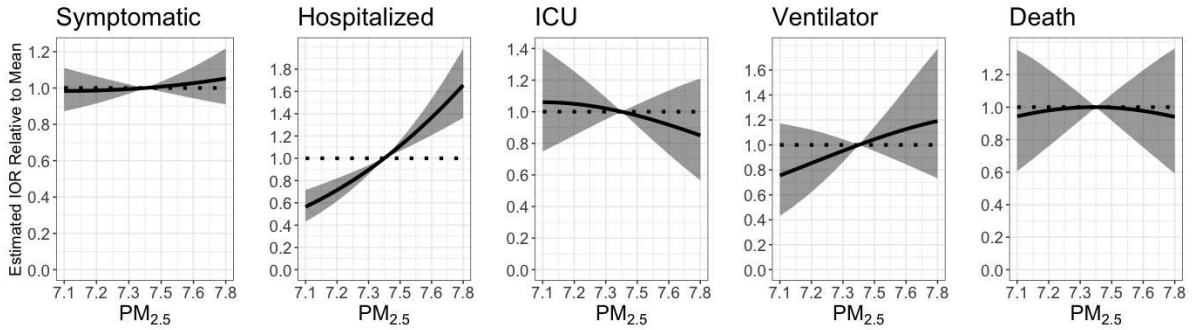
The posterior mean effect of year-prior exposure to $PM_{2.5}$ on the risk of being hospitalized or admitted to the ICU, relative to being asymptomatic, switched from a negative trend to a pos-

itive trend when ozone and temperature moved from their 25th to 75th percentiles. Changes in the effect of $PM_{2.5}$ as co-exposures change indicate interactions among exposures. The $PM_{2.5}$-ozone interaction was associated with a decreased risk of being asymptomatic relative to being hospitalized, admitted to the ICU, placed on a mechanical ventilator, or death as evidenced by the 95% credible interval for the asymptomatic category's exponentiated regression coefficient being below 1.0 (Figure 4.1). Hence, we determined the $PM_{2.5}$-ozone interaction is driving the patterns seen in the effect of $PM_{2.5}$ on risk of severe COVID-19 as ozone and temperature move from low to high levels.

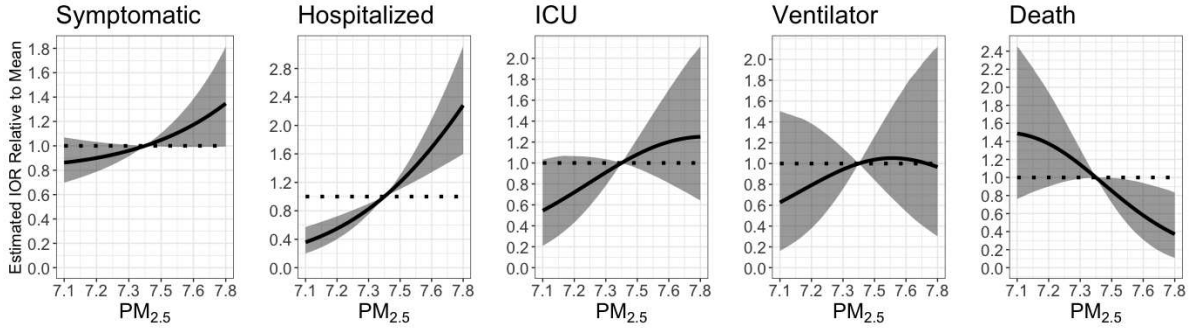
We obtained similar inferences from the IOR. Figure 4.3 shows the IOR associated with annual average $PM_{2.5}$ exposure, holding ozone and temperature at their 25th, 50th, and 75th percentiles. At the 25th percentiles of ozone and temperature (Figure 4.3a), there was a protective effect of $PM_{2.5}$ exposure on the risk of being admitted to the ICU, relative to not being admitted (e.g. being asymptomatic, symptomatic, or hospitalized only). Exposure to $PM_{2.5}$ was also associated with an increased risk of death, relative to not dying. These effects became null as ozone and temperature moved to their 50th percentiles (Figure 4.3b). At the 50th percentile of ozone and temperature, exposure to $PM_{2.5}$ was associated with an increased risk of being hospitalized, relative to not hospitalized. Similar effects of $PM_{2.5}$ occurred at the 75th percentiles of ozone and temperature (Figure 4.3c), with the addition of a suggestive protective effect on the risk of dying, relative to not dying. The complex interaction between $PM_{2.5}$ and ozone was again revealed by the directional switches in the posterior mean trends of $PM_{2.5}$ as ozone and temperature moved from the 25th to 75th percentiles.



(a) 25th percentiles of ozone and temperature



(b) 50th percentiles of ozone and temperature



(c) 75th percentiles of ozone and temperature

Figure 4.3: Results from the analysis of the Denver, Colorado COVID-19 cohort using our proposed method. The figure shows the posterior mean (black line) and 95% credible interval (gray shaded area) of the estimated incremental odds ratio (IOR) for categories symptomatic, hospitalized, admitted to the ICU (ICU), placed on a mechanical ventilator (ventilator) and death, relative to all less severe categories. The IOR was calculated as a function of annual average $PM_{2.5}$ exposure ($\mu g/m^3$) relative to the mean exposure, holding ozone and temperature at their 25^{th} (a), 50^{th} (b), and 75^{th} (c) percentiles.

Similar plots for the effects of year-prior exposure to ozone and temperature are available in Appendix C.3. Overall, there was weaker evidence for the effects of ozone and temperature

on COVID-19 severity. There were suggestive effects of increased ozone exposure associated with an increased risk of dying, relative to being asymptomatic and relative to not dying, when $PM_{2.5}$ and temperature were at their 25^{th} percentiles (Figures C.1 and C.2). These effects became null as $PM_{2.5}$ and temperature moved to their 75^{th} percentiles. Increases in temperature, combined with low levels of $PM_{2.5}$ and ozone, were associated with a decreased risk of being symptomatic relative to asymptomatic (Figure C.3). This effect was attenuated at higher levels of $PM_{2.5}$ and ozone. At high levels of $PM_{2.5}$ and ozone, there was a suggestive protective effect of temperature on risk of being hospitalized and placed on a mechanical ventilator, relative to being asymptomatic (Figure C.3), but not relative to a less severe outcome (Figure C.4). The estimated regression coefficients (Figure 4.1) indicate that interaction effects between $PM_{2.5}$ and temperature and between ozone and temperature may be driving these patterns.

4.5.2 Sensitivity Analysis Results

Results from our logistic regression sensitivity analysis are shown in Appendix C.4 (Table C.8). A one IQR increase in exposure to $PM_{2.5}$ was associated with a 9% increased risk of severe COVID-19 ($\widehat{\text{OR}}$ = 1.09, 95% CI: (1.02, 1.18)). These results mirror the results from our multinomial regression analysis. In both analyses, $PM_{2.5}$ was associated with an increased risk of severe COVID-19. In the logistic analysis, there was a positive estimated effect for the interaction between $PM_{2.5}$ and ozone, and a negative estimated effect for the interaction between ozone and temperature. Notably, the negative interaction effect between ozone and temperature may be due to the fact that annual averages for ozone and temperature were highly negatively correlated ($\rho = -0.86$).

4.6 Discussion

In this paper, we proposed a Bayesian multinomial logistic regression model for data with partially missing outcomes. We implemented Pólya-gamma data augmentation to achieve efficient computation of the posterior distribution. We developed a multiple imputation algorithm

to impute missing outcomes, where the number of missing outcome categories for each case can vary from 2 to the total number of outcomes. Our model is based on the stick-breaking representation of the multinomial distribution, which presents a challenge in interpreting regression coefficients. The stick-breaking multinomial regression approach has historically been used for applications focused on clustering and prediction (Linderman et al., 2015). To our knowledge, we present the first application of this approach in which inference using the odds ratio is the primary goal. We proposed an inferential approach based on visualization of the posterior distribution to retain the familiar logistic regression interpretation of the odds ratios.

In a simulation study, we demonstrated our method's ability to impute missing outcome data and improve estimation over complete case analyses. In eight different scenarios, our proposed method produced unbiased estimates for the regression coefficients that had smaller RMSE and CI width than estimates from respective complete case analyses. Our proposed method leveraged information from various sources to improve imputation. These sources include: partially, as opposed to fully, missing outcomes, a signal in the data, and larger outcome categories. Better imputations resulted in even more efficient inference on the regression coefficients using our method compared to the complete case analysis.

Using our proposed method, we estimated the association between long-term exposure to $PM_{2.5}$, ozone, and temperature and COVID-19 peak severity in a Denver, Colorado cohort. Our model imputed outcomes for the 34,401 cases with partially missing outcome data. In our analysis, we found increased long-term exposure to $PM_{2.5}$, combined with high levels of ozone and temperature, was associated with an increased risk of being hospitalized and admitted to the ICU, relative to being asymptomatic. These associations were null or reversed when ozone and temperature were low, indicating interaction effects between the exposures. Through visualization of the OR and IOR, combined with analysis of the estimated regression coefficients, we identified an interaction effect between $PM_{2.5}$ and ozone. A complete case analysis produced similar results, but with more uncertainty, further exemplifying the estimation gains from our proposed method with imputation. Our results support recent studies that identified an asso-

ciation between increased $PM_{2.5}$ exposure and a lagged effect on COVID-19 mortality (Garcia et al., 2021; Shao et al., 2021). Our individual-level analysis of the Denver, Colorado cohort fills a major gap in the literature. With individual-level data, we controlled for known confounding variables and risk factors, and began to establish a driving association between air pollution exposure and COVID-19 outcomes.

Chapter 5

Conclusion

Exposure to air pollution presents an ongoing threat to human health. Understanding personal exposures and associated health outcomes can help inform air pollution policy and interventions, leading to improved health. To advance the state of the science in air pollution epidemiology, we identified three major gaps in the literature, which we aimed to fill in this dissertation.

In Chapter 2, we highlighted the statistical modeling challenges induced by joint exposure to multiple pollutants, which are further complicated by lack of evaluation of existing methods. We conducted a simulation study comparing five recently developed Bayesian methods for multipollutant mixtures. Our simulation study showed that Bayesian kernel machine regression (BKMR), a nonparametric method, is highly adaptable and can accurately estimate exposure-response functions with complex nonlinear or interaction effects. On the other hand, nonparametric Bayes shrinkage (NPB), a Bayesian linear effect measure modification model, performs well in approximately linear scenarios and can efficiently identify mixture components that have main effects and/or interaction effects on the health outcome in the presence of highly correlated exposures. BKMR may be preferred when the primary goal is to estimate a complex exposure-response function, or to predict health outcomes. NPB may be the preferred choice when the emphasis is on identifying which components of a mixture are associated with the health outcome. Hence, the most appropriate statistical method depends on the research question of interest, as well as the underlying data structure. We applied the methods in an analysis of lung function in children with asthma. Using both BKMR and NPB, we estimated a negative association between nitrogen dioxide and lung function. To promote the use of these contemporary methods for multipollutant mixtures, we developed software to implement each method and post-process results. The software also allows users to reproduce our simulation study as a tool to determine the most appropriate method for their application.

In Chapter 3, we addressed the common problem of missing exposure data in temporal exposure assessments. We developed a Bayesian nonparametric infinite hidden Markov model that leverages information from multiple exposure assessments to identify shared activity patterns, estimate parameters, and impute missing observations. Our proposed model estimates shared hidden states of exposure among multiple time series, and uses the estimated hidden states to inform imputations. In simulation and validation studies, our method outperformed independent analyses of multiple time series, models with no temporal structure, and models with deterministic states of exposure in estimation and imputation. We applied our method to an analysis of 50 sampling days from the Fort Collins Commuter Study, where each sampling day consisted of time-resolved personal exposure to fine particulate matter ($PM_{2.5}$), black carbon, and carbon monoxide. Our model imputed exposure data that were both missing at random and below the limit of detection. Among the 50 sampling days, we identified 53 shared hidden states of exposure. We investigated the hidden states to draw inference on time-activity patterns associated with exposures, and found evidence of a potential cooking activity associated with higher than average pollutant exposures.

We focused on missing health outcome data in Chapter 4. We developed a fully Bayesian method for multinomial logistic regression analysis with partially missing outcome data. We demonstrated our proposed method's estimation gains over a complete case analysis in a variety of simulation scenarios. We then applied our proposed method to estimate the association between long-term exposure to $PM_{2.5}$, ozone, and temperature and COVID-19 peak severity in a Denver, Colorado cohort. By imputing partially missing outcome data, we achieved greater estimation efficiency from a substantially larger sample size ($n = 55273$) compared to a complete case analysis ($n = 20872$). With our novel visualization approach, we made inference on the odds ratios associated with each exposure, and provided sensible interpretation of interaction effects. We found a positive association between $PM_{2.5}$ exposure and increased risk of severe COVID-19 outcomes. We also estimated an interaction effect between $PM_{2.5}$ and ozone. Our results support previous findings from ecological analyses on the relationship between air pol-

lution exposure and COVID-19 endpoints. Our research further contributes to the literature by providing an individual-level analysis, in which we identified an association between air pollution exposure and COVID-19 severity while controlling for individual- and neighborhood-level risk factors.

5.1 Future Work

We identified several areas for future work in environmental mixtures analyses. First, the performance of BKMR and NPB merits a combination of these approaches that draws on the top qualities of each. A possible new direction is a nonlinear extension to NPB that maintains the ease of implementation and inference, but relaxes assumptions on the shape of the exposure-response function. Regarding exposure assessments, our proposed Bayesian infinite hidden Markov model for multiple time series with missing data could be extended to accommodate continuous time series or non-Gaussian emissions. In addition, improvements could be made to speed computation. Extensions to our proposed method for Bayesian multinomial logistic regression analysis include accommodating highly correlated exposures through shrinkage priors or variable selection. Through our work we addressed a number of research problems and provided valuable tools for scientists and applied statisticians. We have also opened the door to promising future directions for statistical methods development in the field of air pollution epidemiology.

Bibliography

Agier, L., Portengen, L., Chadeau-hyam, M., Basagaña, X., Giorgis-allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J. R., Nieuwenhuijsen, M. J., Vineis, P., Vrijheid, M., Slama, R., and Vermeulen, R. (2016). A Systematic Comparison of Linear Regression – Based Statistical Methods to Assess Exposome-Health Associations. *Environmental Health Perspectives*, 124(12):1848–1856.

Altman, R. M. K. (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.

Antonelli, J., Mazumdar, M., Bellinger, D., Christiani, D. C., Wright, R., and Coull, B. A. (2020). Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors. *Annals of Applied Statistics*, 14(1):257–275.

Austin, E., Coull, B., Thomas, D., and Koutrakis, P. (2012). A framework for identifying distinct multipollutant profiles in air pollution data. *Environment International*, 45:112–121.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897.

Barrera-Gómez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J. R., Nieuwenhuijsen, M., Vineis, P., Vrijheid, M., Vermeulen, R., Slama, R., and Basagaña, X. (2017). A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environmental Health: A Global Access Science Source*, 16(74).

Beal, M. J. and Rasmussen, C. E. (2002). The Infinite Hidden Markov Model. *Advances in Neural Information Processing Systems*, pages 577–584.

Benka-Coker, W., Hoskovec, L., Severson, R., Balmes, J., Wilson, A., and Magzamen, S. (2020). The joint effect of ambient air pollution and agricultural pesticide exposures on lung function among children with asthma. *Environmental Research*, 190(February):109903.

Bhaskar, A., Chandra, J., Braun, D., Cellini, J., and Dominici, F. (2020). Air pollution, SARS-CoV-2 transmission, and COVID-19 outcomes: A state-of-the-science review of a rapidly evolving research area. *medRxiv*.

Bobb, J. F. (2017). bkmr: Bayesian Kernel Machine Regression. R package version 0.2.0.

Bobb, J. F., Henn, B. C., Valeri, L., and Coull, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environmental Health: A Global Access Science Source*, pages 1–10.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508.

Bozack, A., Pierre, S., DeFelice, N., Colicino, E., Jack, D., Chillrud, S. N., Rundle, A., Astua, A., Quinn, J. W., McGuinn, L., Yang, Q., Johnson, K., Masci, J., Lukban, L., Maru, D., and Lee, A. G. (2021). Long-Term Air Pollution Exposure and COVID-19 Mortality: A Patient-Level Analysis from New York City. *American Journal of Respiratory and Critical Care Medicine*, pages 1–52.

