Bayesian Modeling of Latent Heterogeneity in Complex Survey Data and Electronic Health Records

Rebecca Anthopolos

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Public Health

in the Department of Biostatistics at the Mailman School of Public Health

COLUMBIA UNIVERSITY

2019

ABSTRACT

Bayesian Modeling of Latent Heterogeneity in Complex Survey Data and Electronic Health Records

Rebecca Anthopolos

In population health, the study of unobserved, or latent, heterogeneity in longitudinal data may help inform public health interventions. Growth mixture modeling is a flexible tool for modeling latent heterogeneity in longitudinal data. However, the application of growth mixture models to certain data types, namely, complex survey data and electronic health records, is underdeveloped. For valid statistical inferences in complex survey data, features of the sample design must be incorporated into statistical analysis. In electronic health records, the application of growth mixture modeling is challenged by high levels of missing values. In this dissertation, I have three goals: First, I propose a Bayesian growth mixture model for complex survey data in which I directly incorporate features of the complex sample design. Second, I extend a Bayesian growth mixture model of multiple longitudinal health outcomes collected in electronic health records to a shared parameter model that can account for different missing data assumptions. Third, I develop open-source software packages in `R` for each method that can be used for model fitting, selection, and checking.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am greatly indebted to my advisors, Qixuan Chen and Ying Wei, for their scientific expertise and professional mentorship. I would like to thank them for their steadfast commitment to the science, and for their thoughtfulness, warmth, and patience. Over these past years, they have truly become role models for me both professionally and personally. Thank you for such a rewarding collaboration.

I would like to thank the members of my dissertation committee other than my advisors, including Gen Li, Melanie M. Wall (chair), and Chunhua Weng.

I would like to express my gratitude to Marie Lynn Miranda for inspiring me with her commitment to scientific research many years ago. Her steadfast confidence in me and her continual career support have been incredible.

I am forever indebted to my husband, Arturas Rozenas, for his love, patience, and support, and my 3 year-old daughter, Zeva Rozenas, for teaching me about life.

Lastly, I would like to thank my parents, Savas and Elene Anthopolos, for always being there for me.

Chapter 1

Introduction

Latent heterogeneity in population health may be a consequence of unobservable subgroups of individuals with distinctive patterning in their longitudinal health trajectories. Subgroup membership may be associated with observed risk factors, such as health or demographic variables. Diverse research areas have sought to improve understanding of latent heterogeneity in longitudinal data. For example, [Elliott *et al.*, 2005] related baseline depression status with longitudinal measurements of mood scores and reactivity to negative events to identify unobserved subgroups of patients with varying risk of depressive disorder. [Neelon *et al.*, 2011] identified unobserved subgroups of pregnant women with distinctive blood pressure trajectories and varying risk of adverse birth outcomes. From a public health perspective, the study of heterogeneity can point to underlying causes of population health and suggest pathways towards improving health outcomes [Galea, 2017]. Trajectory patterns in different subgroups and associated risk factors can be used to target clinical monitoring towards individuals at-risk of adverse health outcomes, and to tailor interventions for specific risk profiles.

Statistical methods for modeling latent heterogeneity in longitudinal data are based on

relaxing the assumption of a single, homogeneous population that underlies conventional growth models. Two commonly used methods are latent class growth analysis (LCGA) and growth mixture models (GMMs) [Muthen *et al.*, 2002; Jung and Wickrama, 2008], which allow probabilistically classifying individuals into different unobserved subgroups – often called "latent classes" – based on individual longitudinal trajectories and risk factors associated with latent class membership. As finite mixture models, both LCGA and GMMs require pre-specifying a fixed number of latent classes K. Then, modeling proceeds in two parts: First, a discrete latent variable for an individual's class membership is introduced via data augmentation. By assuming latent class membership follows a multinomial distribution, the probabilities of latent class membership π_{ik} for individual i in latent class k $(k = 1, \ldots, K)$ can be modeled as a function of hypothesized risk factors. Second, given latent class membership, conditional densities of the longitudinal outcomes $f(\mathbf{y}_i \,|\, \theta_k)$ are specified, enabling estimation of the average longitudinal trajectory in each latent class. The mixture distribution is formed as $f(\mathbf{y}_i \,|\, \theta) = \sum_{k=1}^{K} \pi_{ik} f(\mathbf{y}_i \,|\, \theta_k)$, where the latent class membership probabilities π_{ik} act as mixing weights over the class-specific conditional densities.

The difference between LCGA and GMM concerns the specification of the variance-covariance of longitudinal measurements belonging to the same individual $\mathbf{y}_i$. In LCGA, the covariance between any pair of measurements from the same individual is fixed to zero. Conditional on latent class membership, an individual's longitudinal measurements are assumed to be independent. In contrast, in GMMs, conditional on latent class membership, between-subject heterogeneity is modeled using subject-specific random effects, such as random intercepts or random slopes, as in conventional growth models. Conditional on latent class membership, and subject-specific random effects, longitudinal measurements

from the same individual are assumed to be independent. In this way, GMMs can be viewed as mixtures of random effect models. In this dissertation, I use GMMs because of their general utility in modeling latent heterogeneity in longitudinal data compared to LCGA.

Methods for applying GMMs in certain data types, namely, complex survey data and electronic health records, remain underdeveloped. The application of GMMs to complex survey data is not straightforward because finite mixture modeling assumes that the sample was drawn using simple random sampling. Complex survey data, however, arise from complex sample designs in which different forms of controlled selection, such as unequal selection probabilities, stratification, and clustering, are used to construct a sample. As a probability sample, the selection probabilities of all elements in the population are known. For valid statistical inferences with complex survey data, sample design features must be incorporated into statistical analyses.

In electronic health records (EHRs), the application of GMMs is challenged by the often high prevalence of missing values, which is in part a consequence of the fact that the data were originally collected for clinical and administrative use rather than scientific research. Statistical inferences rely on assumptions about the probability distribution for whether a data point is observed. In the missing data lexicon of Rubin [Rubin, 1976], three missing data mechanisms are possible. Missing completely at random (MCAR) is when the probability of an observed response is unrelated to the value of the data point or to the value of any other observed or unobserved variable. Missing at random (MAR) is when conditional on observed variables, the probability of an observed response is independent of the missing data point or unobserved variables. Under missing not at random (MNAR), the probability of an observed response depends on the missing data point or unobserved

variables, even after conditioning on observed variables. The application of GMMs to EHRs therefore requires assumptions about how the patterns of missing values arose.

In this dissertation, my three aims are to:

1. Propose a new method for applying GMMs to complex survey data while accounting for features of the complex sample design;

2. Propose a new method for applying GMMs to EHRs that can account for different assumptions about the missing data; and,

3. Develop open-source software in the form of `R` packages for each of the proposed new methods.

This dissertation is organized as follows: In Chapter 2, I propose a Bayesian GMM for complex survey data. In Chapter 3, I propose a Bayesian method for applying GMMs to EHRs that can account for different missing data assumptions. In Chapter 4, I explicate two `R` software packages that I developed for the methods in Chapters 2 and 3, which can be used for model fitting, selection, and checking. Finally, in Chapter 5, I conclude with a summary of the contributions of this dissertation.

Chapter 2

A Bayesian Growth Mixture Model for Complex Survey Data: Clustering Post-Disaster PTSD Trajectories

2.1 Introduction

In disaster recovery research, disasters are defined as acute events, such as hurricanes or industrial accidents, that affect many people simultaneously, occur suddenly, and result in at least some primary victims [Norris *et al.*, 2002]. A commonly studied condition of mental health among disaster survivors is post-traumatic stress disorder (PTSD), which manifests through multiple, persistent symptoms like flashbacks and negative thinking [National Institute of Mental Health, 2016]. After a disaster, PTSD trajectories over time exhibit well-documented heterogeneity. The modal trajectory subgroup has been shown to be resilience, which entails early transient perturbations along a relatively stable path of healthy functioning [Bonanno and Diminich, 2013; Norris *et al.*, 2009]. Among other trajectory

subgroups, recovery entails an extended period of dysfunction followed by a gradual return to pre-event functioning, and chronic dysfunction is manifested when an initial stress reaction persists indefinitely [Bonanno and Diminich, 2013; Norris *et al.*, 2009]. The different subgroups of PTSD trajectories are an indication of unobserved, or latent, heterogeneity in the population. For such data from post-disaster studies, a growth mixture model (GMM) can be used to combine subject-level risk factors with longitudinal trajectories to classify subjects probabilistically into different trajectory subgroups, often called "latent classes".

The current study is motivated by the Galveston Bay Recovery Study (GBRS), conducted by the National Center for Disaster Mental Health Research. The GBRS used a stratified multi-stage sample design to collect longitudinal data on PTSD among survivors of Hurricane Ike that struck the Galveston Bay Area of Texas on September 13-14, 2008 [Valliant *et al.*, 2009; Rice, 2016]. To characterize heterogeneity in longitudinal trajectories of PTSD in this population and describe risk factors associated with each trajectory subgroup, I use a GMM to identify latent trajectories and estimate associated risk factors while incorporating the complex sample design.

For valid statistical inferences with complex survey data, sample design features must be incorporated into statistical analyses. Existing methods for finite mixture modeling with complex survey data use pseudo-likelihood with variance estimated via linearization or resampling techniques [Wedel *et al.*, 1998; Patterson *et al.*, 2002; Asparouhov, 2005]. However, large sample approximations are necessary for analyses based on pseudo-likelihoods, and estimation can be challenging for complex models. Alternatively, a Bayesian framework is advantageous because it allows building flexible and complex models and can handle small samples and missing data [Little, 2003; Little, 2004]. In this paper, I propose a Bayesian GMM for complex survey data. I model the hierarchical structure of the data, with repeated

measures of PTSD in different waves of the survey nested within subjects, which are further clustered by area segments and geographic strata. Features of the complex sample design, such as stratification, clustering, and unequal probability sampling, are directly included in the model as covariates or hierarchical variance components. Because classification of disaster survivors into different latent classes may exhibit geographic clustering [Gruebner *et al.*, 2016], I account for spatial correlations among neighboring clustering units in the model for latent class membership. In contrast to existing methods [Muthen and Muthen, 2017], I model longitudinal trajectories as a function of discrete time. My model allows partitioning variability in the probability of latent class membership and PTSD between different aspects of the sample design and other sources. To ease computation, I model latent class membership risk using a multinomial probit model. I show model selection and model checking. For posterior computation, I propose an efficient Markov chain Monte Carlo (MCMC) algorithm. I implement the proposed Bayesian GMM for complex survey data in the `Bsvygmm` package in `R`.

2.2 Motivating Data

The GBRS was a three-wave panel survey conducted in the aftermath of Hurricane Ike. The study aimed to characterize the trajectories and determinants of post-disaster mental health outcomes [Valliant *et al.*, 2009]. The target population comprised persons aged 18 years or older living in Galveston and Chambers counties, Texas, who were present when Hurricane Ike struck, and who had been living in the study area for at least the preceding month [Valliant *et al.*, 2009]. The study area was divided into five geographic strata based on the degree of flood damage and level of poverty from the 2000 US Census (Figure 2.1). Differential sampling rates were used to oversample from strata expected to be worse off from

the storm. Constructed from census block boundaries, 77 area segments were sampled with probability proportional to size (pps) sampling within strata using the number of occupied households from the 2000 US Census as the size variable. Household socioeconomic data were obtained to construct a high risk indicator for developing PTSD. Households at high risk were oversampled.

In the current study, I consider PTSD severity score, which is equal to the sum of responses to 17 symptoms of PTSD, such as "repeated, disturbing memories of Hurricane Ike". Participants rated each symptom on a scale from 1 to 5 corresponding to increasing severity. At wave 1, scores measure PTSD severity since Hurricane Ike. Scores at waves 2 and 3 refer to the time period since the previous interview. At wave 1, various baseline risk factors hypothesized to be associated with mental health wellness were also collected.

2.3 Methods

I formulate the Bayesian GMM for modeling PTSD severity scores among participants in the GBRS that accounts for complex sample design. Assume that there are K latent classes of subjects with distinctive PTSD trajectory patterns across the three survey waves. I first present the latent class membership model in Section 2.3.1, followed by the longitudinal model of PTSD severity scores in Section 2.3.2. In Sections 2.3.3 and 2.3.4, I specify the prior distributions and show posterior computation. In Sections 2.3.5 and 2.3.6, I describe model selection and model checking.