Brandt, E. B. and Mersha, T. B. (2021). Environmental Determinants of Coronavirus Disease 2019 (COVID-19). *Current Allergy and Asthma Reports*, 21(3).

Braun, J. M., Gennings, C., Hauser, R., and Webster, T. F. (2016). What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environmental Health Perspectives*, 124(1):A6–A9.

Bulathsinghala, A. T. and Shaw, I. C. (2014). The toxic chemistry of methyl bromide. *Human and Experimental Toxicology*, 33(1):81–91.

California Department of Pesticide Regulation (2015). California Pesticide Use Reporting Data.

Caliński, T. and Harabasz, J. (1974). Communications in Statistics - Theory and Methods. *Communications in Statistics*, 3(1):1–27.

Carbajal-Arroyo, L., Miranda-Soberanis, V., Medina-Ramón, M., Rojas-Bracho, L., Tzintzun, G., Solís-Gutiérrez, P., Méndez-Ramírez, I., Hurtado-Díaz, M., Schwartz, J., and Romieu, I. (2011). Effect of PM10 and O3 on infant mortality among residents in the Mexico City Metropolitan Area: A case-crossover analysis, 1997-2005. *Journal of Epidemiology and Community Health*, 65(8):715–721.

Carrico, C., Gennings, C., Wheeler, D. C., and Factor-Litvak, P. (2015). Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(1):100–120.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Chan, J. C. C. and Jeliazkov, I. (2009). MCMC estimation of restricted covariance matrices. *Journal of Computational and Graphical Statistics*, 18(2):457–480.

Chiu, Y. H., Bellavia, A., James-Todd, T., Correia, K. F., Valeri, L., Messerlian, C., Ford, J. B., Mínguez-Alarcón, L., Calafat, A. M., Hauser, R., and Williams, P. L. (2018). Evaluating effects of prenatal exposure to phthalate mixtures on birth weight: A comparison of three statistical approaches. *Environment International*, 113(November 2017):231–239.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes Conditional Distribution Modeling With Variable Selection. *Journal of the American Statistical Association*, 104(488):1646–1660.

Ciencewicki, J. and Jaspers, I. (2007). Air pollution and respiratory viral infection. *Inhalation Toxicology*, 19(14):1135–1146.

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics*, 11(6):745–763.

Colovic, M. B., Krsti, D. Z., Lazarevi-Pati, T. D., Bondi, A. M., and Vasi, V. M. (2013). Acetylcholinesterase Inhibitors: Pharmacology and Toxicology. *Current Neuropharmacology*, 11:315–335.

Comunian, S., Dongo, D., Milani, C., and Palestini, P. (2020). Air pollution and covid-19: The role of particulate matter in the spread and increase of covid-19's morbidity and mortality. *International Journal of Environmental Research and Public Health*, 17(12):1–22.

Copat, C., Cristaldi, A., Fiore, M., Grasso, A., Zuccarello, P., Signorelli, S. S., Conti, G. O., and Ferrante, M. (2020). The role of air pollution (PM and NO2) in COVID-19 spread and lethality: A systematic review. *Environmental Research*, 191(110129).

Cui, Y., Zhang, Z. F., Froines, J., Zhao, J., Wang, H., Yu, S. Z., and Detels, R. (2003). Air pollution and case fatality of SARS in the People's Republic of China: An ecologic study. *Environmental Health: A Global Access Science Source*, 2:1–5.

Dahl, D. B. (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. *Bayesian Inference for Gene Expression and Proteomics*, pages 201–218.

Davalos, A. D., Luben, T. J., Herring, A. H., and Sacks, J. D. (2017). Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Annals of Epidemiology*, 27(2):145–153.

Department of Public Health and Environment (2021). Colorado Air Quality.

Di, Q., Dai, L., Wang, Y., Zanobetti, A., Choirat, C., Schwartz, J. D., and Dominici, F. (2017a). Association of Short-term Exposure to Air Pollution With Mortality in Older Adults. *JAMA*, 318(24):2446–2456.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F., and Schwartz, J. D. (2017b). Air Pollution and Mortality in the Medicare Population. *New England Journal of Medicine*, 376(26):2513–2522.

Dias, J. G., Vermunt, J. K., and Ramos, S. (2015). Clustering financial time series: New insights from an extended hidden Markov model. *European Journal of Operational Research*, 243(3):852–864.

Dockery, D. W. and Pope, C. A. (1994). Acute Respiratory Effects of Particulate Air Pollution. *Annual Review of Public Health*, 15(1):107–132.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., and Speizer, F. E. (1993). An Association between Air Pollution and Mortality in Six U.S. Cities. *New England Journal of Medicine*, 329(24):1753–1759.

Domingo, J. L. and Rovira, J. (2020). Effects of air pollutants on the transmission and severity of respiratory viral infections. *Environmental Research*, 187(109650).

Dominici, F., Peng, R. D., Barr, C. D., and Bell, M. L. (2010). Protecting Human Health From Air Pollution. *Epidemiology*, 21(2):187–194.

Dunson, D. B., Herring, A. H., and Engel, S. M. (2008). Bayesian Selection and Clustering of Polymorphisms in Functionally Related Genes. *Journal of the American Statistical Association*, 103(482):534–546.

Engels, J. M. and Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968–976.

Finazzi, F. and Paci, L. (2019). Quantifying personal exposure to air pollution from smartphone-based location data. *Biometrics*, (May):1356–1366.

Fox, E. B., Hughes, M. C., Sudderth, E. B., and Jordan, M. I. (2014). Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *Annals of Applied Statistics*, 8(3):1281–1313.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2 A):1020–1056.

Frontera, A., Cianfanelli, L., Vlachos, K., Landoni, G., and Cremona, G. (2020). Severe air pollution links to higher mortality in COVID-19 patients : The "double-hit" hypothesis. *Journal of Infection*, 81:255–259.

Gale, S. L., Noth, E. M., Mann, J., Balmes, J., Hammond, S. K., and Tager, I. B. (2012). Polycyclic aromatic hydrocarbon exposure and wheeze in a cohort of children with asthma in Fresno, CA. *Journal of Exposure Science and Environmental Epidemiology*, 22(4):386–392.

Garcia, E., Marian, B., Chen, Z., Li, K., Lurmann, F., Gilliland, F., and Eckel, S. P. (2021). Long-term air pollution and COVID-19 mortality rates in California: Findings from the Spring/Summer and Winter surges of COVID-19. *Environmental Pollution*, 292(118396).

Gass, K., Klein, M., Chang, H. H., Dana Flanders, W., and Strickland, M. J. (2014). Classification and regression trees for epidemiologic research: An air pollution example. *Environmental Health: A Global Access Science Source*, 13(1):17.

Gass, K., Klein, M., Sarnat, S. E., Winquist, A., Darrow, L. A., Flanders, W. D., Chang, H. H., Mulholland, J. A., Tolbert, P. E., and Strickland, M. J. (2015). Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: a classification and regression tree approach. *Environmental Health: A Global Access Science Source*, 14(1):58.

Gibson, E. A., Goldsmith, J., and Kioumourtzoglou, M.-A. (2019). Complex Mixtures, Complex Analyses: an Emphasis on Interpretable Results. *Current Environmental Health Reports*, 6(2):53–61.

Global Burden of Diseases 2019 Risk Factors Collaborators (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258):1223–1249.

Good, N., Mölter, A., Ackerson, C., Bachand, A., Carpenter, T., Clark, M. L., Fedak, K. M., Kayne, A., Koehler, K., Moore, B., L'Orange, C., Quinn, C., Ugave, V., Stuart, A. L., Peel, J. L., and Volckens, J. (2016). The Fort Collins Commuter Study: Impact of route type and transport mode on personal exposure to multiple air pollutants. *Journal of Exposure Science and Environmental Epidemiology*, 26(4):397–404.

Hamra, G. B. and Buckley, J. P. (2018). Environmental Exposure Mixtures: Questions and Methods to Address Them. *Current Epidemiology Reports*, 5:160–165.

Health Effects Institute (2018). State of Global Air 2018. Technical Report 4, Boston, MA.

Hensley, A. A. and Djuric, P. M. (2017). Nonparametric learning for Hidden Markov Models with preferential attachment dynamics. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 3854–3858.

Herring, A. H. (2010). Nonparametric bayes shrinkage for assessing exposures to mixtures subject to limits of detection. *Epidemiology*, 21:S71–S76.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168.

Hopke, P. K., Liu, C., and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the arctic. *Biometrics*, 57(1):22–33.

Hoskovec, L. (2019). mmpack: Implement methods for multipollutant mixtures analyses. R package version 0.1.0.

Hoskovec, L. (2021a). pgmultinomr: Bayesian multinomial regression for partially missing outcome data. R package version 0.1.0.

Hoskovec, L. (2021b). psbpHMM: Covariate-dependent iHMMs for multiple time series. R package version 0.1.0.

Houseman, E. A. and Virji, M. A. (2017). A Bayesian approach for summarizing and modeling time-series exposure data with left censoring. *Annals of Work Exposures and Health*, 61(7):773–783.

Junger, W. L. and Ponce de Leon, A. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, 102:96–104.

Koehler, K., Good, N., Wilson, A., Mölter, A., Moore, B. F., Carpenter, T., Peel, J. L., and Volckens, J. (2019). The Fort Collins commuter study: Variability in personal exposure to air pollutants by microenvironment. *Indoor Air*, 29(2):231–241.

Kogevinas, M., Castaño Vinyals, G., Karachaliou, M., Espinosa, A., De Cid, R., Garcia Aymerich, J., O'callaghan Gordo, C., Moncunill, G., Dobaño, C., and Tonne, C. (2021). Ambient air pollution and risk of SARS-CoV-2 infection and of COVID-19 disease in a cohort study in Catalonia (COVICAT Cohort). *ISEE Conference Abstracts*, 2021(1):1–10.

Krall, J. R., Simpson, C. H., and Peng, R. D. (2015). A model-based approach for imputing censored data in source apportionment studies. *Environmental and Ecological Statistics*, 22(4):779–800.

Landguth, E. L., Holden, Z. A., Graham, J., Stark, B., Mokhtari, E. B., Kaleczyc, E., Anderson, S., Urbanski, S., Jolly, M., Semmens, E. O., Warren, D. A., Swanson, A., Stone, E., and Noonan, C. (2020). The delayed effect of wildfire season particulate matter on subsequent influenza season in a mountain west region of the USA. *Environment International*, 139(105668).

Langrock, R., Swihart, B. J., Caffo, B. S., Punjabi, N. M., and Crainiceanu, C. M. (2013). Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in Medicine*, 32(19):3342–3356.

Lenters, V., Portengen, L., Rignell-Hydbom, A., Jönsson, B. A., Lindh, C. H., Piersma, A. H., Toft, G., Bonde, J. P., Heederik, D., Rylander, L., and Vermeulen, R. (2016). Prenatal Phthalate, Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts: Multi-Pollutant Models Based on Elastic Net Regression. *Environmental Health Perspectives*, 124(3):365–372.

Li, Y. and Ghosh, S. K. (2015). Efficient Sampling Methods for Truncated Multivariate Normal and Student-t Distributions Subject to Linear Inequality Constraints. *Journal of Statistical Theory and Practice*, 9(4):712–732.

Linderman, S. W., Johnson, M. J., and Adams, R. P. (2015). Dependent multinomial models made easy: Stick breaking with the Pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, pages 3456–3464.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons, Incorporated, 3 edition.

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., and Richardson, S. (2015). PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. *Journal of Statistical Software*, 64(7):1–30.

Mann, J. K., Balmes, J. R., Bruckner, T. A., Mortimer, K. M., Margolis, H. G., Pratt, B., Katharine Hammond, S., Lurmann, F. W., and Tager, I. B. (2010). Short-term effects of air pollution on wheeze in asthmatic children in Fresno, California. *Environmental Health Perspectives*, 118(10):1497–1502.

Margolis, H. G., Mann, J. K., Lurmann, F. W., Mortimer, K. M., Balmes, J. R., Hammond, S. K., and Tager, I. B. (2009). Altered pulmonary function in children with asthma associated with

highway traffic near residence. *International Journal of Environmental Health Research*, 19(2):139–155.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics*, 11(3):484–498.

Molitor, J., Su, J. G., Molitor, N. T., Rubio, V. G., Richardson, S., Hastie, D., Morello-Frosch, R., and Jerrett, M. (2011). Identifying vulnerable populations through an examination of the association between multipollutant profiles and poverty. *Environmental Science and Technology*, 45(18):7754–7760.

Montañez, G. D., Amizadeh, S., and Laptev, N. (2015). Inertial hidden Markov models: Modeling change in multivariate time series. *Proceedings of the National Conference on Artificial Intelligence*, 3:1819–1825.

Mortimer, K., Neugebauer, R., Lurmann, F., Alcorn, S., Balmes, J., and Tager, I. (2008). Early-lifetime exposure to air pollution and allergic sensitization in children with asthma. *Journal of Asthma*, 45(10):874–881.

Neal, R. M. (2003). Slice Sampling. *Annals of Statistics*, 31(3):705–767.

NIEHS (2012). 2012-2017 Strategic Plan: Advancing Science, Improving Health: A plan for environmental health research. Technical report.

NIEHS (2018). 2018-2023 Strategic Plan: Advancing Environmental Health Sciences Improving Health. Technical report, National Institute of Environmental Health Sciences.

Nikolov, M. C., Coull, B. A., Catalano, P. J., and Godleski, J. J. (2007). An informative Bayesian structural equation model to assess source-specific health effects of air pollution. *Biostatistics*, 8(3):609–624.

Noth, E. M., Hammond, S. K., Biging, G. S., and Tager, I. B. (2011). A spatial-temporal regression model to predict daily outdoor residential PAH concentrations in an epidemiologic study in Fresno, CA. *Atmospheric Environment*, 45(14):2394–2403.

Pachon, J. E., Balachandran, S., Hu, Y., Mulholland, J. A., Darrow, L. A., Sarnat, J. A., Tolbert, P. E., and Russell, A. G. (2012). Development of outcome-based, multipollutant mobile source indicators. *Journal of the Air and Waste Management Association*, 62(4):431–442.

Padula, A. M., Balmes, J. R., Eisen, E. A., Mann, J., Noth, E. M., Lurmann, F. W., Pratt, B., Tager, I. B., Nadeau, K., and Hammond, S. K. (2015). Ambient polycyclic aromatic hydrocarbons and pulmonary function in children. *Journal of Exposure Science and Environmental Epidemiology*, 25(3):295–302.

Pan, L., Wu, S., Li, H., Xu, J., Dong, W., Shan, J., Yang, X., Chen, Y., Shima, M., Deng, F., and Guo, X. (2018). The short-term effects of indoor size-fractioned particulate matter and black carbon on cardiac autonomic function in COPD patients. *Environment International*, 112(38):261–268.

Papathomas, M., Molitor, J., Hoggart, C., Hastie, D., and Richardson, S. (2012). Exploring Data From Genetic Association Studies Using Bayesian Variable Selection and the Dirichlet Process: Application to Searching for Gene × Gene Patterns. *Genetic Epidemiology*, 36(6):663–674.

Pearce, J. L., Waller, L. A., Chang, H. H., Klein, M., Mulholland, J. A., Sarnat, J. A., Sarnat, S. E., Strickland, M. J., and Tolbert, P. E. (2014). Using self-organizing maps to develop ambient air quality classifications: a time series example. *Environmental Health: A Global Access Science Source*, 13(1):56.

Pearce, J. L., Waller, L. A., Mulholland, J. A., Sarnat, S. E., Strickland, M. J., Chang, H. H., and Tolbert, P. E. (2015). Exploring associations between multipollutant day types and asthma

morbidity: Epidemiologic applications of self-organizing map ambient air quality classifications. *Environmental Health: A Global Access Science Source*, 14(55).

Pearce, J. L., Waller, L. A., Sarnat, S. E., Chang, H. H., Klein, M., Mulholland, J. A., and Tolbert, P. E. (2016). Characterizing the spatial distribution of multiple pollutants and populations at risk in Atlanta, Georgia. *Spatial and Spatio-temporal Epidemiology*, 18:13–23.

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.

R Core Team (2018). R: A language and environment for statistical computing.