2.3.1 Latent class membership model

In finite mixture modeling, I can define the mixture density for subject i over K latent classes as $f(y_i \mid \Theta) = \sum_{k=1}^{K} \pi_{ik} f(y_i \mid \theta_k)$, with $\sum_{k=1}^{K} \pi_{ik} = 1$. For $k = 1, \ldots, K$, $f(y_i \mid \theta_k)$ are

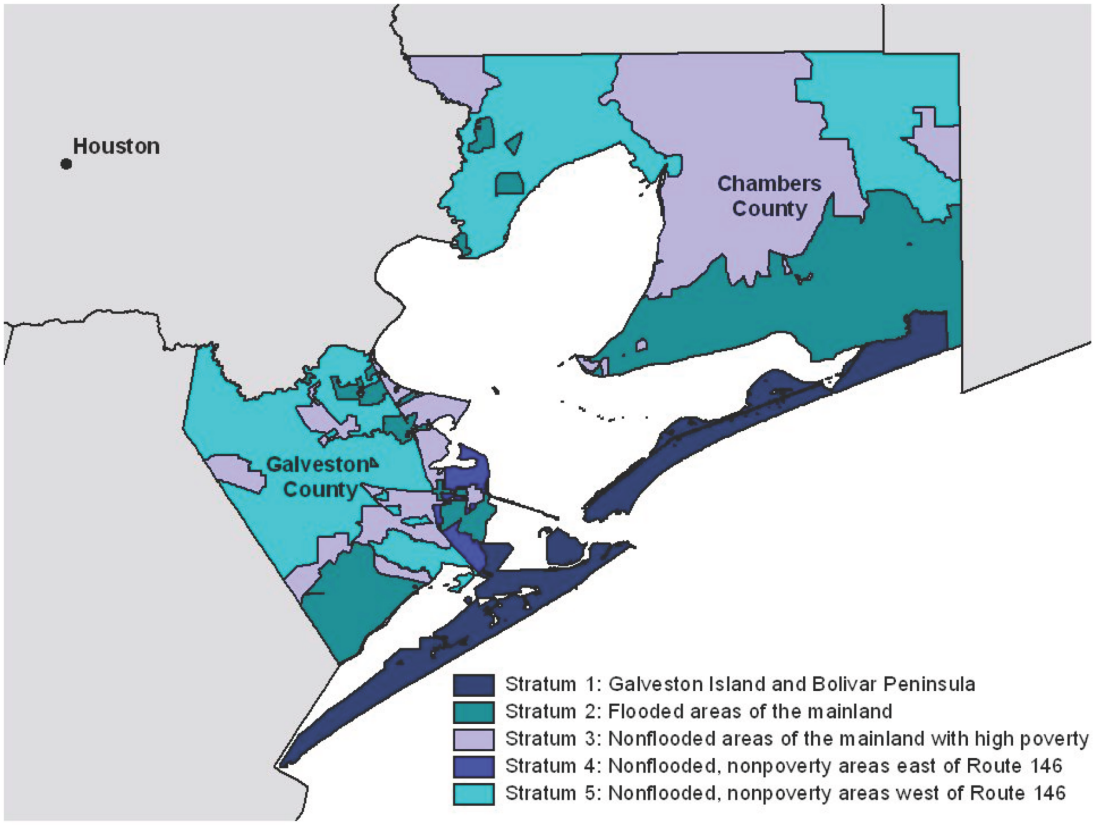

Figure 2.1: Geographic strata in the Galveston Bay Recovery Study. Strata were labeled 1 to 5 in order of decreasing degree of flood damage. Stratum 1 represented Galveston Island and the Bolivar Peninsula, which suffered storm surge damage. Stratum 2 represented flooded areas on the mainland. Stratum 3 indicated non-flooded regions with high poverty, while strata 4 and 5 indicated different non-flooded regions with low poverty.

component densities, and π_{ik} are subject-specific class weights, also called mixing weights, used to "mix" the component densities. As a probability derived from a multinomial regression of latent class membership, π_{ik} can be modeled as a function of different sources of information that may predict a subject's latent class. By adapting an estimation framework for discrete choice models in Bayesian econometrics [McCulloch and Rossi, 1994], I use multinomial probit regression to model π_{ik}.

For participants in the GBRS, let s denote sampling strata ($s = 1, \ldots, S$), j denote area segments ($j = 1, \ldots, J_s$), and i denote subjects ($i = 1, \ldots, n_{sj}$), where n_{sj} is the number of subjects in area segment j of stratum s. To specify the multinomial probit regression, I define c_{sji} to be a discrete variable for latent class membership with values $1, \ldots, K$. Let $\mathbf{z}_{sji} = (z_{sji1}, \ldots, z_{sjiK})^T$ be a column vector of K continuous latent variables associated with c_{sji} such that

$$\mathbf{z}_{sji} = \mu_{sji} + \epsilon_{sji} \quad \text{and} \quad c_{sji} = k \text{ if } max(\mathbf{z}_{sji}) = z_{sjik}, \tag{2.1}$$

where the probability of belonging to latent class k is given by $\pi_{sjik} = Pr(z_{sjik} > z_{sjil} \text{ for all } l \neq k)$ [Albert and Chib, 1993; McCulloch and Rossi, 1994]. μ_{sji} is a mean vector of length K, and ϵ_{sji} is a K-length vector of random errors with $\epsilon_{sji} \sim N_K(0, \mathbf{H})$, where $\mathbf{H}$ is a $K \times K$ variance-covariance matrix.

The model in equation (2.1), however, is unidentifiable without restrictions [Daganzo, 1979; Dansie, 1985; Bunch, 1991]. Following [McCulloch and Rossi, 1994], I use latent class K as the reference class and construct column vector $\mathbf{z}^*_{sji} = (z_{sji1} - z_{sjiK}, \ldots, z_{sji(K-1)} - z_{sjiK})^T$ with

$$\mathbf{z}^*_{sji} \quad = \quad \mu_{\mathbf{sji}}{}^* + \epsilon^*_{sji}. \tag{2.2}$$

In equation (2.2), $\mu_{\mathbf{sji}}{}^* = (\mu^*_{sji1}, \ldots, \mu^*_{sji(K-1)})^T$ and $\epsilon^*_{sji} \sim N_{K-1}(\mathbf{0}, \mathbf{H}^*)$, with $\mathbf{H}^*$ being a

$(K-1) \times (K-1)$ variance-covariance matrix. I then define c^*_{sji} as

$$c^*_{sji} = \begin{cases} K & \text{if } max(\mathbf{z}^*_{sji}) < 0 \\ k & \text{if } max(\mathbf{z}^*_{sji}) = z^*_{sjik} \geq 0 \text{ for } k = 1, \ldots, K-1. \end{cases} \tag{2.3}$$

To address the identifiability problem, previous research has recommended drawing inference only on identifiable subsets of parameters or implementing constraints on the mean structure μ^*_{sji} or the variance-covariance $\mathbf{H}^*$ [Imai and van Dyk, 2005; Koop, 2003; McCulloch and Rossi, 1994]. I choose to implement a constraint on $\mathbf{H}^*$ with $\mathbf{H}^*$ being an identity matrix. For $K = 2$, this is the standard Bayesian probit model for a binary outcome [Albert and Chib, 1993].

I model the mean structure μ^*_{sjik} $(k = 1, \ldots, K-1)$ as

$$\mu^*_{sjik} = \lambda_{sk} + u_{sjk} + \nu_{sjk} + g_k(x_{sj}) + \mathbf{w}^T_{sji}\delta_k \tag{2.4}$$

$$\lambda_{sk} \overset{ind}{\sim} N(0, \gamma^2_k) \tag{2.5}$$

$$u_{sjk} \overset{ind}{\sim} N(0, \tau^2_k) \tag{2.6}$$

$$\nu_{sjk}|\nu_{-sjk} \overset{ind}{\sim} N(\bar{\nu}_{sjk}, \frac{\xi^2_k}{m_{sj}}), \tag{2.7}$$

where λ_{sk} is a stratum-specific intercept that reflects variability in the probability of latent class membership from different strata; u_{sjk} is an area segment-specific intercept that captures correlations among subjects who live in the same area segment; and ν_{sjk} is an area segment-specific intercept that accounts for spatial correlations among neighboring segments. ν_{sjk} is modeled according to an intrinsic conditional autoregressive (ICAR) prior distribution [Besag, 1974; Besag and Kooperberg, 1995], where ξ^2_k is a latent class-specific spatial variance component; m_{sj} is the number of neighbors of segment j in stratum s, with neighboring segments defined by a shared border or vertex; and $\bar{\nu}_{sjk}$ is the sample average of these m_{sj} neighboring segments.

To flexibly capture the effect of pps sampling on probability of latent class membership, I model μ^*_{sjik} as a smoothed function of x_{sj}, the number of occupied households in area segment j of stratum s, using B-splines of polynomial degree m [Chen *et al.*, 2010]. With L pre-specified inner knots and $R = m + L$ degrees of freedom, I have

$$g_k(x_{sj}) = \sum_{r=1}^{R} \alpha_{kr} B_r(x_{sj}), \tag{2.8}$$

where $B_r(x_{sj})$ denotes the r^{th} basis function evaluated at x_{sj}, and $\alpha_k = (\alpha_{k1}, \ldots \alpha_{kR})^T$ are latent class-specific regression coefficients.

Lastly, δ_k contains regression coefficients for corresponding covariates in $\mathbf{w}_{sji}$, including risk factors for PTSD, such as age, and sample design variables, such as the number of household members, that may be associated with latent class membership.

2.3.2 Longitudinal model of PTSD severity scores

Longitudinal PTSD severity scores are modeled conditional on latent class membership. Let t denote the interview wave for $t = 1, 2, 3$. Then, for the i^{th} subject at wave t in area segment j of stratum s and latent class $k = 1, \ldots, K$, I assume

$$\begin{aligned} &[y_{sjit} \mid \mathbf{b}_{sji}, \rho_{sjk}, \zeta_{sk}, c^*_{sji} = k] \\ &= \mathbf{1}_{t=1} b_{1sji} + \mathbf{1}_{t=2} b_{2sji} + \mathbf{1}_{t=3} b_{3sji} + \rho_{sjk} + \zeta_{sk} + \chi_{sjitk}, \end{aligned} \tag{2.9}$$

where

$$\left[\mathbf{b}_{sji} \mid c^*_{sji} = k\right] \overset{ind}{\sim} N_3(\beta_k, \mathbf{\Phi}_k) \tag{2.10}$$

$$\rho_{sjk} \overset{ind}{\sim} N(0, \omega_k^2) \tag{2.11}$$

$$\zeta_{sk} \overset{ind}{\sim} N(0, \psi_k^2) \tag{2.12}$$

$$\chi_{sjitk} \overset{ind}{\sim} N(0, \sigma_k^2). \tag{2.13}$$

In equations (2.9)-(2.13), y_{sjit} is the natural log-transformed PTSD severity score; $\mathbf{b}_{sji} = (b_{1sji},\ b_{2sji}, b_{3sji})^T$ is a column vector that captures the subject-specific latent trajectory centered around the class-specific average growth parameters β_k; and $\mathbf{\Phi}_k$ is a 3×3 unstructured variance-covariance matrix with elements $\phi_{ee'k}$ ($e = 1, 2, 3,\ e^{'} = 1, 2, 3$) that capture between-subject variation in trajectories in class k. Among subjects in area segment j of stratum s in latent class k, ρ_{sjk} is an area segment-specific intercept with a latent class-specific variance ω_k^2. ζ_{sk} is analogously defined at the stratum-level. Finally χ_{sjitk} is an observation-level error with $\chi_{sjitk} \overset{ind}{\sim} N(0, \sigma_k^2)$.

2.3.3 Prior distributions

Bayesian modeling requires specification of prior distributions for all parameters. For each parameter, I use the same prior distribution across mixture components. In the latent class membership model, I follow previous research [Garrett and Zeger, 2000; Elliott *et al.*, 2005] by assigning the probit regression coefficients (δ_k and α_k) independent proper prior distributions $N(0, \mathbf{I})$. After transforming the coefficients to the probability scale, this prior distribution yields a non-informative prior on the probability of latent class membership, with its mode at approximately $\frac{1}{K}$. I assign non-informative uniform prior distributions on the hierarchical standard deviations γ_k, τ_k, and ξ_k [Gelman, 2006].

In the longitudinal model of PTSD, I assign $\beta_k \sim N_3(\mathbf{0}, 10\mathbf{I})$, with $\sqrt{10}$ being over five times the interquartile range of log PTSD severity scores. I assign $\mathbf{\Phi}_k$ an inverse-Wishart prior distribution $IW(\nu_0, \mathbf{S}_0)$, with $\nu_0 = 5$ to indicate lack of knowledge about the latent class-specific variance-covariance and $\mathbf{S}_0$ fixed to a positive definite matrix. As in the latent class membership model, I use uniform prior distributions on the hierarchical standard deviations ψ_k and ω_k. I assign the observation-level variance σ_k^2 an inverse gamma prior

$IG(0.1, 0.1)$.

2.3.4 Posterior computation

Let $\mathbf{\Theta}$ and $\mathbf{\Upsilon}$ be containers for parameters in the longitudinal model for PTSD severity scores and the latent class membership model, respectively. Let $\mathbf{X}$ be a container for model covariates. For posterior computation, I consider the likelihood

$$\begin{aligned} & L(\mathbf{c}^*, \mathbf{\Theta}, \mathbf{\Upsilon} \mid \mathbf{y}; \mathbf{X}) \\ &= \prod_{s=1}^{S} \prod_{j=1}^{J_s} \prod_{i=1}^{n_{sj}} \prod_{k=1}^{K} \Big(\pi_{sjik} \, f(\mathbf{y}_{sji} \mid c^*_{sji}, \mathbf{b}_{sji}, \rho_{sjk}, \zeta_{sk}, \sigma^2_k; \mathbf{X}_{sji}) \\ & \quad \times f(\mathbf{b}_{sji} \mid c^*_{sji}, \beta_k, \mathbf{\Phi}_k) \, f(\rho_{sjk} \mid \omega^2_k) \, f(\zeta_{sk} \mid \psi^2_k) \Big)^{\mathbf{1}_{c^*_{sji}=k}}. \end{aligned}$$

My posterior computation uses Gibbs sampling with closed-form full conditional distributions. To improve the convergence properties of the MCMC sampler, I follow previous research [Frühwirth-Schnatter *et al.*, 2004; Fruhwirth-Schnatter, 2006] by proposing a partly marginalized Gibbs sampler. Using the method of collapsing, I replace the full conditional densities of selected parameters with their marginal densities obtained from integrating out part of the conditioning parameters. Specifically, after I set initial values for the model parameters, the algorithm iterates among the following three steps:

1. For $k = 1, \ldots, K-1$, update parameters in the latent class membership model (2.2) - (2.8), including z^*_{sjik}, λ_{sk}, u_{sjk}, ν_{sjk}, δ_k, α_k, γ^2_k, τ^2_k, and ξ^2_k. Calculate π_{sjik} for $k = 1, \ldots, K$.

2. For $k = 1, \ldots, K$, update parameters in the longitudinal outcomes model (2.9) - (2.13), including $\mathbf{b}_{sji}$, ρ_{sjk}, ζ_{sk}, β_k, σ^2_k, $\mathbf{\Phi}_k$, ω^2_k, and ψ^2_k. In the partly marginalized Gibbs sampler, the full conditional for β_k is replaced with the partially marginalized density obtained from integrating out $\mathbf{b}_{sji}$, ρ_{sjk}, and ζ_{sk}.

3. Using updated parameters from steps 1 and 2, draw the latent class indicators c^*_{sji} $(i = 1, \ldots, n)$ defined in (2.3) from $Multinomial(1; p_{sji1}, \ldots, p_{sjiK})$, with the posterior probabilities of latent class assignment p_{sjik} $(k = 1, \ldots, K)$ given by

$$\begin{aligned} p_{sjik} &= Pr(c^*_{sji} = k \mid \mathbf{y}_{sji}, \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k, \pi_{sjik}; \mathbf{X}_{sji}) \\ &= \frac{\pi_{sjik}\, f(\mathbf{y}_{sji} \mid \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k; \mathbf{X}_{sji})}{\sum_{k=1}^{K} \pi_{sjik}\, f(\mathbf{y}_{sji} \mid \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k; \mathbf{X}_{sji})}, \end{aligned} \tag{2.14}$$

with $f(\mathbf{y}_{sji} \mid .)$ being the partially marginalized density obtained by integrating out $\mathbf{b}_{sji}$, ρ_{sjk}, and ζ_{sk} from $f(\mathbf{y}_{sji}, \mathbf{b}_{sji}, \rho_{sjk}, \zeta_{sk} \mid \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k; \mathbf{X}_{sji})$.

The full MCMC algorithm is detailed in Appendix A.

2.3.5 Model selection

I conduct model selection according to model information criteria and graphical methods. I apply three information criteria: the Bayesian Information Criterion (BIC) [Schwarz, 1978], the integrated classification likelihood using a BIC approximation (ICL-BIC) [Biernacki *et al.*, 2000], and a modified version of the Deviance Information Criterion [Spiegelhalter *et al.*, 2002] for latent variable models known as the DIC4 [Celeux *et al.*, 2006]. Commonly used for model selection in mixture modeling, the BIC combines a measure of goodness of fit with a penalty for model complexity. The ICL-BIC extends the BIC to include a penalty for poorly separated components. Recommended by [Celeux *et al.*, 2006] as an information criterion in the latent variable setting, the DIC4 also penalizes both model complexity and poorly separated components. For each information criterion, models with smaller values are considered preferable. Details about these information criteria can be found in Appendix B.

I use graphical techniques [Garrett and Zeger, 2000] to confirm my selection based on the

information criteria. To examine the extent to which the data are able to distinguish among the assumed number of latent classes, I compare the prior versus posterior distributions for regression coefficients in the latent class membership model. Largely overlapping prior and posterior distributions may suggest that the number of latent classes is too large given the data.

2.3.6 Model checking

I evaluate the overall adequacy of the selected model using Bayesian posterior predictive p-values [Gelman *et al.*, 1996]. At each MCMC iteration, a discrepancy measure is computed using the replicated and observed data. The Bayesian predictive p-value denotes the probability that the discrepancy measure under the replicated data is greater than that under the observed data. A p-value near 0.5 indicates adequate model fit, while a p-value outside the range of 0.05 and 0.95 is considered to suggest a lack of model fit. For my discrepancy measure, I select a weighted mean squared error computed as [Neelon *et al.*, 2011; Gelman *et al.*, 2014]:

$$T = \sum_{k=1}^{K} \sum_{s=1}^{S} \sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} \sum_{t=1}^{3} \frac{(y_{sjit} - \mathbf{1}_{t=1} b_{1sji} - \mathbf{1}_{t=2} b_{2sji} - \mathbf{1}_{t=3} b_{3sji} - \rho_{sjk} - \zeta_{sk})^2}{\sigma_k^2} \times \mathbf{1}_{c^*_{sji}=k}.$$

In addition, I compare plots of the observed data with the posterior predictive distribution to check how well the model captures features of the data.

2.4 Results from the Analysis of the GBRS

I applied the proposed Bayesian GMM to modeling trajectories of PTSD severity scores across the three waves in the GBRS. I considered models with $K = 2, 3, 4$ latent classes. For each K, I fit three different latent class membership models. In the first two models, I

singly included u_{sjk} in equation (2.6) or ν_{sjk} in equation (2.7) to model correlations among subjects in the same area segment or spatial correlations among neighboring area segments, respectively. In the third model, I included both u_{sjk} and ν_{sjk} as in equation (2.4). In fitting each latent class membership model, $\mathbf{w}_{sji}$ from equation (2.4) contained design variables, including high versus low PTSD risk, natural log-transformed weighting adjustment for nonresponse at the household-level, and the number of household members; and PTSD risk factors previously identified in the literature, including demographics, community-level social assets including average collective efficacy and average social support, and pre and peri-disaster mental health factors [Lowe *et al.*, 2015; Gruebner *et al.*, 2016]. To model the effect of pps sampling of area segments with probability of selection proportional to the number of occupied households, I used a cubic B-spline with a set of basis functions for $R = 5$ in equation (2.8).

I ran the MCMC sampler for 30,000 iterations, discarding the first 15,000 as a burn-in. Based on three chains from dispersed initial values, the Gelman-Rubin convergence diagnostic [Gelman *et al.*, 2014] indicated model convergence with values near 1 for all parameters. Trace plots did not show evidence of the label switching problem that can occur in finite mixture modeling applications.

2.4.1 Number of latent classes in the GBRS

The 2-class and 3-class models converged, but the 4-class model could not identify a fourth mixture component in the GBRS data. The 3-class models have a smaller BIC, ICL-BIC, and DIC4 than the 2-class models (Table 2.1). Among the 3-class models, the DIC4 prefers the model with both types of correlations u_{sjk} and ν_{sjk} by a sizeable margin, while the BIC and ICL-BIC are similar for the different correlation structures. I select the 3-class model

Table 2.1: Comparison of information criteria between models with $K = 2, 3$ latent classes. For each number of latent classes, three models to account for different correlations among area segments in the latent class membership model, including u_{sjk} only, ν_{sjk} only, and both u_{sjk} and ν_{sjk}, are compared.

		K	
Criterion	Correlations	2	3
BIC	u_{sjk} only	-408.79	-563.96
	ν_{sjk} only	-430.50	-566.88
	u_{sjk} and ν_{sjk}	-456.71	-557.17
ICL - BIC	u_{sjk} only	-329.71	-401.22
	ν_{sjk} only	-346.34	-388.00
	u_{sjk} and ν_{sjk}	-384.41	-408.37
DIC4	u_{sjk} only	51.73	-558.01
	ν_{sjk} only	-186.81	-688.00
	u_{sjk} and ν_{sjk}	-140.52	-776.40

with both types of correlations.

Figures 2.2 and 2.3 compare the posterior versus prior distributions for the regression coefficients δ_k from the 3-class model with both types of correlations. The posterior distributions are narrow compared to the prior distributions, suggesting that the data contain evidence to estimate this 3-class model.

2.4.2 Latent classes of PTSD severity score trajectories

Figure 2.4 shows the mean trajectory in each latent class using the posterior means for the growth parameters β_k and corresponding 95% credible intervals in vertical bars. The growth parameters in latent class 1 (solid, black) portray a steadily low level of log PTSD severity scores, whereas in latent class 3 (solid, grey), a high level of PTSD persists over time. In