Raanan, R., Balmes, J. R., Harley, K. G., Gunier, R. B., Magzamen, S., Bradman, A., and Eskenazi, B. (2016). Decreased lung function in 7-year-old children with early-life organophosphate exposure. *Thorax*, 71(2):148–153.

Rabiner, L. R. and Juang, B. H. (1986). An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3(1):4–16.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer US, New York, 2 edition.

Roberts, S. and Martin, M. (2005). A critical assessment of shrinkage-based regression approaches for estimating the adverse health effects of multiple air pollutants. *Atmospheric Environment*, 39:6223–6230.

Roberts, S. and Martin, M. (2006a). Investigating the mixture of air pollutants associated with adverse health outcomes. *Atmospheric Environment*, 40(5):984–991.

Roberts, S. and Martin, M. A. (2006b). Using Supervised Principal Components Analysis to Assess Multiple Pollutant Effects. *Environmental Health Perspectives*, 114(12):1877–1882.

Rodríguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1):145–178.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C):53–65.

Sacks, J. D., Ito, K., Wilson, W. E., and Neas, L. M. (2012). Impact of covariate models on the assessment of the air pollution-mortality association in a single-and multipollutant context. *American Journal of Epidemiology*, 176(7):622–634.

Samet, J. M. (2005). The perspective of the National Research Council's Committee on Research Priorities for Airborne Particulate Matter. *Journal of Toxicology and Environmental Health - Part A*, 68(13-14):1063–1067.

Sarkar, A., Bhadra, A., and Mallick, B. K. (2012). Nonparametric Bayesian Approaches to Non-homogeneous Hidden Markov Models. *arXiv: 1205.1839v1*.

Setti, L., Passarini, F., De Gennaro, G., Barbieri, P., Pallavicini, A., Ruscio, M., Piscitelli, P., Colao, A., and Miani, A. (2020a). Searching for SARS-COV-2 on particulate matter: A possible early indicator of COVID-19 epidemic recurrence. *International Journal of Environmental Research and Public Health*, 17(9).

Setti, L., Passarini, F., De Gennaro, G., Barbieri, P., Perrone, M. G., Borelli, M., Palmisani, J., Di Gilio, A., Torboli, V., Fontana, F., Clemente, L., Pallavicini, A., Ruscio, M., Piscitelli, P., and Miani, A. (2020b). SARS-Cov-2RNA found on particulate matter of Bergamo in Northern Italy: First evidence. *Environmental Research*, 188(109754).

Severson, R. (2019). purexposure: Pull and Calculate Exposure to CA Pesticide Use Registry Records. R package version 0.1.0.

Shao, L., Cao, Y., Jones, T., Santosh, M., Silva, L. F., Ge, S., da Boit, K., Feng, X., Zhang, M., and BéruBé, K. (2021). COVID-19 mortality and exposure to airborne PM2.5: A lag time correlation. *Science of The Total Environment*, 2.

Slama, R. and Vrijheid, M. (2015). Some challenges of studies aiming to relate the Exposome to human health. *Occupational and Environmental Medicine*, 72(6):383–384.

Spezia, L., Futter, M. N., and Brewer, M. J. (2011). Periodic multivariate normal hidden Markov models for the analysis of water quality time series. *Environmetrics*, 22(3):304–317.

Sun, Z., Tao, Y., Li, S., Ferguson, K. K., Meeker, J. D., Park, S. K., Batterman, S. A., and Mukherjee, B. (2013). Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health: A Global Access Science Source*, 12(1):85.

Taylor, K. W., Joubert, B. R., Braun, J. M., Dilworth, C., Gennings, C., Hauser, R., Heindel, J. J., Rider, C. V., Webster, T. F., and Carlin, D. J. (2016). Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology: Lessons from an Innovative Workshop. *Environmental Health Perspectives*, 124(12):A227–A229.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic.

United States Census Bureau (2020). American Community Survey (ACS).

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095.

Van Sickle, D., Magzamen, S., and Mullahy, J. (2011). Understanding socioeconomic and racial differences in adult lung function. *American Journal of Respiratory and Critical Care Medicine*, 184(5):521–527.

Vedal, S., Hannigan, M., Dutton, S., Miller, S., Milford, J., Rabinovitch, N., Kim, S.-Y., and Sheppard, L. (2009). The Denver Aerosol Sources and Health (DASH) study: Overview and early findings. *Atmospheric Environment*, 43(9):1666–1673.

Wade, S. and Ghahramani, Z. (2018). Bayesian Cluster Analysis: Point estimation and credible balls (with Discussion). *Bayesian Analysis*, 13(2):559–626.

Walker, K., Herman, M., and Eberwein, K. (2021). tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, 36(1):45–54.

Wang, C. Y. and Hsu, L. (2020). Multinomial logistic regression with missing outcome data: An application to cancer subtypes. *Statistics in Medicine*, 39(24):3299–3312.

Weisskopf, M. G., Seals, R. M., and Webster, T. F. (2018). Bias Amplification in Epidemiologic Analysis of Exposure to Mixtures. *Environmental Health Perspectives*, 126(4):047003.

Winquist, A., Kirrane, E., Klein, M., Strickland, M., Darrow, L. A., Sarnat, S. E., Gass, K., Mulholland, J., Russell, A., and Tolbert, P. (2014). Joint effects of ambient air pollutants on pediatric asthma emergency department visits in atlanta, 1998-2004. *Epidemiology*, 25(5):666–673.

Witte, J. S. and Greenland, S. (1996). Simulation Study of Hierarchical Regression. *Statistics in Medicine*, 15(11):1161–1170.

Wold, S., Ruhe, A., Wold, H., and Dunn III, W. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Society for Industrial and Applied Mathematics*, 5(3):735–743.

Zanobetti, A., Austin, E., Coull, B. A., Schwartz, J., and Koutrakis, P. (2014). Health effects of multi-pollutant profiles. *Environment International*, 71:13–19.

Zhang, Q. and Zhou, M. (2018). Permuted and augmented stick-breaking Bayesian multinomial regression. *Journal of Machine Learning Research*, 18:1–33.

Zhou, M., He, Y., Yu, M., and Hsu, C. H. (2017). A nonparametric multiple imputation approach for missing categorical data. *BMC Medical Research Methodology*, 17(1):1–12.

Appendix A

Model choice for estimating the association between exposure to chemical mixtures and health outcomes: A simulation study

A.1 Demographic Characteristics of the Sample

Table A.1: Descriptive statistics of FACES data. Demographic characteristics of the sample of FACES data used in simulation studies and data analysis.

	n = 153
Age (years), mean (SD)	9 (1.8)
Male gender, %	60.1
Height (inches), mean (SD)	52.3 (4.8)
Ethnicity, %	
Non-Hispanic Black	13.7
Non-Hispanic White	46.4
Hispanic	39.9
BMI (kg/m^2), mean (SD)	18.5 (4.6)
Mother < 12th grade education, %	60.1
Insured, %	94.8
Atopy, %	78.4
Father/Mother smokes currently, %	5.2
Proximity to Freeway (< 1 block away), %	47.7
Severity (GINA ≥ 3), %	17.6
Household income > 30K/year, %	52.3
FEV_1 (L), mean (SD)	1.7 (0.4)

A.2 Hyperparameter Specification

A.2.1 Nonparametric Bayes shrinkage

We specify the following prior hyperparameters for NPB and NPBr in our simulation study and data analysis:

$$\begin{array}{cccc} \boldsymbol{\mu}_\gamma = \mathbf{0} & \kappa^{-2} = 1 & \mu_0 = 0 & \kappa_0^{-2} = 1 \\ \sigma_{\mu 1}^{-2} = 1 & \sigma_{\mu 2}^{-2} = 1 & \alpha_{\phi 1} = 1 & \beta_{\phi 1} = 1 \\ \alpha_{\phi 2} = 1 & \beta_{\phi 2} = 1 & \alpha_\sigma = 1 & \beta_\sigma = 1 \\ \alpha_{\pi 1} = 1 & \beta_{\pi 1} = 1 & \alpha_{\pi 2} = 9 & \beta_{\pi 2} = 1 \\ \alpha_{\alpha 1} = 2 & \beta_{\alpha 1} = 1 & \alpha_{\alpha 2} = 2 & \beta_{\alpha 2} = 1. \end{array}$$

A.2.2 Bayesian Profile Regression

Following Molitor et al. (2011), we specify the following prior hyperparameters for UPR and SPR in our simulation study and data analysis:

$$\begin{array}{cccc} \alpha_\alpha = 2 & \beta_\alpha = 1 & \alpha_\kappa = \frac{7}{2} & \beta_\kappa = \frac{43.75}{2} \\ \alpha_\phi = \frac{7}{2} & \beta_\phi = \frac{43.75}{2} & \alpha_\sigma = 2.5 & \beta_\sigma = 2.5 \\ \alpha_\rho = 0.5 & \beta_\rho = 0.5 & r = p & C = 20. \end{array}$$

We also set $\boldsymbol{\nu}_0$ to the vector of empirical exposure means, $\boldsymbol{\Lambda}_0$ to the diagonal matrix where each non-zero element is the square of the observed range for each exposure, and $\mathbf{R}$ to the empirical covariance matrix of the exposure data.

A.3 Additional Simulation Study Results

Table A.2: Grouping structure in fixed profiles scenario. Table shows summary statistics for the Calinski-Harabasz index, silhouette statistic, and number of clusters to maximize the gap width (Gap clusters) for 200 simulated data sets used in the simulation study.

	Min	1st quartile	Median	Mean	3rd quartile	Max
Calinski-Harabasz	6.0647	18.1074	22.5437	21.3772	25.9485	35.0493
Silhouette	0.0015	0.1099	0.1463	0.1515	0.1878	0.2900
Gap clusters	2	4	6	6.07	9	10

Table A.3: Summary of method performance in the linear scenario. Results from simulation study across 200 simulated data sets in scenario h_1: linear. Reported values are means (standard errors) across all data sets for: root mean squared error (RMSE) and coverage (Cvg) for the exposure-response function, true selection rate for main effects (TSR), false selection rate for main effects (FSR), true selection rate for interactions (TSR_{int}), and false selection rate for interactions (FSR_{int}).

	NPBr	NPB	UPR	SPR
RMSE	1.02 (0.02)	0.54 (0.01)	2.01 (0.04)	1.59 (0.04)
Cvg	0.73 (0.01)	0.95 (0.01)	0.56 (0.01)	0.54 (0.01)
TSR	0.85 (0.01)	0.92 (0.01)	0.25 (0.02)	0.63 (0.02)
FSR	0.35 (0.02)	0.10 (0.01)	0.26 (0.02)	0.53 (0.02)
TSR_{int}	–	0.59 (0.02)	–	–
FSR_{int}	–	0.02 (0.00)	–	–

	BKMR	LM	LM-int
RMSE	0.55 (0.01)	1.01 (0.02)	0.73 (0.01)
Cvg	0.96 (0.01)	0.73 (0.01)	0.95 (0.01)
TSR	1.00 (0.00)	0.84 (0.01)	0.68 (0.01)
FSR	0.39 (0.02)	0.29 (0.02)	0.04 (0.01)
TSR_{int}	–	–	0.32 (0.02)
FSR_{int}	–	–	0.04 (0.00)

Table A.4: Summary of method performance in the nonlinear scenario. Results from simulation study across 200 simulated data sets in scenario h_2: nonlinear. Reported values are means (standard errors) across all data sets for: root mean squared error (RMSE) and coverage (Cvg) for the exposure-response function, true selection rate for main effects (TSR), false selection rate for main effects (FSR), true selection rate for interactions (TSR_{int}), and false selection rate for interactions (FSR_{int}).

	NPBr	NPB	UPR	SPR
RMSE	0.77 (0.01)	0.69 (0.01)	1.42 (0.03)	1.27 (0.03)
Cvg	0.80 (0.01)	0.86 (0.01)	0.56 (0.01)	0.58 (0.01)
TSR	0.79 (0.02)	0.78 (0.02)	0.27 (0.02)	0.68 (0.02)
FSR	0.22 (0.02)	0.16 (0.02)	0.24 (0.02)	0.58 (0.02)
TSR_{int}	–	0.25 (0.03)	–	–
FSR_{int}	–	0.01 (0.00)	–	–

	BKMR	LM	LM-int
RMSE	0.59 (0.01)	0.78 (0.01)	0.89 (0.02)
Cvg	0.92 (0.01)	0.81 (0.01)	0.91 (0.01)
TSR	0.96 (0.01)	0.78 (0.02)	0.54 (0.02)
FSR	0.48 (0.02)	0.17 (0.01)	0.08 (0.01)
TSR_{int}	–	–	0.20 (0.03)
FSR_{int}	–	–	0.07 (0.01)

Table A.5: Summary of method performance in the fixed profiles scenario. Results from simulation study across 200 simulated data sets in scenario h_3: fixed profiles. Reported values are means (standard errors) across all data sets for: root mean squared error (RMSE) and coverage (Cvg) for the exposure-response function, true selection rate for main effects (TSR), false selection rate for main effects (FSR), true selection rate for interactions (TSR_{int}), and false selection rate for interactions (FSR_{int}).

	NPBr	NPB	UPR	SPR
RMSE	1.11 (0.02)	1.02 (0.02)	1.41 (0.02)	1.38 (0.02)
Cvg	0.66 (0.01)	0.75 (0.01)	0.55 (0.01)	0.54 (0.01)
TSR	0.66 (0.02)	0.68 (0.02)	0.27 (0.03)	0.68 (0.02)
FSR	0.11 (0.01)	0.13 (0.01)	0.25 (0.02)	0.59 (0.01)
TSR_{int}	–	0.06 (0.02)	–	–
FSR_{int}	–	0.02 (0.00)	–	–

	BKMR	LM	LM-int
RMSE	0.69 (0.01)	1.13 (0.02)	0.99 (0.02)
Cvg	0.91 (0.01)	0.70 (0.01)	0.91 (0.00)
TSR	0.97 (0.01)	0.69 (0.02)	0.56 (0.02)
FSR	0.64 (0.03)	0.14 (0.01)	0.14 (0.01)
TSR_{int}	–	–	0.12 (0.02)
FSR_{int}	–	–	0.11 (0.01)

A.4 Null, Complex Mixture, and Large Sample Size Simulation Studies

A.4.1 Design

We conducted three additional simulation studies to assess robustness of our results. For all of the additional simulations, we reported results from 100 simulated data sets.

First, we included a null scenario, $h_4(\mathbf{x})$, where none of the exposures are associated with the response. That is,

$$h_4(\mathbf{x}) = 0. \tag{A.1}$$

This scenario uses the same exposure data, covariates, and residual variance described in scenarios 1-3 in the main text.

Second, we included a complex mixtures scenario, $h_5(\mathbf{x})$, where we simulated data for seven additional pollutants to have a total of 14 mixture components. For each data set, the first seven pollutants are the exposures from the FACES data set as described in scenarios 1-3 in the main text. The exposure values for the seven additional pollutants were simulated as random linear combinations of the FACES exposure data using N(0,1) weights plus N(0,1) noise. All exposures were then scaled to have mean 0 and variance 1. We simulated the response as a linear function of 10 main effects and two pairwise interactions. Specifically,

$$\begin{aligned} h_5(\mathbf{x}) \quad = \quad & x_1 - x_2 + x_3 - x_4 + 1.4x_5 + 1.5x_6 + 1.2x_7 - \\ & 1.4x_8 - 1.5x_9 - 1.2x_{10} + 0.7x_1x_2 - 0.5x_3x_4. \end{aligned} \tag{A.2}$$

The ten active mixture components $x_1, \ldots, x_{10}$ were randomly selected for each data set. All 14 pollutants were included in the models as predictors. All other details of the data generating mechanism are the same as previously described for the other scenarios.

Third, we replicated the simulation scenarios 1-3 in the main text but used a larger sample size. We repeatedly sampled from the FACES exposure and covariate data to create a sample of size $n = 1000$ for each data set. All other details are described in the main text.

A.4.2 Results

The methods performed more similarly to each other in the null scenario compared to the other scenarios (Table A.6). NPBr, NPB, and BKMR had the lowest RMSE for the exposure-response function and LM-int had the highest RMSE. All methods except SPR achieved the nominal coverage level. FSR was lowest for NPBr and NPB, meaning these methods were the best at not selecting any mixture components into the model when none are associated with the response. FSR was highest for SPR.