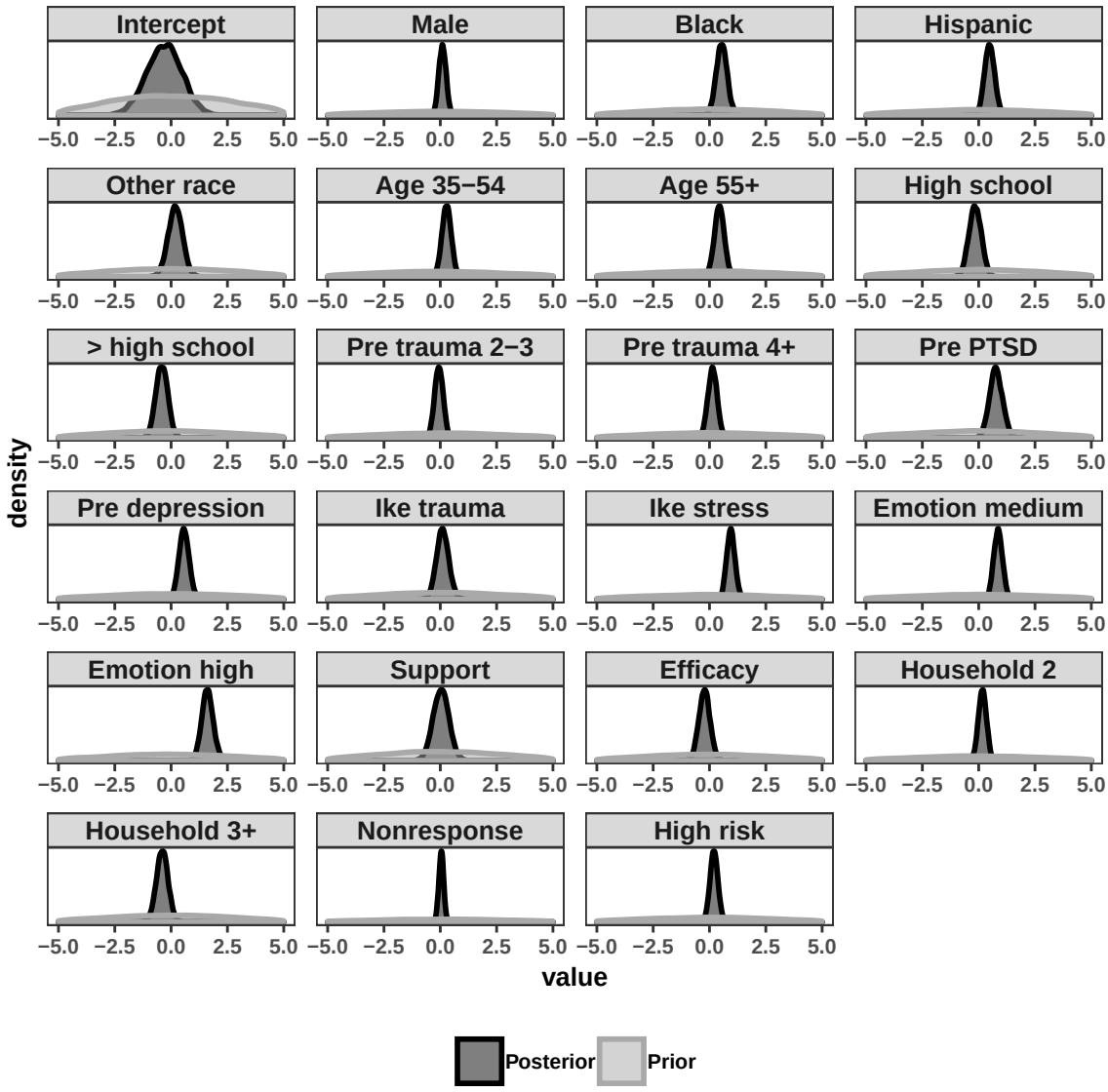

Figure 2.2: Posterior versus prior densities of regression coefficients δ_1 from the latent class membership model comparing the likelihood of being in the recovery versus resilient subgroup. The model includes both correlations among subjects in the same area segment and spatial correlations among area segments.

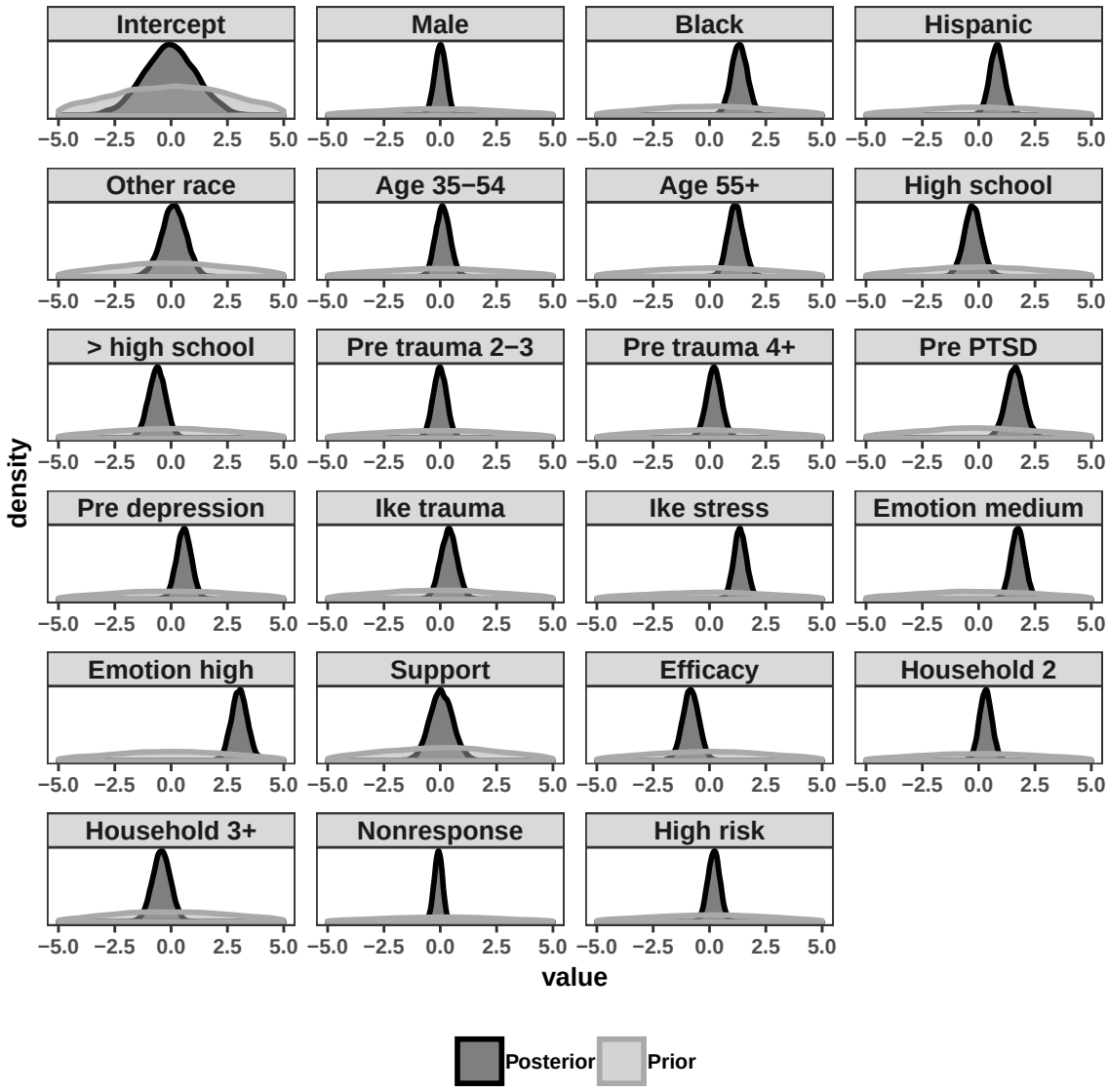

Figure 2.3: Posterior versus prior densities of regression coefficients δ_2 from the latent class membership model comparing the likelihood of being in the chronic versus resilient subgroup. The model includes both correlations among subjects in the same area segment and spatial correlations among area segments.

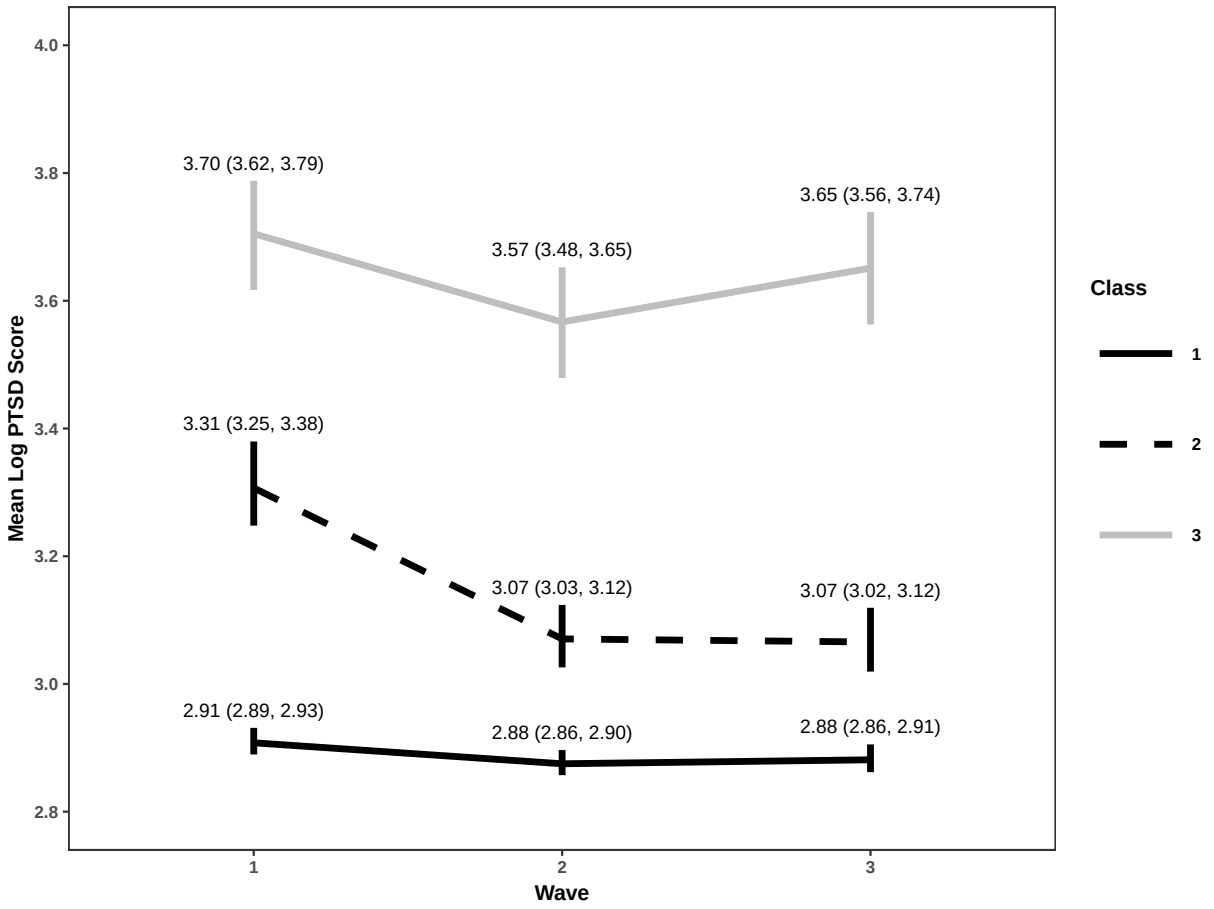

Figure 2.4: Mean log PTSD severity score trajectory in each latent class based on the posterior mean and 95% credible interval of β_k in the longitudinal model of PTSD.

latent class 2 (dashed), after decreasing from medium high in wave 1 to medium low in wave 2, the trend remains steady at wave 3. Figure 2.4 demonstrates that based on taxonomy used in disaster recovery research, latent classes 1, 2, and 3 can be interpreted as the *resilient*, *recovery*, and *chronic* subgroups of PTSD severity score trajectories, respectively.

Table 2.2 presents the hierarchical variance components at the observation-level, subject-level, area segment-level, and stratum-level in the model of longitudinal PTSD severity scores. At each level, the resilience subgroup is characterized by very small variability, and

the chronic subgroup exhibits larger variability relative to the other subgroups. Notwithstanding, for all three subgroups, variation at the area segment and stratum-level is minimal, as evidenced by the lower bound of the 95% credible interval being nearly zero.

Table 2.2: Variance components in the longitudinal model for PTSD severity score trajectories.

	Resilience	Recovery	Chronic
Variance	Posterior Mean (95% CrI)	Posterior Mean (95% CrI)	Posterior Mean (95% CrI)
Observation-level:			
σ_k^2	0.003 (0.003, 0.005)	0.02 (0.014, 0.029)	0.061 (0.04, 0.087)
Subject-level:			
ϕ_{11k}	0.008 (0.005, 0.014)	0.044 (0.028, 0.063)	0.064 (0.031, 0.109)
ϕ_{12k}	0.002 (0, 0.006)	-0.003 (-0.013, 0.009)	0.028 (0.004, 0.06)
ϕ_{13k}	0.002 (0, 0.006)	-0.001 (-0.01, 0.01)	-0.004 (-0.029, 0.023)
ϕ_{22k}	0.006 (0.004, 0.01)	0.022 (0.013, 0.037)	0.047 (0.022, 0.086)
ϕ_{23k}	0.002 (0, 0.006)	0.004 (-0.003, 0.014)	0.006 (-0.014, 0.03)
ϕ_{33k}	0.007 (0.004, 0.011)	0.017 (0.01, 0.029)	0.047 (0.021, 0.084)
Area segment-level:			
ω_k^2	0.002 (0, 0.004)	0.006 (0.001, 0.014)	0.022 (0.006, 0.045)
Stratum-level:			
ψ_k^2	0.003 (0, 0.015)	0.004 (0, 0.021)	0.024 (0, 0.13)

I classified subjects into the three subgroups of PTSD trajectories based on the maximum of the average posterior probabilities of belonging to each latent class over MCMC draws. Of 563 subjects, 274 (nearly 50%) are classified in the resilient subgroup, followed by 178 and 111 (approximately 30% and 20%) in the recovery and chronic subgroups, respectively. For the 274 subjects classified in the resilient subgroup, the median (mean) of the average posterior probabilities of belonging to this subgroup is 0.98 (0.94). The corresponding median (mean) for the recovery and chronic subgroups are 0.87 (0.84) and 0.99 (0.93), respectively.

2.4.3 Predicting latent class membership

Associations of the probability of latent class membership with PTSD risk factors and sample design variables are presented in Figure 2.5 and 2.6. In Figure 2.5, compared to the resilient subgroup, subjects in both the recovery and chronic subgroups are more likely to be older in age, to be black or Hispanic race, and to have higher peri-emotional reactions, Ike-related stress, and pre-Ike depression and PTSD. However, these associations are in general more pronounced in the chronic subgroup than in the recovery subgroup. Increasing average collective efficacy is associated with lower likelihood of being in the chronic versus resilient subgroup. Although none of the associations between the sample design variables and probability of latent class membership are significant, subjects with high PTSD risk tend to be more likely to belong in the recovery or chronic subgroups, and subjects in a household with three or more members tend to be more likely to belong to the resilient subgroup. Figure 2.6 presents the probability of latent class membership as a smoothed function of the number of occupied households. Among subjects from a given area segment, the probability of belonging to the resilient subgroup tends to increase with

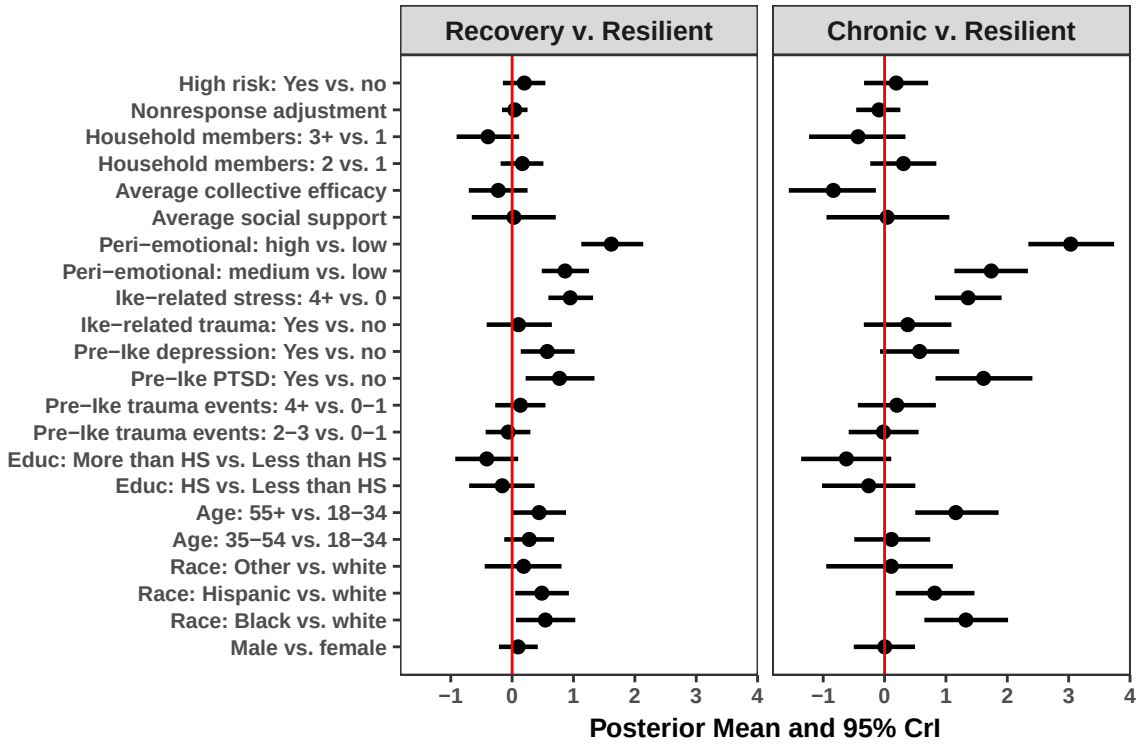

Figure 2.5: Probit regression coefficients, along with 95% credible intervals, for covariates in the latent class membership model.

the number of occupied households at the area segment-level, whereas I observe a modest negative association in the recovery subgroup and no association in the chronic subgroup.

Figure 2.7 shows variability in latent class membership among strata and area segments. Stratum 1 exhibits some evidence of being associated with higher probability of belonging to the recovery or chronic subgroups. This is consistent with the sample design because stratum 1 contained Galveston Island and the Bolivar Peninsula, which suffered severe damage from the storm. No difference is observed among the remaining four strata. Conditional on stratum, I also observe moderate variability in latent class membership among area segments, as measured by the sum of the terms u_{sjk} and ν_{sjk} in equation (2.4).

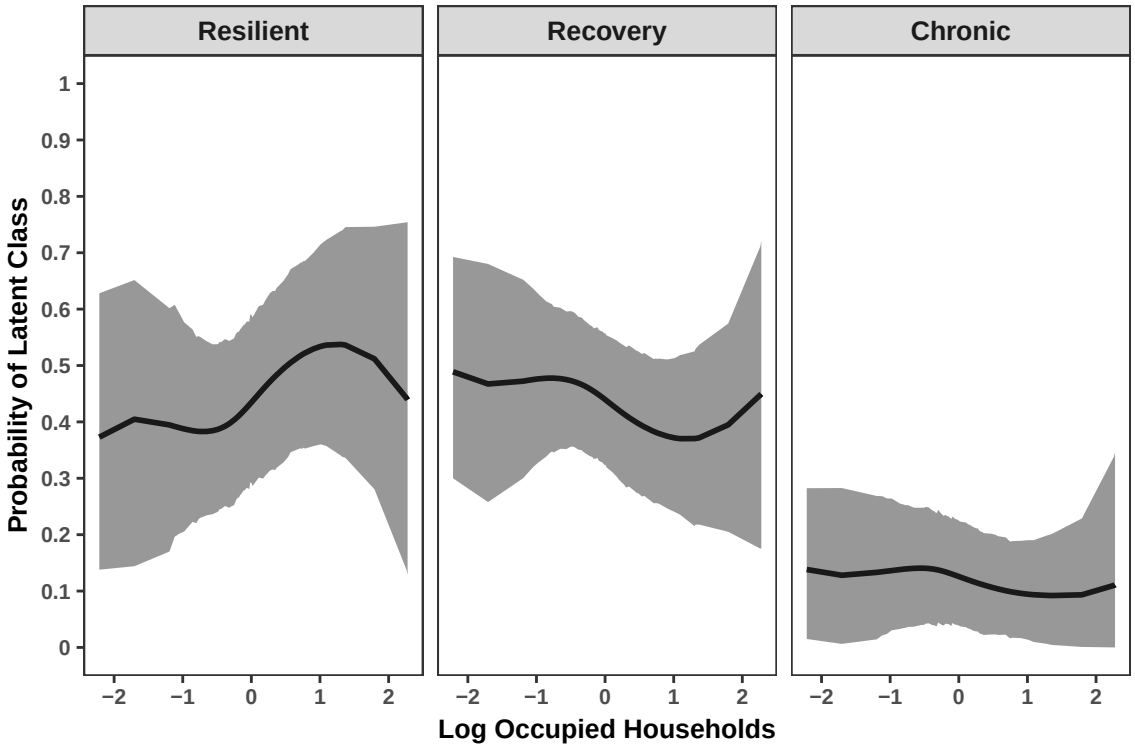

Figure 2.6: Probability of belonging to each latent class as a cubic B-spline of log occupied households, with knots at the distribution tertiles. The shaded region is the 95% highest posterior density interval of the B-spline.

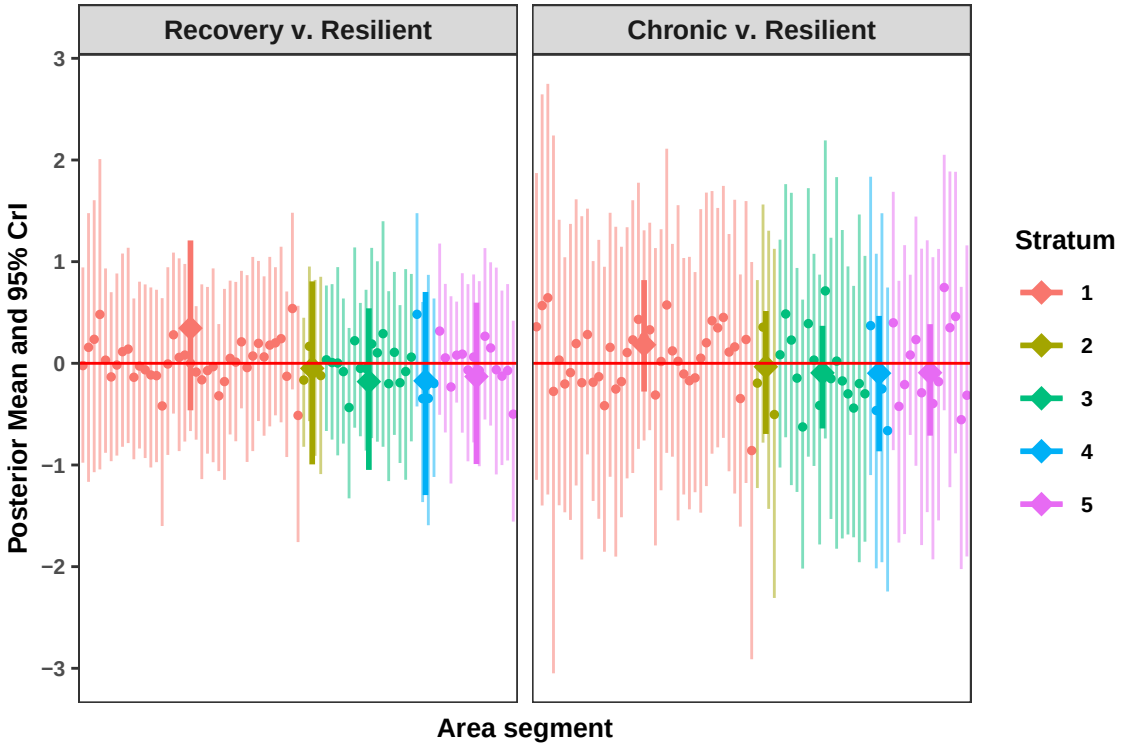

Figure 2.7: Estimation of stratum and area segment-specific intercepts using the posterior mean (diamond) and 95% credible interval (vertical bar) in the latent class membership model. The area segment-specific intercepts are the sum of u_{sjk} and ν_{sjk}. Coloring denotes area segments from the same stratum. The five stratum-specific intercepts are in bold font.

2.4.4 Model checking in the GBRS

Figure 2.8 presents a scatter plot of the predicted versus observed discrepancy measure T across MCMC samples. The Bayesian predictive p-value of 0.83 represents the proportion of samples above the diagonal, suggesting adequate overall model fit. Figure 2.9 shows histograms of the observed data overlaid by the posterior predictive distributions by subgroup and wave. The selected model fits the data reasonably well except for some observations with very high PTSD severity scores in the chronic subgroup.

2.5 Discussion

To my knowledge, this is the first study that uses Bayesian hierarchical modeling for incorporating a complex sample design into a finite mixture model, and specifically, a growth mixture model. By modeling variance components hierarchically to reflect the hierarchy of the data structure, with repeated measures nested within subjects, which are further clustered by area segments and strata, my method enables partitioning the variance across different levels of the data. In addition to modeling the effect of area segments using the typical independent random intercepts, I account for spatial correlations among neighboring area segments, thus relaxing the assumption that class membership risk is independent in geographic space. I develop an efficient Gibbs sampler including only closed-form full conditional distributions by using a probit link to model latent class membership probabilities, which largely reduces computational burden as compared to a logit link. My user-friendly `R` package `Bsvygmm` can be used for model fitting, selection, and checking.