Results from the complex mixture scenario are shown in Table A.7. Here NPB estimated the exposure-response function with lowest RMSE and near-optimal coverage. BKMR had the next lowest RMSE. LM-int achieved the nominal coverage, but with substantially higher RMSE. NPBr and BKMR had highest TSR, followed by NPB and LM. NPB, UPR, LM, and LM-int all had mean FSR at or below 0.10. NPB outperformed LM-int in variable selection rates for interactions. Overall, NPB and BKMR were the top-performing methods in simultaneously estimating the exposure-response function and identifying active mixture components in this complex mixture scenario.

For the larger sample size simulation, our results remain generally the same as in our original simulation study (Table A.8). NPB performed best in the linear scenario, followed by BKMR and LM-int. BKMR performed best in the nonlinear and fixed profiles scenarios. TSR improved for all methods in all scenarios. With the increased sample size, UPR and SPR often selected all of the mixture components into the model, as evidenced by both high TSR and FSR.

Table A.6: Summary of method performance in the null scenario. Results from simulation study across 100 simulated data sets in the null scenario. Reported values are means (standard errors) across all data sets for: root mean squared error (RMSE) and coverage (Cvg) for the exposure-response function, false selection rate for main effects (FSR), and false selection rate for interactions (FSR_{int}). True selection rates were not reported since there were no active mixture components in the exposure-response function.

	NPBr	NPB	UPR	SPR
RMSE	0.23 (0.02)	0.24 (0.02)	0.28 (0.02)	0.56 (0.06)
Cvg	0.98 (0.02)	0.98 (0.02)	0.98 (0.01)	0.74 (0.05)
FSR	0.00 (0.00)	0.00 (0.00)	0.28 (0.03)	0.74 (0.02)
FSR_{int}	–	0.00 (0.00)	–	–

	BKMR	LM	LM-int
RMSE	0.25 (0.02)	0.44 (0.02)	0.77 (0.03)
Cvg	0.97 (0.01)	0.96 (0.01)	0.95 (0.01)
FSR	0.30 (0.04)	0.03 (0.01)	0.08 (0.01)
FSR_{int}	–	–	0.07 (0.01)

Table A.7: Summary of method performance in the complex mixture scenario. Results from simulation study across 100 simulated data sets in the complex mixture scenario. Reported values are means (standard errors) across all data sets for: root mean squared error (RMSE) and coverage (Cvg) for the exposure-response function, true selection rate for main effects (TSR), false selection rate for main effects (FSR), true selection rate for interactions (TSR_{int}), and false selection rate for interactions (FSR_{int}).

	NPBr	NPB	UPR	SPR
RMSE	1.00 (0.03)	0.69 (0.03)	3.22 (0.12)	2.97 (0.12)
Cvg	0.77 (0.02)	0.91 (0.01)	0.46 (0.01)	0.32 (0.02)
TSR	0.62 (0.02)	0.58 (0.02)	0.00 (0.00)	0.29 (0.04)
FSR	0.19 (0.03)	0.10 (0.02)	0.00 (0.00)	0.31 (0.05)
TSR_{int}	–	0.39 (0.03)	–	–
FSR_{int}	–	0.01 (0.00)	–	–

	BKMR	LM	LM-int
RMSE	0.86 (0.05)	1.00 (0.03)	1.68 (0.06)
Cvg	0.90 (0.01)	0.80 (0.02)	0.96 (0.01)
TSR	0.65 (0.02)	0.56 (0.01)	0.23 (0.01)
FSR	0.28 (0.04)	0.08 (0.01)	0.04 (0.01)
TSR_int	–	–	0.07 (0.02)
FSR_int	–	–	0.04 (0.01)

Table A.8: Summary of method performance in large sample size (n = 1000) simulation study. Results from the large sample size simulation study across 100 simulated data sets in all three exposure-response scenarios. Reported values are means across all data sets for: root mean squared error (RMSE) and coverage (Cvg) for the exposure-response function, true selection rate for main effects (TSR), false selection rate for main effects (FSR), true selection rate for interactions (TSR_{int}), and false selection rate for interactions (FSR_{int}). Results for top-performing methods are listed in bold.

Method	RMSE	Cvg	TSR	FSR	TSR_{int}	FSR_{int}
$h_1(\mathbf{x})$: linear with multiplicative interactions						
NPBr	0.91	0.37	0.96	0.69	–	–
NPB	**0.14**	**0.96**	**1.00**	**0.01**	**1.00**	**0.00**
UPR	1.53	0.28	**1.00**	1.00	–	–
SPR	1.40	0.28	**1.00**	1.00	–	–
BKMR	0.23	**0.96**	**1.00**	0.04	–	–
LM	0.90	0.37	0.96	0.68	–	–
LM-int	0.30	**0.95**	**0.99**	0.04	0.92	0.06
$h_2(\mathbf{x})$: nonlinear with multiplicative interactions						
NPBr	0.65	0.45	0.96	0.44	–	–
NPB	0.48	0.67	0.96	**0.24**	**0.74**	**0.20**
UPR	1.08	0.32	**0.99**	1.00	–	–
SPR	1.10	0.33	**1.00**	1.00	–	–
BKMR	**0.29**	**0.92**	**1.00**	**0.25**	–	–
LM	0.65	0.46	0.95	0.49	–	–
LM-int	0.58	0.71	0.86	0.27	0.62	0.25
$h_3(\mathbf{x})$: constant function of fixed profiles						
NPBr	1.08	0.33	0.87	**0.50**	–	–
NPB	0.75	0.57	0.93	**0.51**	0.54	**0.39**
UPR	1.15	0.35	**0.99**	1.00	–	–
SPR	1.17	0.35	**0.99**	1.00	–	–
BKMR	**0.43**	**0.87**	**0.99**	0.67	–	–
LM	1.09	0.33	0.89	0.54	–	–
LM-int	0.77	0.63	0.83	0.53	**0.58**	0.44

A.5 Additional Data Analysis Results

Table A.9: Additional results from analysis of FACES data set using LM-int. Table shows main effect and interaction regression coefficient estimates ($\hat{\beta}$), 95% confidence intervals (CI), and p-values. The regression coefficient $\hat{\beta}$ is the expected change in FEV_1 for a 1 standard deviation increase in the square root transformed exposures.

	$\hat{\beta}$	95% CI	p-value
Main Effects			
C	0.05	(-0.08 , 0.19)	0.44
MeBr	0.17	(0.05 , 0.29)	0.01
OP	0.02	(-0.17 , 0.22)	0.80
O_3	-0.13	(-0.32 , 0.06)	0.17
NO_2	-0.68	(-1.10 , -0.25)	0.00
$PM_{2.5}$	-0.11	(-0.48 , 0.26)	0.55
PM_{10}	0.50	(0.08 , 0.93)	0.02
Interactions			
C:MeBr	-0.04	(-0.14 , 0.07)	0.51
C:OP	0.15	(-0.18 , 0.47)	0.38
C:O_3	-0.01	(-0.18 , 0.16)	0.91
C:NO_2	-0.06	(-0.35 , 0.23)	0.67
C:$PM_{2.5}$	0.28	(0.01 , 0.54)	0.04
C:PM_{10}	-0.08	(-0.31 , 0.14)	0.48
MeBr:OP	0.01	(-0.26 , 0.28)	0.93
MeBr:O_3	-0.03	(-0.20 , 0.15)	0.77
MeBr:NO_2	-0.11	(-0.43 , 0.21)	0.50
MeBr:$PM_{2.5}$	0.18	(-0.06 , 0.42)	0.14
MeBr:PM_{10}	0.08	(-0.11 , 0.28)	0.41
OP:O_3	-0.04	(-0.20 , 0.12)	0.63
OP:NO_2	-0.10	(-0.33 , 0.12)	0.37
OP:$PM_{2.5}$	-0.23	(-0.58 , 0.12)	0.19
OP:PM_{10}	0.31	(-0.01 , 0.62)	0.05
O_3:NO_2	-0.12	(-0.54 , 0.29)	0.56
O_3:$PM_{2.5}$	0.04	(-0.23 , 0.30)	0.78
O_3:PM_{10}	-0.02	(-0.31 , 0.27)	0.88
NO_2:$PM_{2.5}$	-0.27	(-0.70 , 0.16)	0.21
NO_2:PM_{10}	0.33	(-0.05 , 0.72)	0.09
$PM_{2.5}$:PM_{10}	0.01	(-0.38 , 0.40)	0.95

Table A.10: Additional results from analysis of FACES data set using NPB. Table shows main effect and interaction regression coefficient estimates ($\hat{\beta}$), 95% credible intervals, and posterior inclusion probabilities (PIP). The regression coefficient $\hat{\beta}$ is the expected change in FEV_1 for a 1 standard deviation increase in the square root transformed exposures.

	$\hat{\beta}$	95% CI	PIP
Main Effects			
C	0.00	(0.00 , 0.03)	0.07
MeBr	0.00	(-0.01 , 0.00)	0.06
OP	0.01	(0.00 , 0.11)	0.16
O_3	-0.01	(-0.12 , 0.01)	0.11
NO_2	-0.12	(-0.36 , 0.00)	0.60
$PM_{2.5}$	0.00	(-0.09 , 0.05)	0.12
PM_{10}	0.02	(-0.01 , 0.2)	0.19
Interactions			
C:MeBr	0.00	(0.00 , 0.00)	0.02
C:OP	0.00	(0.00 , 0.00)	0.02
C:O_3	0.00	(0.00 , 0.00)	0.01
C:NO_2	0.00	(0.00 , 0.00)	0.01
C:$PM_{2.5}$	0.00	(0.00 , 0.00)	0.01
C:PM_{10}	0.00	(0.00 , 0.00)	0.01
MeBr:OP	0.00	(0.00 , 0.00)	0.01
MeBr:O_3	0.00	(0.00 , 0.00)	0.01
MeBr:NO_2	0.00	(0.00 , 0.00)	0.01
MeBr:$PM_{2.5}$	0.00	(0.00 , 0.00)	0.02
MeBr:PM_{10}	0.00	(0.00 , 0.00)	0.02
OP:O_3	0.00	(0.00 , 0.00)	0.02
OP:NO_2	0.00	(0.00 , 0.00)	0.01
OP:$PM_{2.5}$	0.00	(0.00 , 0.00)	0.01
OP:PM_{10}	0.00	(0.00 , 0.00)	0.01
O_3:NO_2	0.00	(0.00 , 0.00)	0.01
O_3:$PM_{2.5}$	0.00	(0.00 , 0.00)	0.01
O_3:PM_{10}	-0.01	(-0.09 , 0.00)	0.06
NO_2:$PM_{2.5}$	0.01	(0.00 , 0.16)	0.11
NO_2:PM_{10}	0.01	(0.00 , 0.13)	0.12
$PM_{2.5}$:PM_{10}	0.00	(0.00 , 0.06)	0.05

Table A.11: Variable selection results from FACES data analysis using BKMR with component-wise variable selection. Table shows posterior inclusion probabilities (PIP) for each exposure.

Exposure	PIP
C	0.08
MeBr	0.06
OP	0.10
O_3	0.16
NO_2	0.96
$PM_{2.5}$	0.20
PM_{10}	0.34

Table A.12: Variable selection results from FACES data analysis using BKMR with hierarchical variable selection. Table shows posterior inclusion probabilities for each group (Group PIP) as well as conditional posterior inclusion probabilities for each exposure given the group to which it belongs is included (Conditional PIP). Component-wise PIPs are calculated from the group and conditional PIPs by multiplying the group PIP by the conditional PIP for each exposure.

Exposure	Group PIP	Conditional PIP	Component-wise PIP
C	0.20	0.18	0.03
MeBr	0.20	0.13	0.03
OP	0.20	0.70	0.14
O_3	0.98	0.00	0.00
NO_2	0.98	0.98	0.96
$PM_{2.5}$	0.98	0.01	0.01
PM_{10}	0.98	0.00	0.00

Table A.13: Variable selection results from FACES data analysis using UPR and SPR. Table shows posterior inclusion probabilities (PIP) for each exposure in each method.

	PIP	
Exposure	UPR	SPR
C	0.03	0.02
MeBr	0.21	0.71
OP	0.57	0.51
O_3	0.54	0.75
NO_2	0.61	0.67
$PM_{2.5}$	0.56	0.63
PM_{10}	0.24	0.03

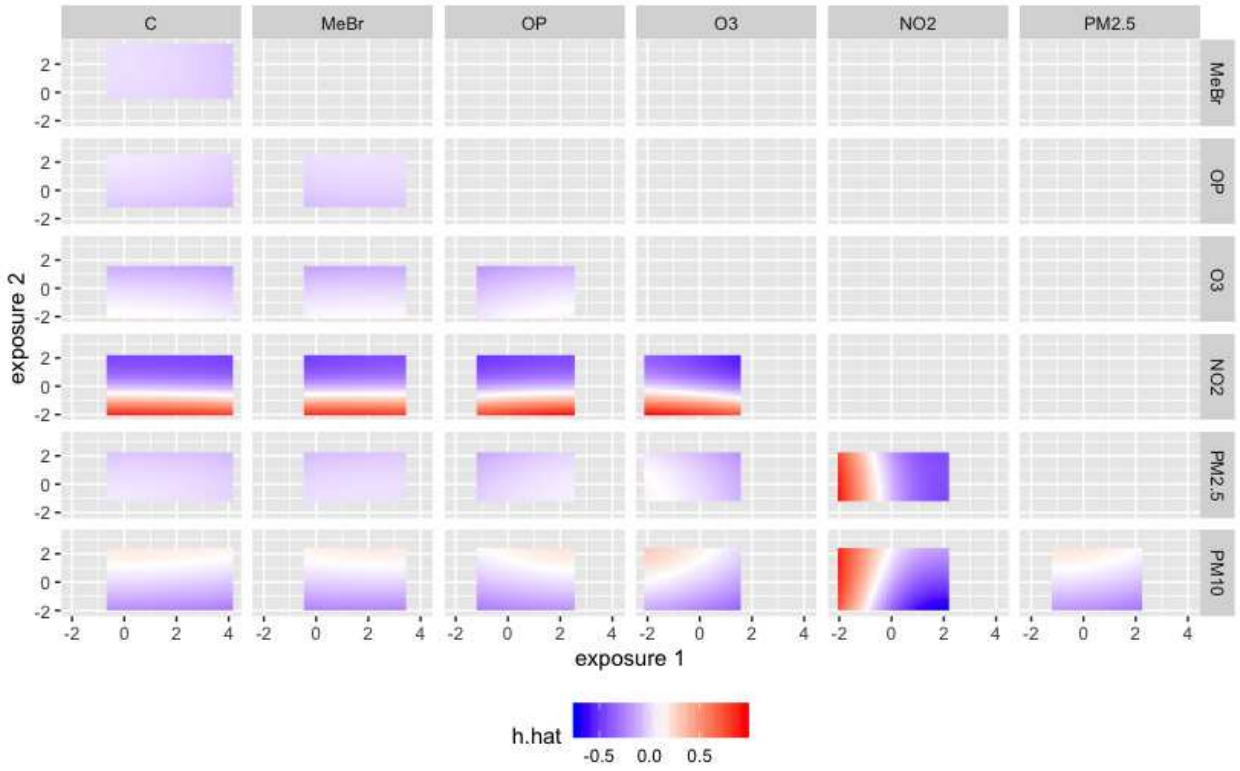


Figure A.1: Estimated bivariate exposure-response function from FACES data analysis using BKMR. Each grid panel is an image plot of the predicted exposure-response function $\hat{h}$ for varying levels of two exposures, while holding all other exposures at their median value. Evidence of an interaction would be reflected by changes in the predicted exposure-response function with changes in the levels of both exposure 1 and exposure 2. Figure shows no notable evidence of interactions, but does depict the main effect of NO_2.