Applying my proposed model to the GBRS, I found three clinically meaningful subgroups of PTSD severity score trajectories, namely, resilience, recovery, and chronic. Incorporating

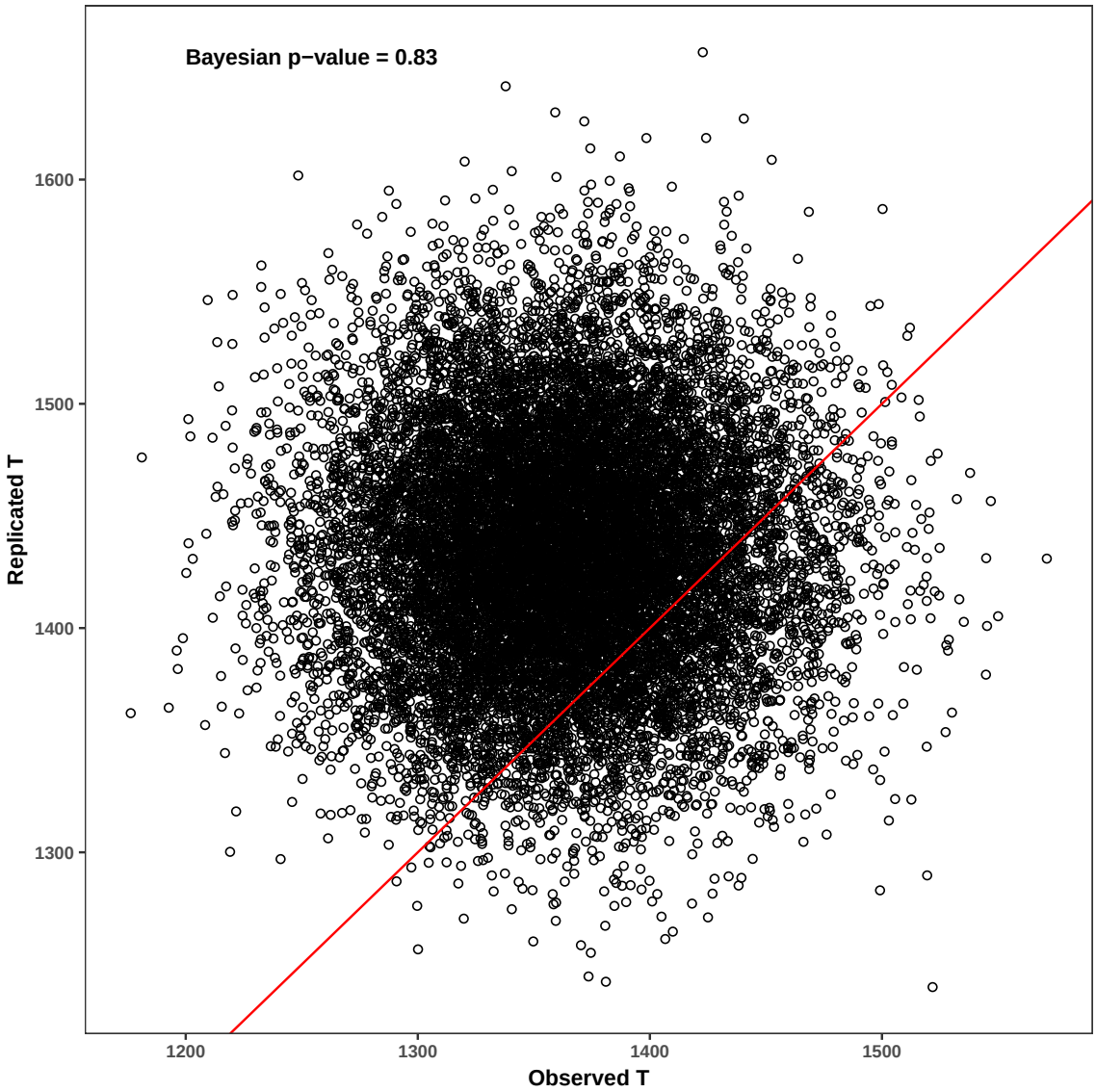

Figure 2.8: Posterior predictive checking for the selected model using a Bayesian posterior predictive p-value. Observed T is computed using the observed data. Replicated T is computed using the replicated datasets from the posterior predictive distribution.

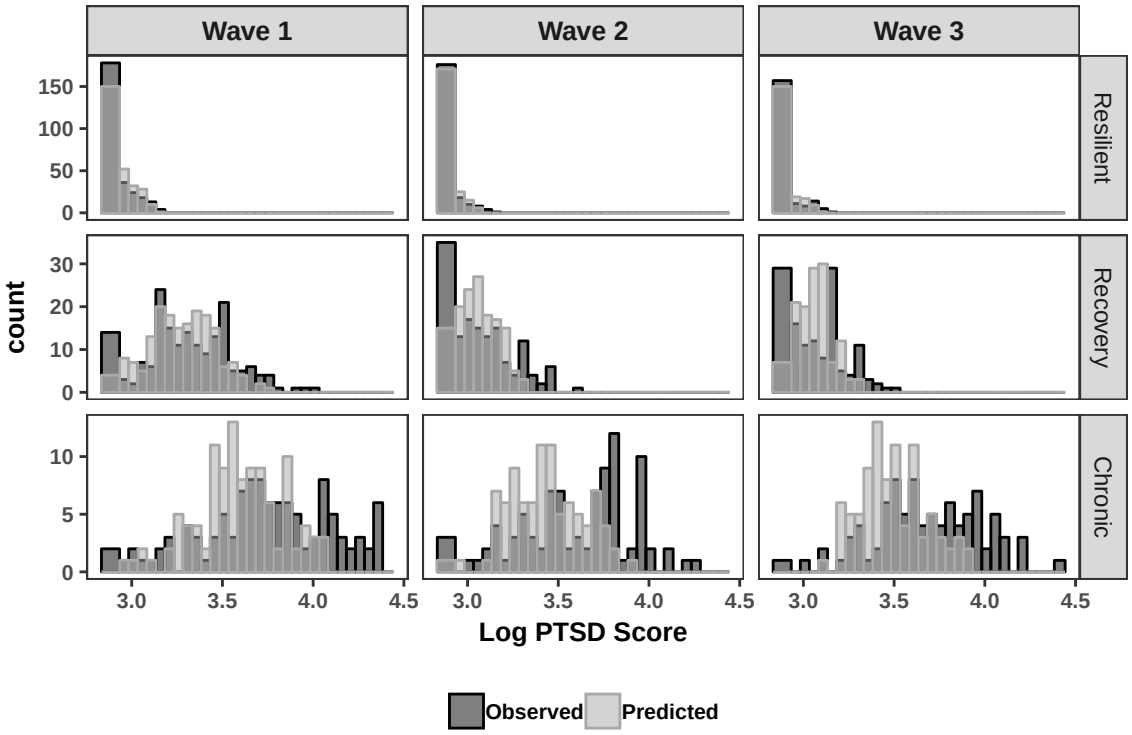

Figure 2.9: Histograms of the observed data and the posterior predictive distribution of log PTSD severity scores by subgroup and wave of survey. The posterior predictive distribution is summarized using the median draw over MCMC samples.

the sample design can affect estimation of the optimal number of latent classes, latent class proportions, latent class-specific regression coefficients, and actual subject classification [Patterson *et al.*, 2002; Wedel *et al.*, 1998]. For example, in previous research with the GBRS, [Lowe *et al.*, 2015] used latent class growth analysis (LCGA) implemented by the *TRAJ* procedure in SAS [Jones *et al.*, 2001; Jones and Nagin, 2007] and found a fourth subgroup (5% of the study sample) that exhibited a delayed PTSD score trajectory, defined as initially low symptomatology that increases over time. The difference in the number of latent classes may be, in part, because given latent class, LCGA does not account for correlation among repeated measures of PTSD scores from the same subject, or because [Lowe *et al.*, 2015]'s analysis ignored features of the complex sample design. In a sensitivity analysis (see Appendix C), I fit alternative Bayesian GMMs assuming $K = 2, 3, 4$ latent classes that removed all information about the complex sample design. Based on the BIC and ICL-BIC, the 3-class model was preferred, but the 4-class model was selected according to the DIC4 (Table C.1). However, the additional class in the 4-class model exhibited low posterior probability of class assignment. In comparing the 3-class models with and without complex sample design, estimation of the average latent class-specific trajectories and variance components was similar (see Figure C.1 corresponding to Figure 2.4 and Table C.2 corresponding to Table 2.2), and only 37 of the 563 subjects in the GBRS differed in their latent class membership.

A Bayesian hierarchical modeling approach requires careful consideration of model specification of the design features. In my proposed GMM, I use design variables – in addition to PTSD-related risk factors measured at baseline – to predict latent class membership. To flexibly model the effect of pps sampling on the probability of class membership, I include the continuous size variable, the number of occupied households from the census,

using B-splines. Because the design variables are not strongly associated with the probability of latent class membership, I do not include interaction effects. Stratum and area segment-level effects are included in both the latent class membership model and the longitudinal model of PTSD scores. My analysis, however, reveals little variability in PTSD scores at the stratum and area segment-levels after conditioning on latent class, which may suggest removing these variance components from the longitudinal model of PTSD scores. By including the other sample design variables only in the latent class membership model, I assume that they are independent of PTSD trajectories given latent class. Overall, my Bayesian GMM analysis suggests that design features play a relatively small role in predicting PTSD score trajectories. This may explain why the 3-class models with and without accounting for the complex sample design yield similar mean and variance estimates. Unlike the pseudo-likelihood method, in my Bayesian approach, variance estimation will not be inflated when design features are unnecessarily included in the model.

In the GBRS, not all subjects who participated in the baseline survey completed the two follow-up surveys. I assumed that PTSD scores are missing at random; however, particularly with mental health data, the probability of a missing value may depend on unobserved PTSD scores even after conditioning on latent class. In future research, I will conduct sensitivity analyses to assess the missing at random assumption. In addition, my analysis suggests that subjects in the chronic subgroup have a more dispersed PTSD distribution than assumed in my normal model. Future research may explore other distributional assumptions.

My proposed GMM for analyzing complex survey data using a Bayesian approach has practical utility for planning and allocating post-disaster services. Classification of disaster survivors into their trajectory subgroups provides information about the extent to which

different types of interventions are needed, the efficacious timing of these interventions, and the tailoring of these interventions to specific risk profiles. Especially important in the context of precision public health, my proposed GMM provides subject-specific inference. For example, I can use subject-specific predictions to identify individuals who have higher than average PTSD scores compared to other individuals with similar geographic, demographic, and health characteristics. Post-disaster services can be targeted not only within subgroup, but also within geographic areas, for individuals with specific combinations of risk factors, and for individuals themselves. Moreover, information about predominant levels of variability in PTSD can suggest cost-effective scales at which to implement an intervention, which is critical post-disaster when resources are scarce.

Chapter 3

Modeling Heterogeneity and Missing Data in Electronic Health Records

3.1 Introduction

Longitudinal data collected in electronic health records (EHRs) are big data. As EHRs are increasingly adopted in US health systems, an estimated one billion patient visits may be documented per year [Hripcsak and Albers, 2013]. A natural feature of such data may be unobserved, or "latent" heterogeneity, whereby unobservable subgroups of patients are characterized by distinctive patterning in their longitudinal health trajectories. Researchers from diverse biomedical fields, such as psychology [Elliott *et al.*, 2005] and maternal and infant health [Neelon *et al.*, 2011], have used growth mixture models (GMMs) [Muthen *et al.*, 2002; Verbeke and Lesaffre, 1996] to analyze latent heterogeneity in longitudinal data. GMMs enable classifying subjects into different subgroups, often called latent classes, according to individual longitudinal trajectories and risk factors hypothesized to be associated with class

membership.

Despite the potential for new scientific insights from analyzing the vast amounts of data in EHRs, one of the primary challenges faced by researchers is how to handle the often large numbers of missing values [Weiskopf and Weng, 2013]. Unlike in longitudinal data collected in a designed study, in EHRs, two patient-led missing data processes drive the generation of missing values, namely, the visit process and the response process given a clinic visit. In the absence of follow-up times fixed *a priori* by the study design, the visit process refers to the probability that patients themselves decide to visit the clinic, which may be based on a patient's own prerogative, physician recommendation, or a combination thereof. Without a set of variables for data collection fixed before study onset, the response process given a clinic visit refers to the probability of observing a response on a given EHR variable, conditional on a patient visiting the clinic. This may be based, in part, on a patient's stated medical reasons for the visit, in addition to clinical judgement. Multiplied over huge patient populations in EHRs, the visit process and the response process given a clinic visit spawn innumerable patterns in missingness over time, which may themselves be characterized by latent heterogeneity.

For valid statistical inferences with EHRs, the missing data mechanisms for the visit process and the response process given a clinic visit require careful attention. When the probability of a missed visit is related to the underlying process generating the longitudinal outcomes, the visit process is characterized as a special case of missing not at random (MNAR), termed informative [Wu and Carroll, 1988; Follmann and Wu, 1995]. In EHR-based research, because a patient's underlying health status may be associated with when and how often the patient visits the clinic, longitudinal data analysis may be subject to an informative visit process. Existing methods to handle an informative visit process rely on

shared parameter modeling [Wu and Carroll, 1988; Follmann and Wu, 1995] in which the longitudinal outcomes and visit processes are jointly modeled on the basis of a conditional independence assumption that includes – at a minimum – shared continuous or discrete latent variables [Liang *et al.*, 2009; Sun *et al.*, 2007; McCulloch *et al.*, 2016; Lin *et al.*, 2004]. However, to my knowledge, no methods have been developed for the setting of EHRs where in addition to the visit process, the response process given a clinic visit may exhibit an MNAR mechanism.

In this paper, I propose a Bayesian shared parameter model to model latent heterogeneity in multiple longitudinal health outcomes in EHRs, while accounting for MNAR missing data mechanisms for the visit process and response process given a clinic visit. My focus is on longitudinal health outcomes in EHRs for which there is a clinically prescribed visit schedule, which I use to construct time windows of observation to measure each patient's visit process. For example, my data application is on early childhood weight and height measurements, which according to the American Academy of Pediatrics, should be collected according to the well-child check schedule [American Academy of Pediatrics, 2018]. Conditional on observing a visit in a given clinical time window, I measure the response process for each health outcome.

The proposed shared parameter model links GMMs of the longitudinal health outcomes, the visit process, and the response process given a clinic visit using a discrete latent variable to indicate the latent class to which each patient belongs. Conditional on a patient's latent class membership, the longitudinal health outcomes, the visit process, and the response process given a clinic visit are assumed to be independent. The use of the discrete latent class variable [Lin *et al.*, 2004; Roy, 2003] to link the health outcomes and missing data processes confers three main advantages in the EHR setting: First, I can relax the as-

sumption of a single, homogeneous patient population in modeling longitudinal trajectories of health outcomes, the visit process, and the response process given a clinic visit, while having population-averaged inferences at my disposal if I so desire. Second, I can tractably summarize the innumerable patterns of missing values from the visit process and response process given a clinic visit into a small number of latent classes. Third, I can easily alter my MNAR assumption about the visit process or response process given a clinic visit to handle ignorable missing data mechanisms. For model estimation, I developed an efficient Markov chain Monte Carlo (MCMC) algorithm that is based on easily sampled closed-form full conditional distributions. I developed the `R` package `EHRMiss` for model fitting, selection, and checking.

3.2 Statistical Method

I formulate the proposed model of longitudinal health outcomes among patients in EHRs accounting for MNAR missing data mechanisms for the visit process and the response process given a clinic visit. First, in Section 3.2.1, I present the Bayesian multivariate GMM for complete data. In Section 3.2.2, I extend the complete-data model to account for a nonignorable visit process and response process given a clinic visit, followed by an explication of the missing data mechanisms in Section 3.2.3. Sections 3.2.4 and 3.2.5 detail prior distributions and posterior computation, respectively, followed by model selection in 3.2.6 and model checking in 3.2.7.

3.2.1 Complete-data model

Suppose there are K latent classes of patients with distinctive patterning in their trajectories of R health outcomes collected over J prescribed time windows for clinical observation. The

Bayesian multivariate GMM for the complete-data model comprises two submodels, namely, the latent class membership model and the longitudinal health outcomes model. I begin with the latent class membership model.

Let c_i be a discrete latent variable taking values $k = 1, \ldots, K$ to indicate the latent class membership of patient i $(i = 1, \ldots, n)$. I assume that

$$c_i \sim Multinomial\left(1;\, \pi_{i1}, \ldots, \pi_{iK}\right), \tag{3.1}$$

where π_{ik} are patient-specific latent class membership probabilities that I model by adapting a multinomial probit regression framework [McCulloch and Rossi, 1994].

To connect π_{ik} with latent class membership c_i, I introduce K latent normal random variables ξ^*_{ik} $(k = 1, \ldots, K)$ with unknown mean and variance-covariance, where $\pi_{ik} = Pr(\xi^*_{ik} > \xi^*_{il} \text{ for all } l \neq k)$. Following standard practice, I define latent class K as the reference level by taking the difference $\xi_{ik} = \xi^*_{ik} - \xi^*_{iK}$ for $k = 1, \ldots, K-1$. Then, I specify the multinomial probit model as

$$\xi_{ik} = \mathbf{w}_i \delta_k^T + \epsilon_{ik}, \tag{3.2}$$

with

$$c_i \;=\; \begin{cases} K & \text{if } max(\xi_{i1}, \ldots, \xi_{iK-1}) < 0 \\ k & \text{if } max(\xi_{i1}, \ldots, \xi_{iK-1}) = \xi_{ik} \geq 0 \text{ for } k = 1, \ldots, K-1. \end{cases} \tag{3.3}$$

In (3.2), the latent normal random variables ξ_{ik} are modeled as a function of $\mathbf{w}_i$ $(1 \times s)$, which includes patient-level risk factors and a column of ones for an intercept. Corresponding regression coefficients are contained in δ_k. ϵ_{ik} $(k = 1, \ldots, K-1)$ are normal random errors with mean zero, whose variance-covariance I restrict to the identity matrix in order to address identifiability issues in the multinomial probit [Daganzo, 1979; Dansie, 1985; Bunch, 1991]. For $K = 2$, this set-up corresponds to the standard Bayesian probit model

for a binary outcome [Albert and Chib, 1993]. Equation (3.3) defines c_i according to the values of the latent normal random variables ξ_{ik} $(k = 1, \ldots, K-1)$.

The multivariate model of longitudinal health outcomes is specified conditional on latent class membership. Let $y_{1ij}, \ldots, y_{Rij}$ be longitudinal measurements on R health outcomes in clinical time window j. Then,

$$\left[\begin{array}{c|c} y_{1ij} & \\ \vdots & c_i = k \\ y_{Rij} & \end{array}\right] \sim MVN_R \left(\begin{bmatrix} \beta_{1k}\mathbf{x}_{ij}^T + \mathbf{b}_{1i}\mathbf{z}_{ij}^T \\ \vdots \\ \beta_{Rk}\mathbf{x}_{ij}^T + \mathbf{b}_{Ri}\mathbf{z}_{ij}^T \end{bmatrix}, \boldsymbol{\Sigma}_k \right) \tag{3.4}$$

$$\left[\begin{array}{c|c} \mathbf{b}_{1i} & \\ \vdots & c_i = k \\ \mathbf{b}_{Ri} & \end{array}\right] \sim MVN_{Rq} \left(\begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \boldsymbol{\Psi}_k \right). \tag{3.5}$$

In (3.4), conditional on latent class membership, the longitudinal health outcomes y_{rij} $(r = 1, \ldots, R)$ are modeled as a polynomial function of a patient's age in window j, with polynomial terms and a column of ones for an intercept included in $\mathbf{x}_{ij}^T$ $(p \times 1)$. The corresponding regression coefficients in β_{rk} $(1 \times p)$ capture the average health trajectory in latent class k, and $\boldsymbol{\Sigma}_k$ is an $R \times R$ latent class-specific variance-covariance among y_{rij} $(r = 1, \ldots, R)$. For each outcome r, $\mathbf{b}_{ri} = (b_{ri1}, \ldots, b_{riq})^T$ $(1 \times q)$ are patient-specific random effects associated with $\mathbf{z}_{ij}^T$, the columns of which are a subset of $\mathbf{x}_{ij}^T$. As shown in (3.5), $\mathbf{b}_{ri}$ are modeled conditional on a patient's latent class membership, thus reflecting patient-specific variability around the average health trajectory in a given latent class. The latent class-specific variance-covariance $\boldsymbol{\Psi}_k$ is an $Rq \times Rq$ block diagonal matrix with entries $\boldsymbol{\Psi}_{kr}$ $(q \times q)$, the elements of which compose a variance-covariance for $\mathbf{b}_{ri}$ $(i = 1, \ldots, n)$. Note that for simplicity, I assume that $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are the same for all health outcomes r; however, this is not required.

3.2.2 Nonignorable visit process and response processes given a clinic visit

I extend the complete-data model in (3.1) - (3.5) to allow for missing values from the visit process and the response process given a clinic visit. To account for nonignorable missing data mechanisms for the visit process and the response process given a clinic visit, I build a shared parameter model through which the longitudinal health outcomes, visit process, and response process given a clinic visit are linked via the discrete latent variable c_i for a patient's latent class membership.

To specify the full data, corresponding to the elements $y_{ri1}, \ldots, y_{riJ}$, let d_{ij} $(j = 1, \ldots, J)$ be an indicator for the visit process such that $d_{ij} = 1$ if patient i has a clinic visit during time window j, and 0 otherwise. The response process for the r^{th} health outcome given a clinic visit is defined for the subset of time windows when patient i visits the clinic. Let $A = \{j : \ d_{ij} = 1 \text{ for } j = 1, \ldots, J\}$, and let the total number of clinic visits for patient i be $n_i = \sum_{j=1}^{J} d_{ij}$. Then, for $l = 1, \ldots, n_i$, define $m_{riA(l)} = 1$ if a response is observed for health outcome r at window $A(l)$, and 0 otherwise. The full data are given by y_{rij}, d_{ij}, and $m_{riA(l)}$.

Using a probit link function, I model the probability of a clinic visit for patient i in time window j as

$$\left[\ d_{ij} \ \middle| \ c_i = k \ \right] \sim Bernoulli \left(\Phi\{\mathbf{x}_{ij}\phi_k^T + \mathbf{z}_{ij}\tau_i^T\} \right) \tag{3.6}$$

$$\left[\ \tau_i \ \middle| \ c_i = k \ \right] \sim MVN_q \left(\mathbf{0}, \mathbf{\Omega}_k \right), \tag{3.7}$$

where $\Phi\{.\}$ is the cumulative distribution function of the standard normal distribution. In (3.6) and (3.7), ϕ_k $(1 \times p)$ are latent class-specific regression coefficients associated with $\mathbf{x}_{ij}$ that capture the average visit process trajectory in latent class k. τ_i $(1 \times q)$ are patient-

specific random effects associated with $\mathbf{z}_{ij}$ and modeled conditional on latent class membership with a $q \times q$ variance-covariance $\mathbf{\Omega}_k$. τ_i $(i = 1, \ldots, n)$ reflect within latent class variability around the average visit process trajectory.

Analogous to the visit process model, the probability of response for health outcome r in window $A(l)$ is specified as

$$\left[\; m_{riA(l)} \; \middle| \; c_i = k \; \right] \sim Bernoulli \left(\Phi\{\mathbf{x}_{iA(l)}\lambda_{rk}^T + \mathbf{z}_{iA(l)}\kappa_{ri}^T\} \right) \tag{3.8}$$

$$\left[\; \kappa_{ri} \; \middle| \; c_i = k \; \right] \sim MVN_q \left(\mathbf{0}, \, \mathbf{\Theta}_{rk} \right), \tag{3.9}$$

where λ_{rk} $(1 \times p)$ represent the latent class-specific average response process for health outcome r; and, κ_{ri} are patient-specific random effects associated with $\mathbf{z}_{iA(l)}$ that are modeled with a latent class-specific variance-covariance $\mathbf{\Theta}_{rk}$ $(q \times q)$. As in the visit process model, κ_{ri} $(i = 1, \ldots, n)$ capture variability within a latent class around the average response trajectory. To simplify notation, I have assumed that the visit process and response process given a clinic visit use the same design matrices as in the longitudinal health outcomes model in (3.4), but this is unnecessary in practice.