Appendix B

Infinite Hidden Markov Models for Multiple Multivariate Time Series with Missing Data

B.1 Convergence Diagnostics for Simulation Study

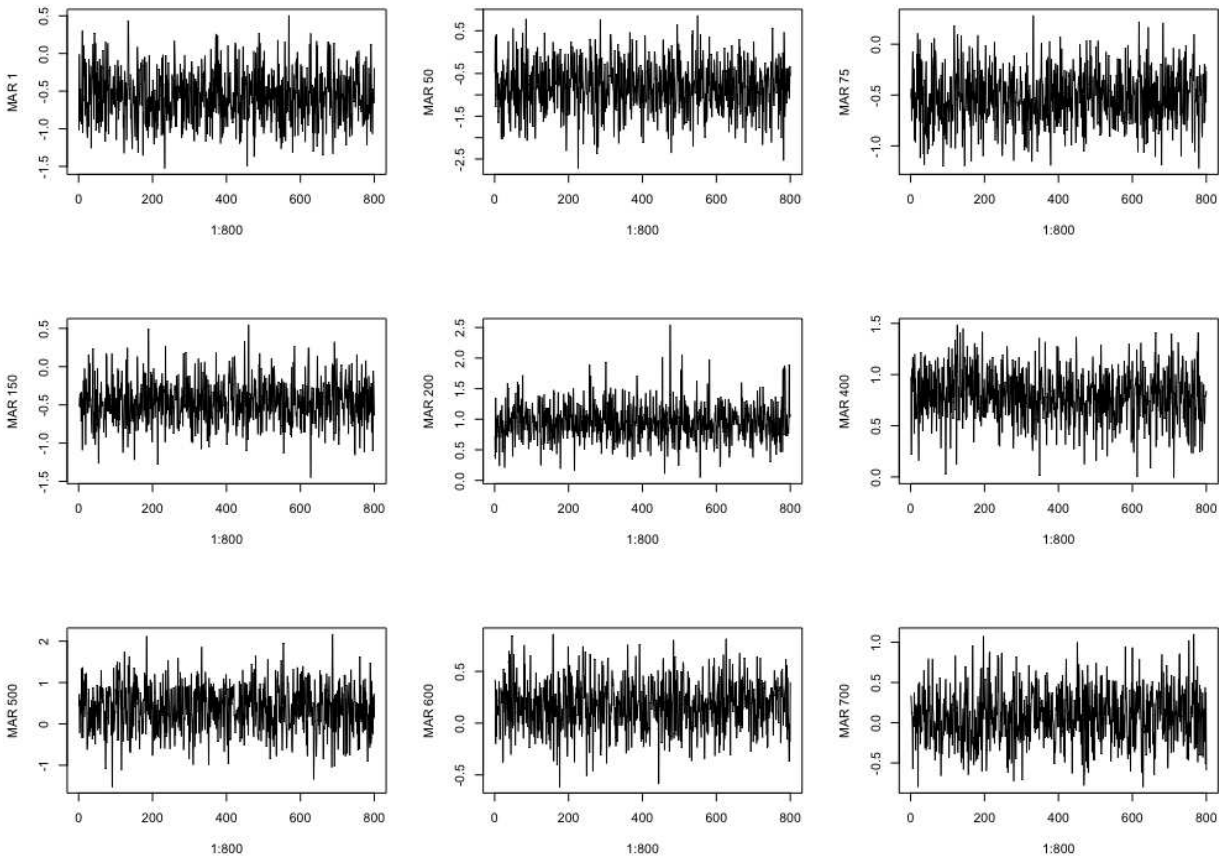


Figure B.1: Convergence diagnostics from our proposed joint cyclical model in the simulation study. The figure shows traceplots of imputed values for a sample of 9 missing at random (MAR) observations. Imputations were taken at 800 equally spaced iterations in an MCMC chain of 10,000 total iterations.

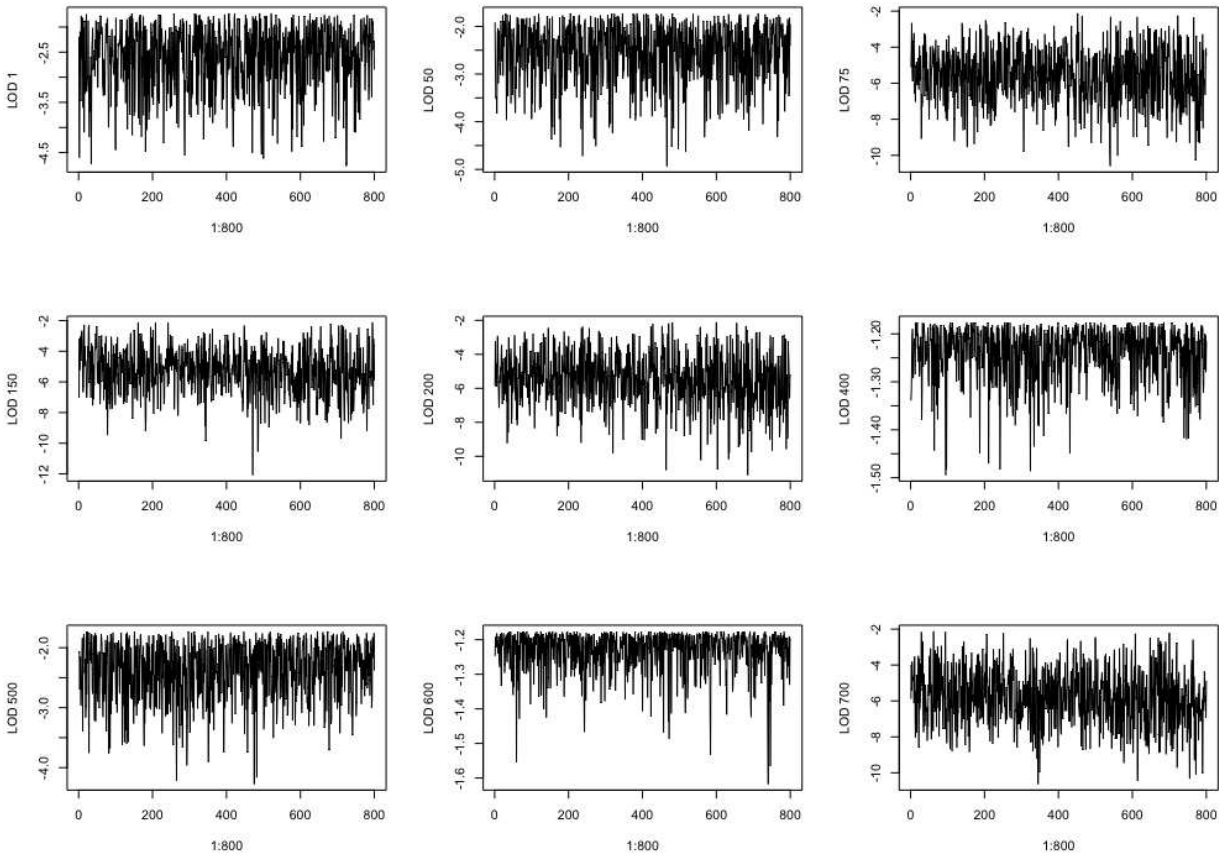


Figure B.2: Convergence diagnostics from our proposed joint cyclical model in the simulation study. The figure shows traceplots of imputed values for a sample of 9 below LOD observations. Imputations were taken at 800 equally spaced iterations in an MCMC chain of 10,000 total iterations.

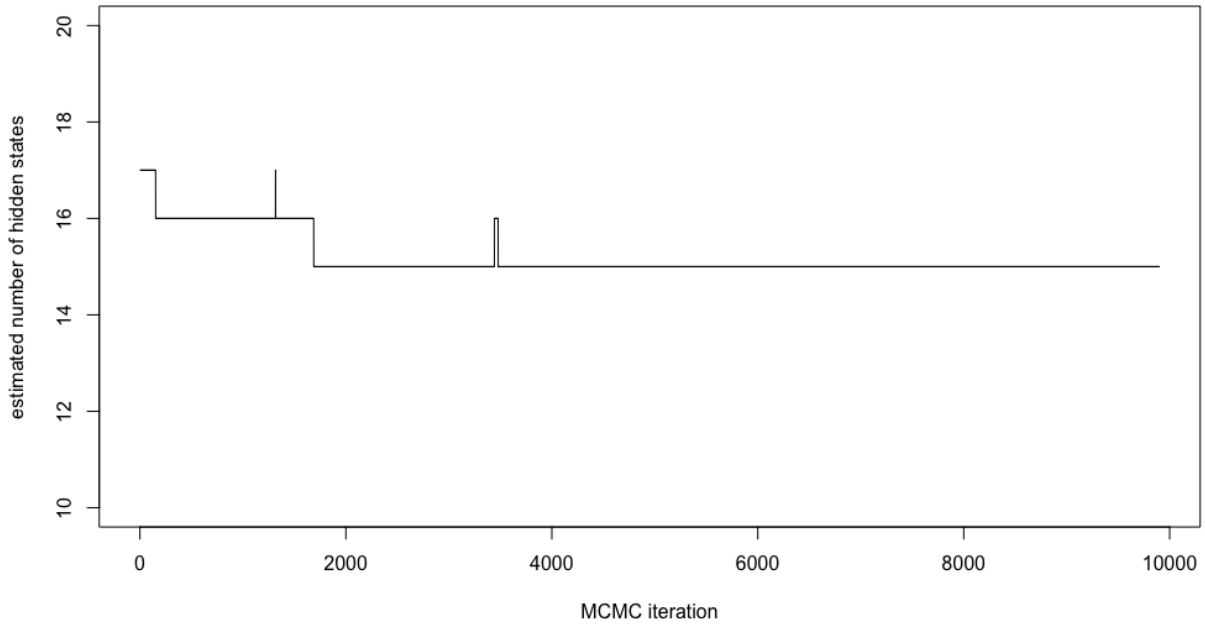


Figure B.3: Convergence diagnostics from our proposed joint cyclical model in the simulation study. The figure show a traceplot of estimated number of hidden states for 10,000 iterations of the MCMC sampler.

B.2 Formula for Calculating Mean Squared Error for Estimated State-Specific Means

Here we describe the formula for calculating MSE for state-specific means in our simulation study. Let N be the total number of sampling days (i.e. $N = \sum_{i=1}^{n} S_i$). For iterations $b = 1, \ldots, B$ post burn-in, we calculated

$$\boldsymbol{\mu}_{\text{MSE}} = \left(\frac{1}{BNTp}\right) \sum_{b=1}^{B} \sum_{i=1}^{n} \sum_{s=1}^{S_i} \sum_{t=1}^{T} \left(\hat{\boldsymbol{\mu}}_{\hat{z}_{ist}}^{(b)} - \boldsymbol{\mu}_{z_{ist}}\right)' \left(\hat{\boldsymbol{\mu}}_{\hat{z}_{ist}}^{(b)} - \boldsymbol{\mu}_{z_{ist}}\right), \tag{B.1}$$

where $\hat{\boldsymbol{\mu}}_{\hat{z}_{ist}}^{(b)}$ is the vector of estimated exposure means for the state to which observation $\mathbf{y}_{ist}$ is assigned at iteration b and $\boldsymbol{\mu}_{z_{ist}}$ is the vector of true exposure means for the state in which observation $\mathbf{y}_{ist}$ truly belongs.

B.3 Additional Simulation Study Results

Table B.1: Standard errors from the shared trends scenario simulation study. The two variations of our proposed joint iHMM approach are the model with cyclical trends (joint cyclical) and the model with no covariates (joint no covariates). We include the model with cyclical trends fit independently to each time series (indep. cyclical) and the model with no covariates fit independently to each time series (indep. no covariates). Last is the Dirichlet process mixture model (joint DPMM) fit jointly to all time series. The table shows the standard error for the following measures: estimated number of hidden states ($\hat{K}$); Hamming distance, which is a measure of the distance between the estimated hidden state trajectories and the true hidden state trajectories; MSE for the state-specific means ($\boldsymbol{\mu}_{\text{MSE}}$); MSE and bias for the MAR and below LOD data imputations. Results are shown for four levels of missing data: 0%, 5%, 10%, and 20%.

	Method	$\hat{K}$	Hamming	$\boldsymbol{\mu}_{\text{MSE}}$	MAR MSE	LOD MSE	MAR bias	LOD bias
0%	joint cyclical	0.34	0.02	0.01	–	–	–	–
	joint no covariates	0.38	0.01	0.01	–	–	–	–
	indep. cyclical	1.55	0.01	0.01	–	–	–	–
	indep. no covariates	1.68	0.00	0.01	–	–	–	–
	joint DPMM	1.25	0.02	0.01	–	–	–	–
5%	joint cyclical	0.37	0.02	0.02	0.04	0.82	0.01	0.12
	joint no covariates	0.29	0.02	0.01	0.07	0.93	0.02	0.13
	indep. cyclical	1.77	0.01	0.01	0.04	0.58	0.01	0.08
	indep. no covariates	1.99	0.01	0.01	0.04	0.51	0.01	0.07
	joint DPMM	1.61	0.01	2.12	30.93	96.85	0.68	1.37
10%	joint cyclical	0.40	0.02	0.09	0.19	2.32	0.01	0.20
	joint no covariates	0.42	0.02	0.05	0.13	1.22	0.02	0.15
	indep. cyclical	1.86	0.01	0.02	0.04	0.79	0.01	0.12
	indep. no covariates	1.77	0.01	0.02	0.04	0.60	0.01	0.10
	joint DPMM	1.45	0.02	5.83	34.34	115.52	0.60	1.74
20%	joint cyclical	0.42	0.02	0.07	0.13	0.87	0.02	0.14
	joint no covariates	0.38	0.02	0.05	0.08	0.65	0.02	0.11
	indep. cyclical	2.38	0.01	0.03	0.11	0.94	0.01	0.12
	indep. no covariates	1.66	0.01	0.02	0.06	0.76	0.01	0.11
	joint DPMM	1.20	0.01	11.90	52.05	123.41	0.95	1.30

Table B.2: Results from the distinct trends scenario simulation study. The two variations of our proposed joint iHMM approach are the model with cyclical trends (joint cyclical) and the model with no covariates (joint no covariates). We include the model with cyclical trends fit independently to each time series (indep. cyclical) and the model with no covariates fit independently to each time series (indep. no covariates). Last is the Dirichlet process mixture model (joint DPMM) fit jointly to all time series. The table shows the following measures: mean estimated number of hidden states ($\hat{K}$); mean Hamming distance, which is a measure of the distance between the estimated hidden state trajectories and the true hidden state trajectories; mean MSE for the state-specific means ($\boldsymbol{\mu}_{\text{MSE}}$); mean MSE and bias for the MAR and below LOD data imputations. Results are shown for four levels of missing data: 0%, 5%, 10%, and 20%.

	Method	$\hat{K}$	Hamming	$\boldsymbol{\mu}_{\text{MSE}}$	MAR MSE	LOD MSE	MAR bias	LOD bias
0%	joint cyclical	11.67	0.44	0.11	–	–	–	–
	joint no covariates	11.18	0.45	0.10	–	–	–	–
	indep. cyclical	122.82	0.52	0.37	–	–	–	–
	indep. no covariates	94.92	0.61	0.45	–	–	–	–
	joint DPMM	27.33	0.38	0.06	–	–	–	–
5%	joint cyclical	11.45	0.50	0.11	0.60	3.08	-0.08	-0.85
	joint no covariates	11.20	0.51	0.11	0.63	0.72	-0.03	-0.46
	indep. cyclical	127.19	0.51	0.28	1.03	5.63	-0.05	-1.28
	indep. no covariates	102.02	0.59	0.35	1.18	4.56	-0.05	-1.15
	joint DPMM	47.21	0.60	6.93	86.17	300.86	-2.13	-6.52
10%	joint cyclical	11.43	0.54	0.23	0.83	4.65	-0.10	-1.20
	joint no covariates	11.38	0.52	0.27	1.05	5.74	-0.11	-1.27
	indep. cyclical	121.55	0.53	0.38	1.32	9.68	-0.08	-1.66
	indep. no covariates	95.50	0.62	0.44	1.44	8.05	-0.09	-1.47
	joint DPMM	57.33	0.60	24.46	165.96	504.72	-4.41	-10.37
20%	joint cyclical	11.92	0.60	0.54	1.20	6.10	-0.13	-1.16
	joint no covariates	12.03	0.60	1.10	2.31	12.34	-0.21	-1.67
	indep. cyclical	111.00	0.57	0.46	1.46	6.55	-0.11	-1.28
	indep. no covariates	83.20	0.67	0.55	1.62	5.37	-0.10	-1.09
	joint DPMM	60.80	0.66	59.39	264.84	496.36	-7.43	-10.69

B.4 Sensitivity Analysis of Multiple Imputation Approach

In our validation study using FCCS data (Section 3.5.1), we assessed the sensitivity of our imputation approach to the hyperparameter λ in the emission distribution. In the main analysis, we used $\lambda = 10$. As a sensitivity analysis, we tested the additional values of $\lambda = 1$ and $\lambda = 25$. Results were similar when $\lambda = 25$ (Table B.3). When $\lambda = 1$ (Table B.4), we obtained a few data

sets with very high MSE for imputations in all of our proposed iHMMs. Our method estimates an unknown number of hidden states, where both the number of hidden states and the size of each state are estimated from the data. Some very small states can be estimated. In these small states with little data informing the emission distribution, the prior distribution is very influential. When $\lambda = 1$, the prior emission distribution is such that the variation of the state-specific means around the prior mean is just as large as the variation of the data within each state. Hence, small states can have estimated means that are far away from the prior mean when there is high variability in the data or a large percentage of missing data within the state. We found that the data sets with particularly poor imputations estimated more small clusters with a large percentage of missing data than did data sets with better imputations. The fixed-state methods were not sensitive to the specification of λ because the large amount of data in each state (i.e. all data pooled together or pooled within each microenvironment) outweighs the prior. For applications, the hyperparameter λ can be chosen through a validation study such as this one or it can be modeled as a state-specific parameter.

Table B.3: Results from the sensitivity analysis of the imputation validation using FCCS data with $\lambda = 25$. The table shows the minimum (min), median, mean, and maximum (max) mean squared error (MSE) for imputations of MAR and below LOD data. The five variations of our proposed joint iHMM approach include the model with no covariates (joint no covariates), the model with cyclical trends (joint cyclical), the model with subject-specific cyclical trends (joint s.s. cyclical), the model with microenvironments as categorical predictors (joint microenv.), and the model with subject-specific microenvironment effects (joint s.s. microenv.) In the pooled approach, a single multivariate normal distribution was fit to all data. In the stratified approach, multivariate normal distributions were fit to all data within each FCCS assigned microenvironment. Last is the Dirichlet process mixture model (joint DPMM) fit jointly to all time series.

	MAR MSE				LOD MSE			
	min	median	mean	max	min	median	mean	max
joint no covariates	1.11	1.24	1.30	1.92	0.87	2.10	2.17	5.29
joint cyclical	1.05	1.30	1.31	1.79	0.64	2.15	2.18	5.63
joint s.s. cyclical	0.99	1.15	1.19	1.49	0.97	2.31	2.39	4.12
joint microenv.	1.06	1.29	1.29	1.54	1.02	1.90	2.54	6.61
joint s.s. microenv.	1.05	1.23	1.23	1.47	1.04	2.11	2.45	4.74
pooled	2.12	2.18	2.20	2.31	1.10	1.13	1.13	1.19
stratified	1.97	2.04	2.05	2.18	1.10	1.12	1.13	1.19
joint DPMM	338.25	626.41	662.13	1407.50	777.49	1305.84	1736.05	6017.40

Table B.4: Results from the sensitivity analysis of the imputation validation using FCCS data with $\lambda = 1$. The table shows the minimum (min), median, mean, and maximum (max) mean squared error (MSE) for imputations of MAR and below LOD data. The five variations of our proposed joint iHMM approach include the model with no covariates (joint no covariates), the model with cyclical trends (joint cyclical), the model with subject-specific cyclical trends (joint s.s. cyclical), the model with microenvironments as categorical predictors (joint microenv.), and the model with subject-specific microenvironment effects (joint s.s. microenv.) In the pooled approach, a single multivariate normal distribution was fit to all data. In the stratified approach, multivariate normal distributions were fit to all data within each FCCS assigned microenvironment. Last is the Dirichlet process mixture model (joint DPMM) fit jointly to all time series.

	MAR MSE				LOD MSE			
	min	median	mean	max	min	median	mean	max
joint no covariates	0.91	1.63	6.47	68.71	0.91	3.58	70.56	858.42
joint cyclical	0.86	1.34	3.39	28.00	0.71	2.88	25.49	386.98
joint s.s. cyclical	0.84	1.36	3.63	19.58	0.67	4.72	72.71	698.08
joint microenv.	0.99	1.35	2.55	8.40	0.79	4.60	15.36	93.12
joint s.s. microenv.	0.87	1.16	1.99	12.34	0.67	1.96	12.10	131.42
pooled	2.12	2.18	2.20	2.31	1.11	1.13	1.13	1.19
stratified	1.97	2.04	2.05	2.18	1.10	1.12	1.12	1.18
joint DPMM	36.08	50.46	61.36	124.03	87.19	137.65	170.58	494.62

B.5 Convergence Diagnostics for Case Study

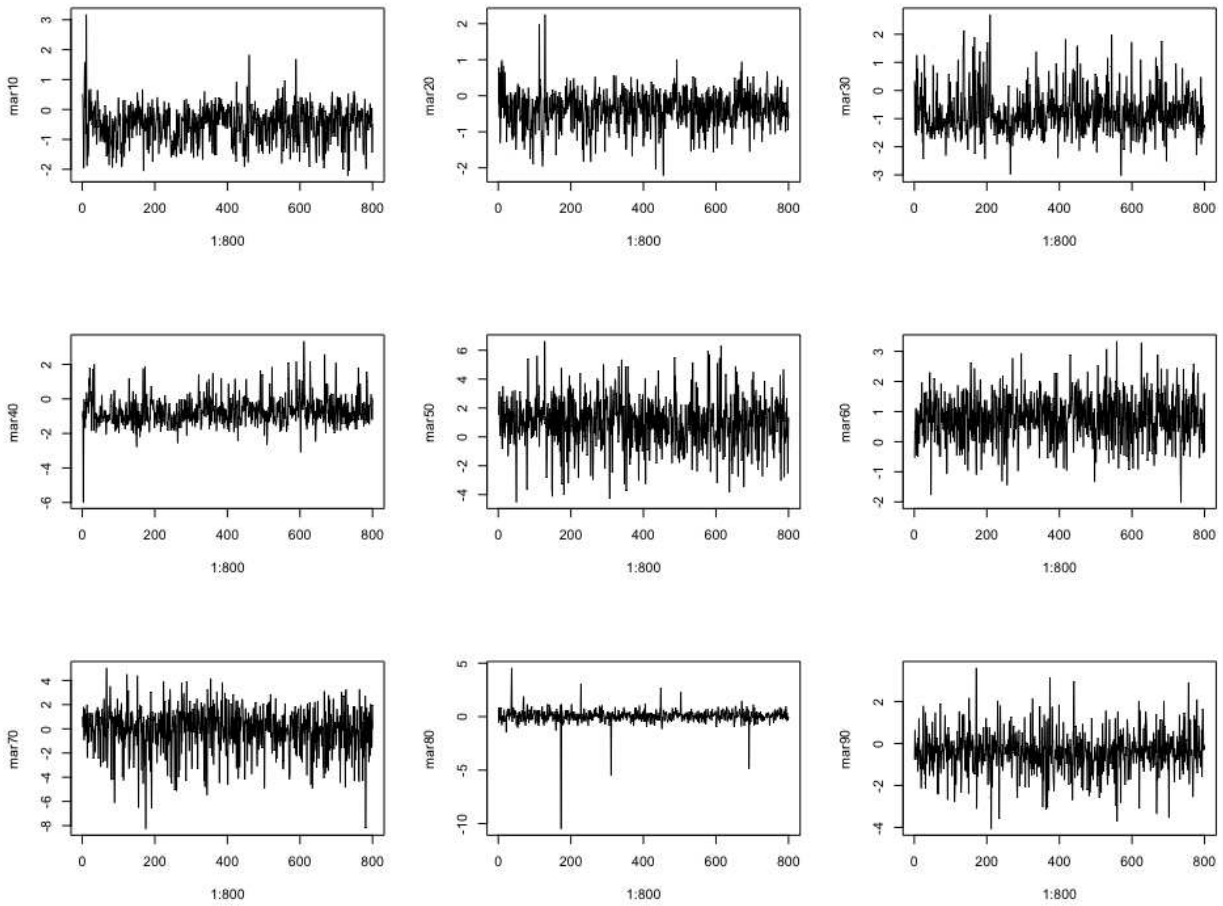


Figure B.4: Convergence diagnostics from our proposed joint subject-specific cyclical model in the case study of the Fort Collins Commuter Study data. The figure shows traceplots of imputed values for a sample of 9 missing at random (MAR) observations. Imputations were taken at 800 equally spaced iterations in an MCMC chain of 10,000 total iterations.

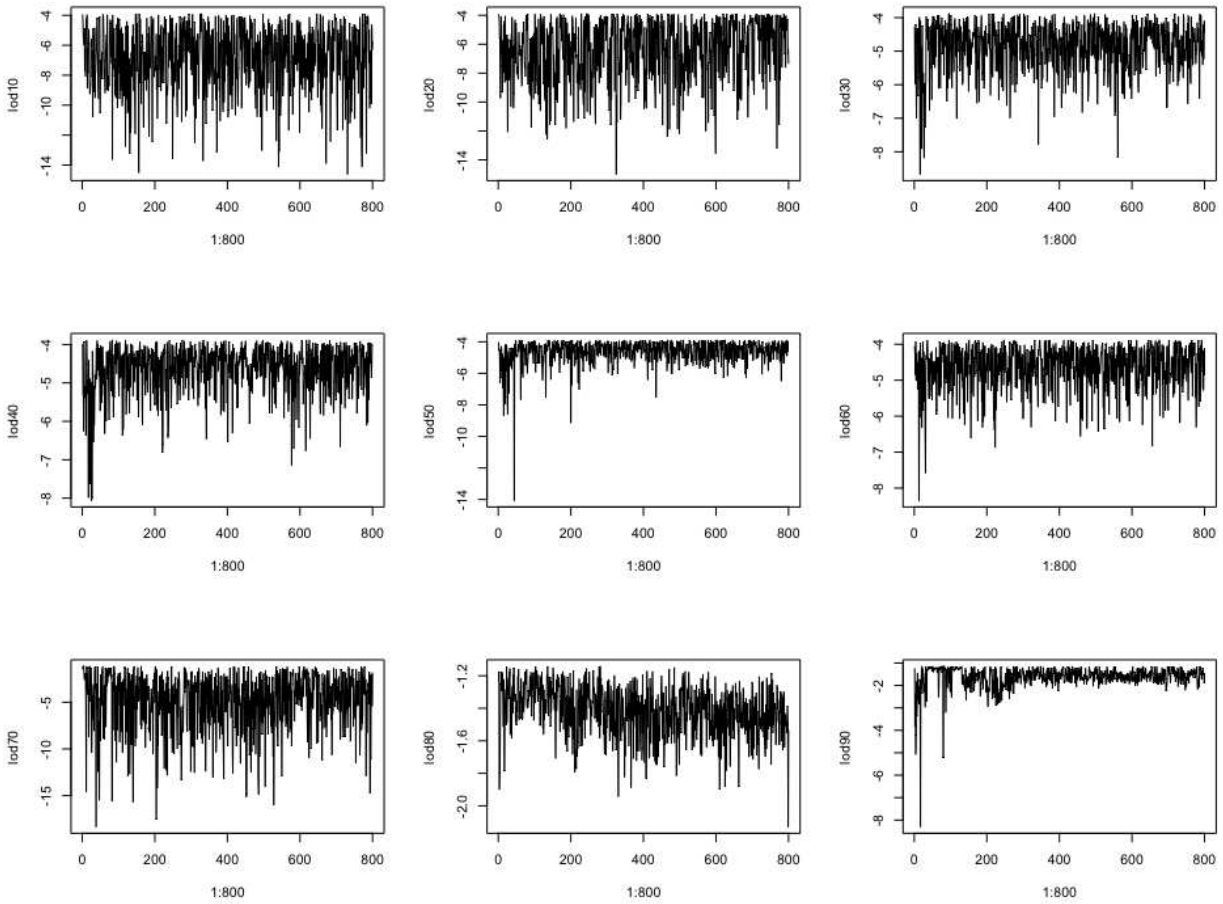


Figure B.5: Convergence diagnostics from our proposed joint subject-specific cyclical model in the case study of the Fort Collins Commuter Study data. The figure shows traceplots of imputed values for a sample of 9 below limit of detection (LOD) observations. Imputations were taken at 800 equally spaced iterations in an MCMC chain of 10,000 total iterations.

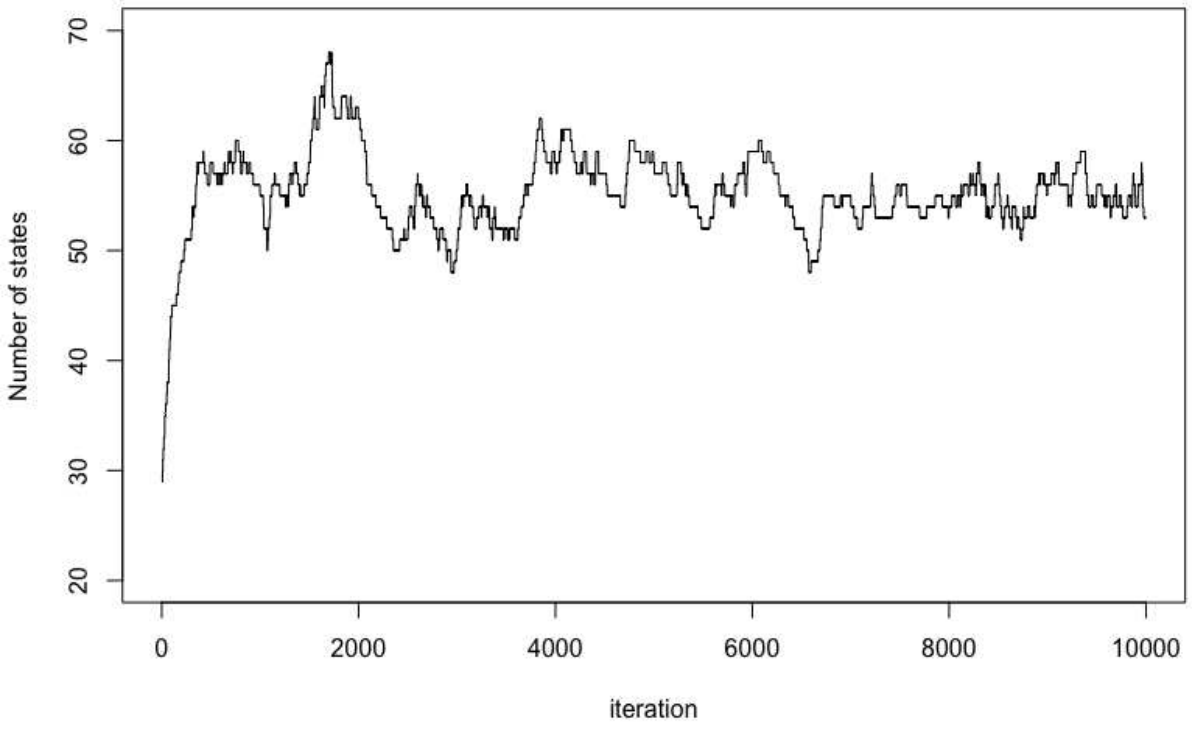


Figure B.6: Convergence diagnostics from our proposed joint subject-specific cyclical model in the case study of the Fort Collins Commuter Study data. The figure show a traceplot of estimated number of hidden states for 10,000 iterations of the MCMC sampler.

B.6 Additional Case Study Results

Table B.5: Results from the joint subject-specific cyclical model applied to the FCCS data. The table shows the number of sampling days, unique people, and time points assigned to each of the 53 hidden states estimated from the data.

hidden state	days	people	time points	hidden state	days	people	time points
1	43	9	1196	28	24	9	104
2	38	9	1193	29	12	6	89
3	33	9	999	30	5	5	83
4	32	9	897	31	16	7	81
5	40	9	864	32	9	5	63
6	33	9	800	33	4	3	35
7	34	9	643	34	7	4	31
8	38	9	637	35	6	4	27
9	36	9	574	36	20	7	27
10	32	9	541	37	4	2	22
11	35	9	485	38	7	3	21
12	31	9	458	39	8	6	19
13	41	9	446	40	5	4	17
14	23	8	439	41	11	6	15
15	29	9	420	42	12	6	14
16	28	9	411	43	3	1	13
17	29	9	389	44	9	7	13
18	22	8	332	45	10	7	12
19	25	9	305	46	8	5	12
20	30	9	258	47	7	5	10
21	30	9	243	48	9	4	10
22	29	9	236	49	7	4	8
23	11	5	236	50	6	4	6
24	16	7	184	51	6	4	6
25	28	8	183	52	2	2	6
26	17	8	145	53	4	4	4
27	17	8	138				

Appendix C

Association Between Air Pollution and COVID-19 Disease Severity via Bayesian Multinomial Logistic Regression with Partially Missing Outcomes

C.1 Demographic Characteristics of the Sample

Table C.1: Demographic characteristics of the Denver, CO COVID-19 cohort.

	n = 55273
Peak severity, %	
Asymptomatic	5.4
Symptomatic	28.3
Hospitalized	2.1
Admitted to ICU	0.4
Placed on mechanical ventilator	0.4
Death	1.2
Partially Unknown	62.2
Age (years), mean (SD)	37.1 (18.3)
Gender, %	
Male	48.1
Female	50.9
Other	<0.01
Unknown	1.0
Race/Ethnicity, %	
Non-Hispanic Black	5.6
Non-Hispanic White	30.3
Hispanic	40.1
Asian	2.4
American Indian	0.7
Multiple	2.5
Unknown	18.4
Pregnant, %	
Yes	0.7
No	63.4
Unknown	35.9
Census-tract variables, mean (SD)	
Median income	39,473 (13,889)
Percent unemployed	3.8 (2.5)
Percent low education	13.8 (12.0)
Percent poverty	13.9 (8.1)
Annual average exposures, mean (SD)	
$PM_{2.5}$ ($\mu g/m^3$)	7.4 (0.2)
1-hour daily maximum ozone (ppb)	48.9 (1.1)
Temperature (degrees Fahrenheit)	51.4 (0.5)

C.2 Additional Simulation Study Results

Table C.2: Simulation study results for the data probabilities setting. Table shows mean across 500 data sets for each measure in four simulation scenarios (“partially missing, signal," “fully missing, signal," “partially missing, null," and “fully missing, null"). The measures are root mean squared error (RMSE), bias, 95% credible interval width (width), and coverage (cov) for covariate regression coefficients. The table shows results from our proposed method and the complete case analysis for missing data levels of 0%, 20%, 50%, and 80%.

		proposed method				complete case analysis			
		RMSE	bias	width	cov	RMSE	bias	wid	cov
partially missing, signal	0%	0.29	0.00	0.71	0.95	0.29	0.00	0.71	0.95
	20%	0.31	0.00	0.75	0.95	0.32	0.00	0.79	0.95
	50%	0.36	0.00	0.85	0.94	0.39	0.00	0.97	0.95
	80%	0.44	0.00	1.01	0.93	0.55	-0.00	1.41	0.95
fully missing, signal	0%	0.29	0.00	0.71	0.95	0.29	0.00	0.71	0.95
	20%	0.32	0.00	0.76	0.95	0.32	0.00	0.79	0.95
	50%	0.38	0.00	0.89	0.93	0.39	0.00	0.97	0.95
	80%	0.51	0.00	1.20	0.91	0.55	0.00	1.41	0.96
partially missing, null	0%	0.26	0.00	0.58	0.94	0.26	0.00	0.58	0.94
	20%	0.30	0.00	0.64	0.94	0.30	0.00	0.66	0.94
	50%	0.36	0.00	0.79	0.93	0.40	0.00	0.88	0.94
	80%	0.46	0.00	1.06	0.94	0.58	0.01	1.41	0.95
fully missing, null	0%	0.26	0.00	0.58	0.94	0.26	0.00	0.58	0.94
	20%	0.30	0.00	0.65	0.93	0.30	0.00	0.66	0.94
	50%	0.37	0.00	0.81	0.93	0.40	0.00	0.88	0.94
	80%	0.48	0.00	1.16	0.93	0.58	0.01	1.41	0.95

Table C.3: Simulation study results for the equal probabilities setting. Table shows mean across 500 data sets for each measure in four simulation scenarios ("partially missing, signal," "fully missing, signal," "partially missing, null," and "fully missing, null"). The measures are root mean squared error (RMSE), bias, 95% credible interval width (width), and coverage (cov) for covariate regression coefficients. The table shows results from our proposed method and the complete case analysis for missing data levels of 0%, 20%, 50%, and 80%.

		proposed method				complete case analysis			
		RMSE	bias	width	cov	RMSE	bias	wid	cov
partially missing, signal	0%	0.11	0.00	0.29	0.94	0.11	0.00	0.29	0.94
	20%	0.11	0.00	0.30	0.94	0.12	0.00	0.32	0.95
	50%	0.13	0.00	0.34	0.93	0.15	0.00	0.40	0.95
	80%	0.16	0.00	0.41	0.92	0.23	0.01	0.63	0.95
fully missing, signal	0%	0.11	0.00	0.29	0.94	0.11	0.00	0.29	0.94
	20%	0.12	0.00	0.31	0.94	0.12	0.00	0.32	0.95
	50%	0.14	0.00	0.38	0.92	0.15	0.00	0.40	0.95
	80%	0.22	0.01	0.55	0.90	0.23	0.01	0.63	0.95
partially missing, null	0%	0.06	-0.00	0.16	0.95	0.06	-0.00	0.16	0.95
	20%	0.06	-0.00	0.18	0.94	0.07	-0.00	0.18	0.95
	50%	0.08	-0.00	0.21	0.93	0.08	-0.00	0.23	0.95
	80%	0.10	-0.00	0.26	0.93	0.14	-0.01	0.38	0.94
fully missing, null	0%	0.06	0.00	0.16	0.95	0.06	0.00	0.16	0.95
	20%	0.07	0.00	0.18	0.94	0.07	0.00	0.18	0.95
	50%	0.08	0.00	0.22	0.93	0.08	0.00	0.23	0.95
	80%	0.13	0.00	0.33	0.91	0.14	-0.01	0.38	0.94

Table C.4: Summary of imputation performance in the data probabilities setting. Results are shown for 20% and 50% missing data and four simulation scenarios (“partially missing, signal," “fully missing, signal," “partially missing, null," and “fully missing, null"). The table shows mean across 500 data sets for precision and recall for each outcome category.

			outcome category					
			1	2	3	4	5	6
partially missing, signal	20% missing	precision	0.92	0.70	0.54	0.45	0.33	0.32
		recall	0.92	0.69	0.54	0.44	0.33	0.33
	50% missing	precision	0.92	0.69	0.54	0.44	0.33	0.32
		recall	0.92	0.69	0.54	0.44	0.33	0.32
fully missing, signal	20% missing	precision	0.85	0.48	0.31	0.24	0.15	0.15
		recall	0.85	0.48	0.31	0.23	0.15	0.15
	50% missing	precision	0.85	0.48	0.31	0.23	0.15	0.14
		recall	0.85	0.48	0.31	0.23	0.15	0.14
partially missing, null	20% missing	precision	0.88	0.50	0.28	0.14	0.07	0.07
		recall	0.88	0.50	0.28	0.13	0.07	0.07
	50% missing	precision	0.88	0.50	0.28	0.14	0.08	0.06
		recall	0.88	0.50	0.28	0.14	0.08	0.06
fully missing, null	20% missing	precision	0.77	0.16	0.05	0.01	0.01	0.01
		recall	0.77	0.16	0.05	0.01	0.01	0.01
	50% missing	precision	0.77	0.16	0.05	0.01	0.01	0.01
		recall	0.77	0.16	0.05	0.01	0.01	0.01

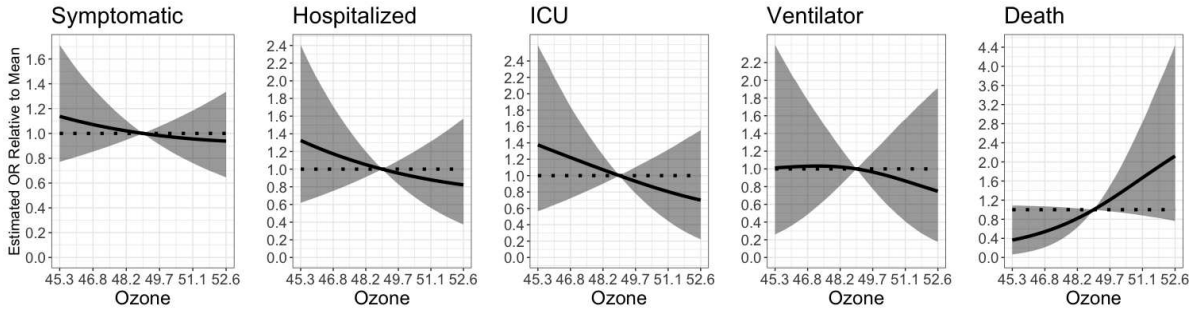
Table C.5: Summary of imputation performance in the equal probabilities setting. Results are shown for 20% and 50% missing data and four simulation scenarios (“partially missing, signal," “fully missing, signal," “partially missing, null," and “fully missing, null"). The table shows mean across 500 data sets for precision and recall for each outcome category.

			outcome category					
			1	2	3	4	5	6
partially missing, signal	20% missing	precision	0.71	0.67	0.63	0.63	0.59	0.62
		recall	0.71	0.68	0.64	0.63	0.59	0.62
	50% missing	precision	0.71	0.67	0.63	0.63	0.59	0.62
		recall	0.71	0.68	0.63	0.63	0.59	0.62
fully missing, signal	20% missing	precision	0.56	0.51	0.45	0.45	0.39	0.43
		recall	0.56	0.51	0.45	0.45	0.39	0.42
	50% missing	precision	0.56	0.51	0.45	0.45	0.39	0.42
		recall	0.56	0.51	0.45	0.44	0.39	0.42
partially missing, null	20% missing	precision	0.28	0.31	0.35	0.36	0.27	0.35
		recall	0.28	0.31	0.35	0.36	0.27	0.35
	50% missing	precision	0.28	0.31	0.35	0.36	0.27	0.35
		recall	0.29	0.31	0.35	0.36	0.27	0.34
fully missing, null	20% missing	precision	0.14	0.16	0.19	0.19	0.14	0.18
		recall	0.14	0.16	0.19	0.19	0.14	0.18
	50% missing	precision	0.14	0.16	0.19	0.19	0.14	0.18
		recall	0.14	0.16	0.19	0.19	0.14	0.18

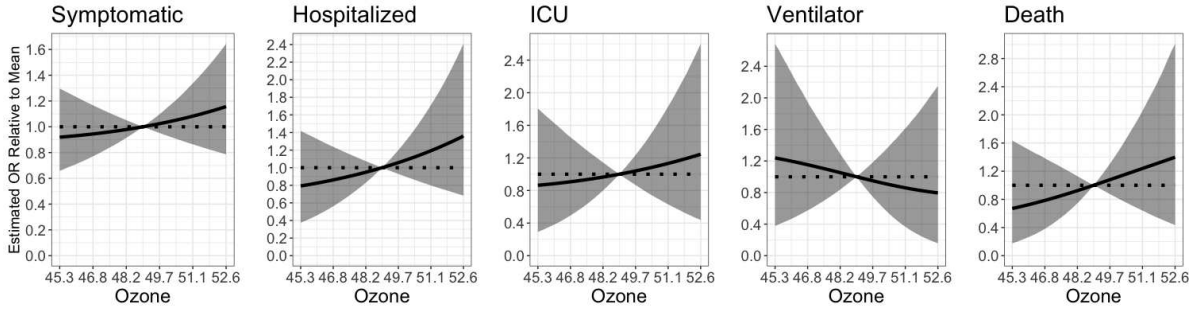
C.3 Additional Data Analysis Results

Table C.6: Results from the analysis of the Denver, CO COVID-19 cohort using our proposed method. The table shows the posterior mean, 0.025 quantile (lwr), and 0.975 quantile (upr) for the estimated exponentiated regression coefficients of the main effects of $PM_{2.5}$, ozone, and temperature and all pairwise interactions for each category. Estimated exponentiated regression coefficients with 95% posterior credible intervals that do not cross 1.00 are denoted in bold.

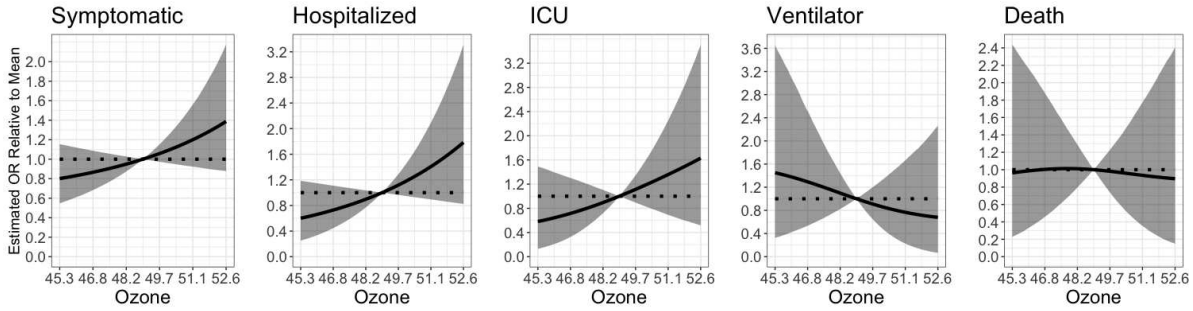
		mean	lwr	upr
$PM_{2.5}$	Symptomatic	**0.94**	**0.90**	**0.98**
	Asymptomatic	**0.85**	**0.77**	**0.94**
	Hospitalized	**1.24**	**1.08**	**1.43**
	ICU	0.93	0.73	1.20
	Ventilator	1.14	0.87	1.53
Ozone	Symptomatic	1.02	0.89	1.17
	Asymptomatic	0.96	0.73	1.27
	Hospitalized	1.03	0.72	1.45
	ICU	1.02	0.54	1.92
	Ventilator	0.66	0.33	1.31
Temperature	Symptomatic	0.94	0.86	1.04
	Asymptomatic	1.19	0.96	1.47
	Hospitalized	0.98	0.74	1.30
	ICU	1.07	0.67	1.72
	Ventilator	0.73	0.43	1.23
$PM_{2.5}$*Ozone	Symptomatic	0.97	0.87	1.08
	Asymptomatic	**0.69**	**0.55**	**0.87**
	Hospitalized	**1.45**	**1.06**	**1.97**
	ICU	**1.83**	**1.01**	**3.33**
	Ventilator	1.73	0.92	3.37
$PM_{2.5}$*Temperature	Symptomatic	1.03	0.98	1.09
	Asymptomatic	0.95	0.85	1.06
	Hospitalized	0.96	0.83	1.10
	ICU	**1.48**	**1.12**	**2.00**
	Ventilator	1.27	0.92	1.74
Ozone*Temperature	Symptomatic	**1.08**	**1.01**	**1.15**
	Asymptomatic	1.11	0.97	1.28
	Hospitalized	0.92	0.77	1.10
	ICU	1.15	0.84	1.59
	Ventilator	0.83	0.57	1.21



(a) 25th percentiles of $PM_{2.5}$ and temperature

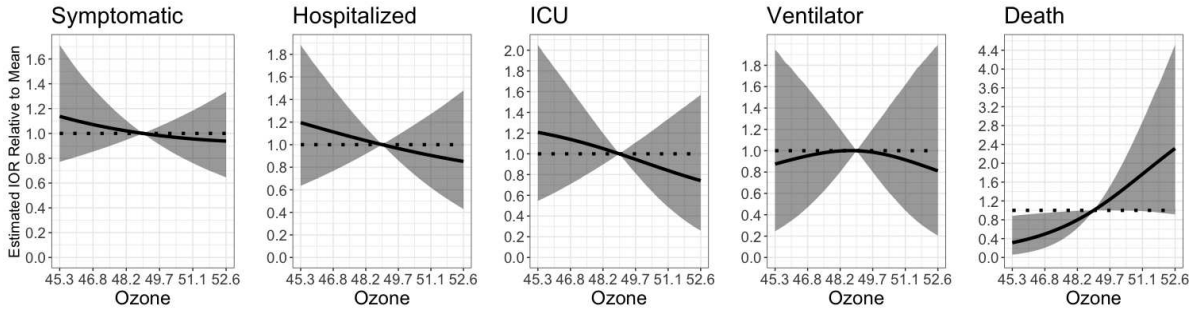


(b) 50th percentiles of $PM_{2.5}$ and temperature

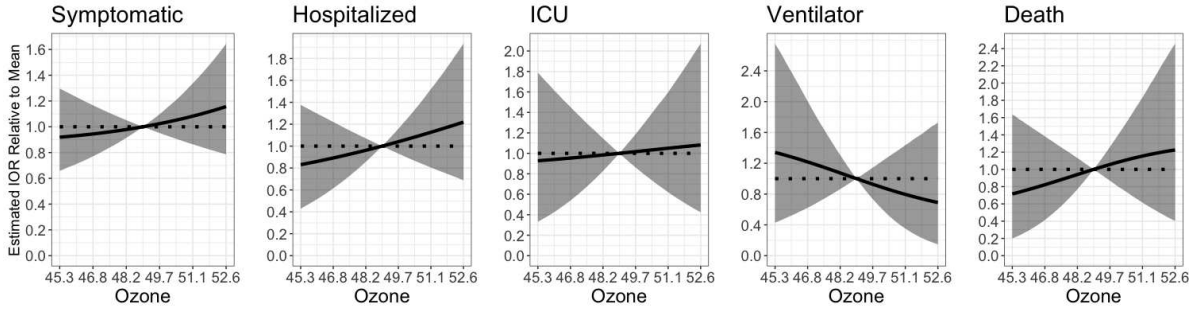


(c) 75th percentiles of $PM_{2.5}$ and temperature

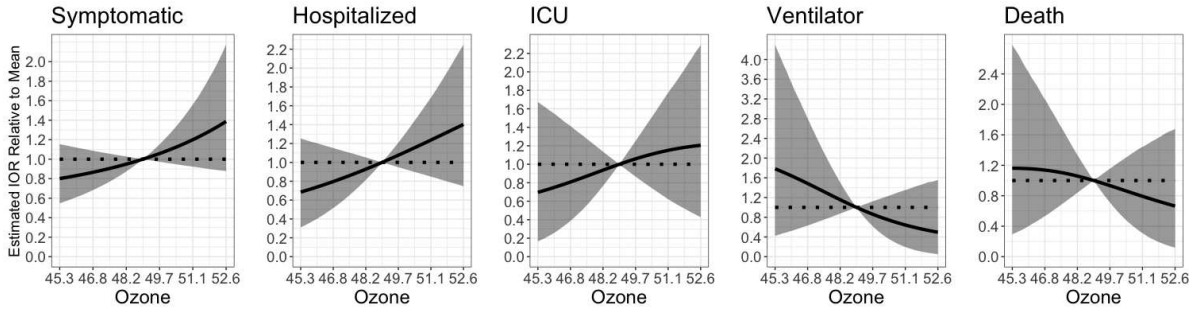
Figure C.1: Results from the analysis of the Denver, CO COVID-19 cohort using our proposed method. The figure shows the posterior mean (black line) and 95% credible interval (gray shaded area) of the estimated odds ratio (OR) for categories symptomatic, hospitalized, admitted to the ICU (ICU), placed on a mechanical ventilator (ventilator) and death, relative to asymptomatic. The OR was calculated as a function of annual average ozone exposure (ppb) relative to the mean exposure, holding $PM_{2.5}$ and temperature at their 25^{th} (a), 50^{th} (b), and 75^{th} (c) percentiles.



(a) 25th percentiles of $PM_{2.5}$ and temperature

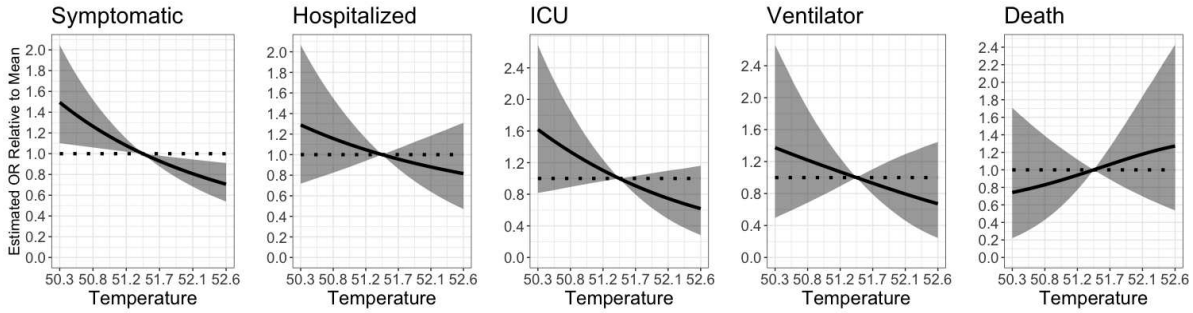


(b) 50th percentiles of $PM_{2.5}$ and temperature

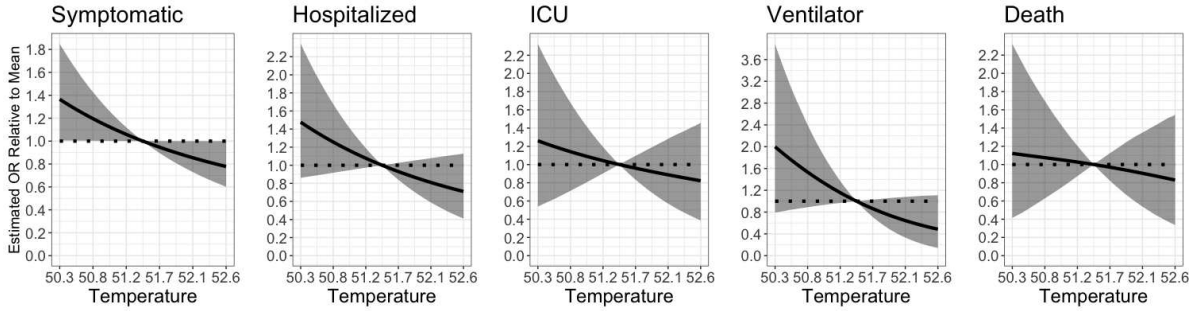


(c) 75th percentiles of $PM_{2.5}$ and temperature

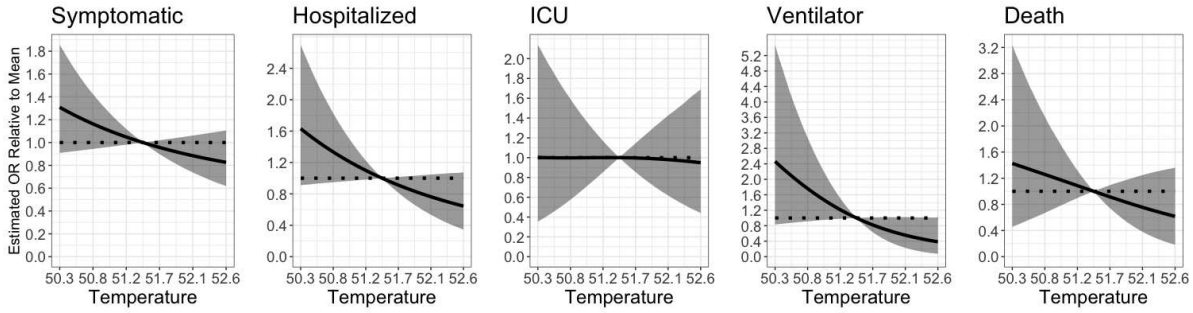
Figure C.2: Results from the analysis of the Denver, CO COVID-19 cohort using our proposed method. The figure shows the posterior mean (black line) and 95% credible interval (gray shaded area) of the estimated incremental odds ratio (IOR) for categories symptomatic, hospitalized, admitted to the ICU (ICU), placed on a mechanical ventilator (ventilator) and death, relative to all less severe categories. The IOR was calculated as a function of annual average ozone exposure (ppb) relative to the mean exposure, holding $PM_{2.5}$ and temperature at their 25^{th} (a), 50^{th} (b), and 75^{th} (c) percentiles.



(a) 25th percentiles of $PM_{2.5}$ and ozone

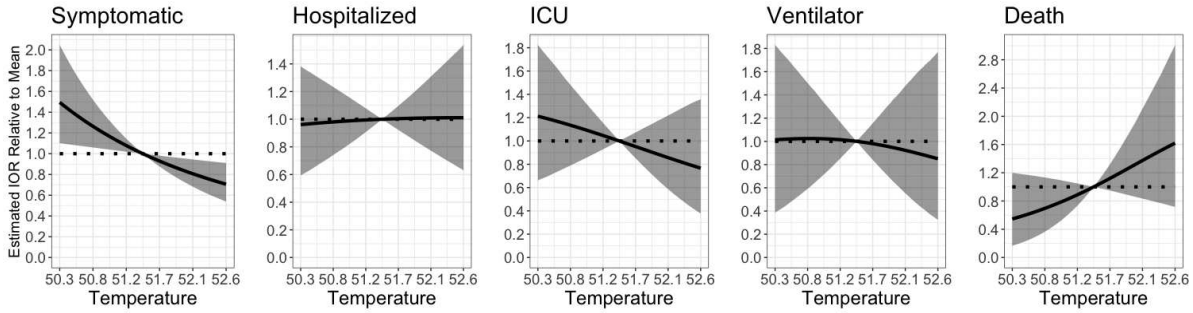


(b) 50th percentiles of $PM_{2.5}$ and ozone

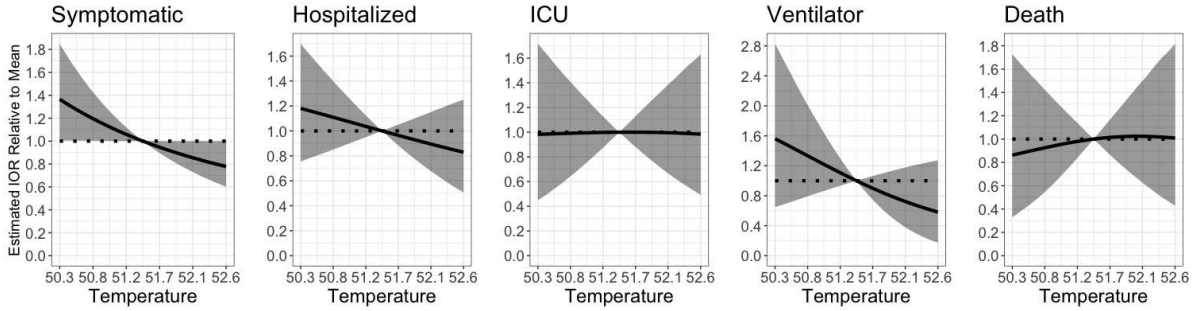


(c) 75th percentiles of $PM_{2.5}$ and ozone

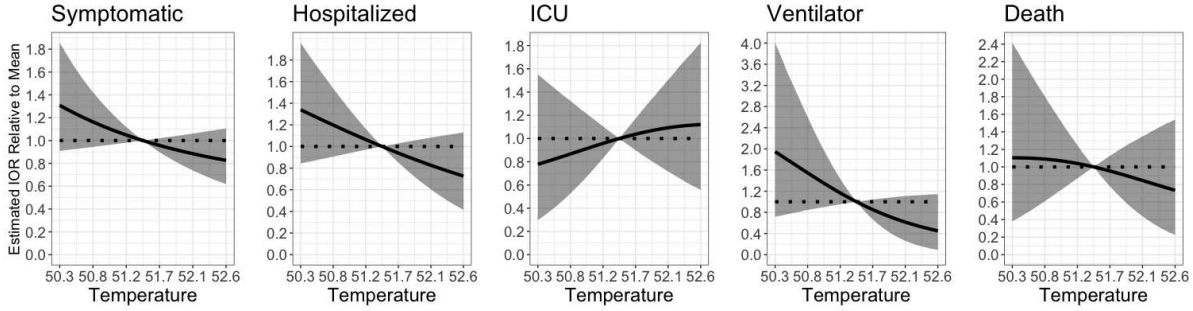
Figure C.3: Results from the analysis of the Denver, CO COVID-19 cohort using our proposed method. The figure shows the posterior mean (black line) and 95% credible interval (gray shaded area) of the estimated odds ratio (OR) for categories symptomatic, hospitalized, admitted to the ICU (ICU), placed on a mechanical ventilator (ventilator) and death, relative to asymptomatic. The OR was calculated as a function of annual average temperature (degrees Fahrenheit) relative to the mean exposure, holding $PM_{2.5}$ and ozone at their 25^{th} (a), 50^{th} (b), and 75^{th} (c) percentiles.



(a) 25th percentiles of $PM_{2.5}$ and ozone



(b) 50th percentiles of $PM_{2.5}$ and ozone



(c) 75th percentiles of $PM_{2.5}$ and ozone

Figure C.4: Results from the analysis of the Denver, CO COVID-19 cohort using our proposed method. Figure shows the posterior mean (black line) and 95% credible interval (gray shaded area) of the estimated incremental odds ratio (IOR) for categories symptomatic, hospitalized, admitted to the ICU (ICU), placed on a mechanical ventilator (ventilator) and death, relative to all less severe categories. The IOR was calculated as a function of annual average temperature (degrees Fahrenheit) relative to the mean exposure, holding $PM_{2.5}$ and ozone at their 25^{th} (a), 50^{th} (b), and 75^{th} (c) percentiles.

Table C.7: Results from the complete case analysis of the Denver, CO COVID-19 cohort. The table shows the posterior mean, 0.025 quantile (lwr), and 0.975 quantile (upr) for the estimated exponentiated regression coefficients of the main effects of $PM_{2.5}$, ozone, and temperature and all pairwise interactions for each category. Estimated exponentiated regression coefficients with 95% posterior credible intervals that do not cross 1.00 are denoted in bold.

		mean	lwr	upr
$PM_{2.5}$	Symptomatic	**0.94**	**0.90**	**0.99**
	Asymptomatic	**0.88**	**0.79**	**0.98**
	Hospitalized	**1.27**	**1.10**	**1.46**
	ICU	0.95	0.74	1.22
	Ventilator	1.16	0.87	1.54
Ozone	Symptomatic	1.04	0.90	1.21
	Asymptomatic	0.94	0.69	1.28
	Hospitalized	1.08	0.74	1.57
	ICU	1.04	0.54	1.98
	Ventilator	0.75	0.38	1.51
Temperature	Symptomatic	0.98	0.87	1.09
	Asymptomatic	1.20	0.95	1.53
	Hospitalized	1.04	0.77	1.40
	ICU	1.11	0.68	1.80
	Ventilator	0.83	0.48	1.43
$PM_{2.5}$*Ozone	Symptomatic	0.99	0.88	1.12
	Asymptomatic	0.74	0.57	0.96
	Hospitalized	**1.48**	**1.08**	**2.03**
	ICU	1.83	0.99	3.43
	Ventilator	1.67	0.86	3.16
$PM_{2.5}$*Temperature	Symptomatic	1.05	0.99	1.11
	Asymptomatic	0.96	0.84	1.09
	Hospitalized	0.96	0.83	1.12
	ICU	**1.46**	**1.09**	**2.02**
	Ventilator	1.23	0.89	1.72
Ozone*Temperature	Symptomatic	**1.11**	**1.04**	**1.19**
	Asymptomatic	1.12	0.96	1.30
	Hospitalized	0.95	0.78	1.15
	ICU	1.17	0.85	1.66
	Ventilator	0.89	0.61	1.31

C.4 Sensitivity Analysis Results

Table C.8: Results from the sensitivity analysis of complete cases in the Denver, CO COVID-19 cohort using the logistic regression model. COVID-19 peak severity outcomes were collapsed to two categories: severe (hospitalized, admitted to the ICU, placed on a mechanical ventilator, or died) and not severe (asymptomatic or symptomatic). The table shows the posterior mean, 0.025 quantile (lwr), and 0.975 quantile (upr) for the estimated exponentiated regression coefficients of the main effects of $PM_{2.5}$, ozone, and temperature and all pairwise interactions for each category. Estimated exponentiated regression coefficients with 95% posterior credible intervals that do not cross 1.00 are denoted in bold.

	mean	lwr	upr
$PM_{2.5}$	**1.09**	**1.02**	**1.18**
Ozone	0.96	0.77	1.18
Temperature	0.86	0.73	1.02
$PM_{2.5}$*Ozone	1.18	1.00	1.40
$PM_{2.5}$*Temperature	1.00	0.92	1.09
Ozone*Temperature	**0.86**	**0.77**	**0.94**