3.2.3 Missing data mechanism

Let $\mathbf{y}_{ij} = (y_{1ij}, \ldots, y_{RiJ})^T$ and $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{iJ}^T)$. Let $\mathbf{b}_i = (\mathbf{b}_{1i}^T, \ldots, \mathbf{b}_{Ri}^T)^T$. Let $\mathbf{d}_i = (d_{i1}, \ldots, d_{iJ})^T$, and $\mathbf{m}_{ri} = (m_{riA(1)}, \ldots, m_{riA(n_i)})^T$ for $r = 1, \ldots, R$. Let there be a partition of the longitudinal health outcomes $\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^m)$ for observed (o) and missing (m) components, where $\mathbf{y}_i^m$ can be decomposed between missed clinic visits, $\mathbf{y}_i^{m1}$, and missed responses given a clinic visit, $\mathbf{y}_i^{m2}$. To examine the missing data mechanism, I consider the

joint density

$$
\begin{aligned}
& f(\mathbf{y}_i^o, \mathbf{y}_i^{m2}, \mathbf{b}_i;\, \mathbf{d}_i, \tau_i;\, \mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri}, \kappa_{1i}, \ldots, \kappa_{Ri};\, c_i \,|\, rest) \\
&= \int_{\mathbf{y}_i^{m1}} f(\mathbf{y}_i^o, \mathbf{y}_i^{m1}, \mathbf{y}_i^{m2}, \mathbf{b}_i;\, \mathbf{d}_i, \tau_i;\, \mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri}, \kappa_{1i}, \ldots, \kappa_{Ri};\, c_i \,|\, rest)\, \partial \mathbf{y}_i^{m1},
\end{aligned}
$$

with the factorization

$$
\begin{aligned}
& f(\mathbf{y}_i^o, \mathbf{y}_i^{m2}, \mathbf{b}_i;\, \mathbf{d}_i, \tau_i;\, \mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri}, \kappa_{1i}, \ldots, \kappa_{Ri};\, c_i \,|\, rest) \\
&= f(\mathbf{y}_i^o, \mathbf{y}_i^{m2} \,|\, c_i, \mathbf{b}_i)\, f(\mathbf{b}_i \,|\, c_i) \\
&\times f(\mathbf{d}_i \,|\, c_i, \tau_i)\, f(\tau_i \,|\, c_i) \\
&\times f(\mathbf{m}_{1i} \,|\, c_i, \kappa_{1i})\, f(\kappa_{1i} \,|\, c_i) \ldots f(\mathbf{m}_{Ri} \,|\, c_i, \kappa_{Ri})\, f(\kappa_{Ri} \,|\, c_i) \\
&\times f(c_i).
\end{aligned}
$$

Conditional on latent class membership, the longitudinal health outcomes, visit process, and response process given a clinic visit are assumed to be independent. The MNAR mechanism is evident because the visit process and the response process given a clinic visit depend on $\mathbf{y}_i^{m2}$ indirectly through latent class membership.

The proposed shared parameter model can be easily altered to accommodate an MAR mechanism for one or both of the visit process and response process given a clinic visit. For example, the visit process is MAR if $f(\mathbf{d}_i, \tau_i \,|\, c_i, rest) = f(\mathbf{d}_i, \tau_i \,|\, rest)$. Conditional on observed information, the visit process and the associated patient-specific random effects are assumed to be independent of latent class. Under an MAR mechanism and the assumption of separable parameter spaces, the visit process can be ignored in statistical analysis.

3.2.4 Prior specification

To complete the Bayesian model specification, I assign prior distributions to all of the parameters. For each parameter, I use the same prior distribution across mixture components.

In the latent class membership model, I follow previous research [Garrett and Zeger, 2000; Elliott *et al.*, 2005] by assigning the probit regression coefficients δ_k in (3.2) a prior distribution $MVN_s(\mathbf{0}, \mathbf{I})$. On the probability scale, this prior distribution yields a non-informative prior on the probability of latent class membership, with its mode at approximately $\frac{1}{K}$.

In the longitudinal health outcomes model (3.4), I assign the latent-class specific regression coefficients β_{rk} a diffuse prior distribution of the form $MVN_p(\mathbf{0}, \mathbf{\Sigma}_\beta)$, where $\mathbf{\Sigma}_\beta$ is a diagonal variance-covariance with some large variance. I assign the observation-level variance-covariance $\mathbf{\Sigma}_k$ an inverse-Wishart prior distribution $IW(\nu_\Sigma, \mathbf{S}_\Sigma^{-1})$, where ν_Σ is the degrees of freedom and $\mathbf{S}_\Sigma^{-1}$ is a positive definite matrix. In (3.5), for the hierarchical variance-covariance of the patient-specific random effects $\mathbf{\Psi}_k$, I assign each of the constituent variance-covariances $\mathbf{\Psi}_{kr}$ an inverse-Wishart prior distribution.

Like the model of longitudinal health outcomes, for the visit process model in (3.6) and (3.7) and the response process model in (3.8) and (3.9), I use diffuse normal prior distributions on the latent class-specific regression coefficients ϕ_k and λ_{rk}, and inverse-Wishart prior distributions on the hierarchical variance-covariances $\mathbf{\Omega}_k$ and $\mathbf{\Theta}_{rk}$.

3.2.5 Posterior computation

Let $\mathbf{y}_{iA(l)} = (y_{1iA(l)}, \ldots, y_{RiA(l)})^T$, and $\beta_k = (\beta_{1k}^T, \ldots, \beta_{Rk}^T)^T$. Assuming prior independence, I specify the joint posterior distribution as

$$
\begin{aligned}
& p(\mathbf{c};\, \beta, \mathbf{b}, \boldsymbol{\Sigma}, \boldsymbol{\Psi};\, \phi, \tau, \boldsymbol{\Omega};\, \lambda, \kappa, \boldsymbol{\Theta} \,|\, \mathbf{y}, \mathbf{d}, \mathbf{m}; \mathbf{x}, \mathbf{z}, \mathbf{w}) \\
&= \prod_{k=1}^{K} \left\{ \prod_{i=1}^{n} \pi_{ik} \left[\left(\prod_{j=1}^{J} f(d_{ij} \,|\, \tau_i, \phi_k)\, f(\tau_i \,|\, \boldsymbol{\Omega}_k) \right) \right.\right. \\
&\times \left.\left. \prod_{l=1}^{n_i} \left(f(\mathbf{y}_{iA(l)} \,|\, \mathbf{b}_i, \beta_k, \boldsymbol{\Sigma}_k)\, f(\mathbf{b}_i \,|\, \boldsymbol{\Psi}_k) \prod_{r=1}^{R} f(m_{riA(l)} \,|\, \kappa_{ri}, \lambda_{rk})\, f(\kappa_{ri} \,|\, \boldsymbol{\Theta}_{rk}) \right) \right]^{\mathbf{1}_{c_i=k}} \right. \\
&\times \left. p(\beta_k)\, p(\boldsymbol{\Sigma}_k)\, p(\boldsymbol{\Psi}_k)\, p(\phi_k)\, p(\boldsymbol{\Omega}_k) \prod_{r=1}^{R} p(\lambda_{rk})\, p(\boldsymbol{\Theta}_{rk}) \right\} \prod_{k=1}^{K-1} p(\delta_k),
\end{aligned}
$$

where for notational simplicity, the design matrices in the conditional densities for d_{ij}, $\mathbf{y}_{iA(l)}$, and $m_{riA(l)}$ are suppressed.

For posterior computation, I propose an MCMC algorithm that uses easily sampled closed-form full conditionals. After assigning initial values to model parameters, the algorithm iterates among the following steps:

1. For $k = 1, \ldots, K-1$, update δ_k for the latent class membership model in (3.2). Compute π_{ik} for $k = 1, \ldots, K$ in (3.1).

2. For $k = 1, \ldots, K$, update parameters for the longitudinal health outcomes model in (3.4) and (3.5), including β_{rk}, $\mathbf{b}_{ri}$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\Psi}_k$.

3. For $k = 1, \ldots, K$, update parameters for the visit process model in (3.6) and (3.7), including ϕ_k, τ_i, and $\boldsymbol{\Omega}_k$.

4. For $k = 1, \ldots, K$, update parameters for the model of the response process given a clinic visit in (3.8) and (3.9), including λ_{rk}, κ_{ri}, and $\boldsymbol{\Theta}_{rk}$.

5. Sample latent class indicators c_i for $i = 1, \ldots, n$ from $Multinomial(1; p_{i1}, \ldots, p_{iK})$, where $p_{i1}, \ldots, p_{iK}$ are the posterior probabilities of latent class assignment given by

$$\begin{aligned} & p_{ik} \qquad (3.10) \\ &= Pr(c_i = k \,|\, \pi_{ik};\, \mathbf{y}_i^*, \mathbf{b}_i;\, \mathbf{d}_i, \tau_i;\, \mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri}, \kappa_{1i}, \ldots, \kappa_{Ri};\, rest) \\ &\propto \pi_{ik}\, f(\mathbf{y}_i^* \,|\, \mathbf{b}_i, \beta_k, \boldsymbol{\Sigma}_k^*)\, f(\mathbf{b}_i \,|\, \boldsymbol{\Psi}_k) \\ &\times f(\mathbf{d}_i \,|\, \tau_i, \phi_k)\, f(\tau_i \,|\, \boldsymbol{\Omega}_k) \\ &\times \prod_{r=1}^{R} f(\mathbf{m}_{ri} \,|\, \kappa_{ri}, \lambda_{rk})\, f(\kappa_{ri} \,|\, \boldsymbol{\Theta}_{rk}), \end{aligned}$$

where $\mathbf{y}_i^* = (\mathbf{y}_{iA(1)}^T, \ldots, \mathbf{y}_{iA(n_i)}^T)$, and $\boldsymbol{\Sigma}_k^*$ is an $n_i R \times n_i R$ block diagonal matrix with elements $\boldsymbol{\Sigma}_k$ $(R \times R)$ for each $\mathbf{y}_{iA(l)}$ $(l = 1, \ldots, n_i)$.

The full MCMC algorithm is detailed in Appendix D.

3.2.6 Model selection

I use model selection as a tool to guide sensitivity analysis about missing data assumptions. First, I select the optimal number of latent classes among models with the same assumed missing data mechanism. Then, assuming each of the selected number of latent classes, I fit models varying the missing data assumptions. This approach enables investigating the sensitivity of statistical inferences to missing data assumptions given the selected number of latent classes.

To conduct model selection, I use two model information criteria and a graphical technique known as latent class identifiability displays (LCIDs) [Garrett and Zeger, 2000]. The model information criteria include the Bayesian Information Criterion (BIC) [Schwarz, 1978], and a modified version of the Deviance Information Criterion (DIC) [Spiegelhalter *et al.*, 2002] known as the DIC3 [Celeux *et al.*, 2006]. I calculate the BIC using the

marginal density of $\mathbf{y}_i^*$, $\mathbf{d}_i$, $\mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri}$ after integrating out latent class membership c_i and the random effects $\mathbf{b}_i$, τ_i, $\kappa_{1i}, \ldots, \kappa_{Ri}$ for each of the outcomes, given by

$$
\begin{aligned}
& f(\mathbf{y}_i^*, \mathbf{d}_i, \mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri} \,|\, \pi;\, \beta, \boldsymbol{\Sigma}, \boldsymbol{\Psi};\, \phi, \boldsymbol{\Omega};\, \lambda, \boldsymbol{\Theta}) \qquad (3.11) \\
&= \sum_{k=1}^{K} \pi_{ik} \left(\int_{b_i} f(\mathbf{y}_i^* \,|\, \mathbf{b}_i, \beta_k, \boldsymbol{\Sigma}_k^*)\, f(\mathbf{b}_i \,|\, \boldsymbol{\Psi}_k)\, \partial \mathbf{b}_i \right) \\
&\times \left(\int_{\tau_i} f(\mathbf{d}_i \,|\, \tau_i, \phi_k)\, f(\tau_i \,|\, \boldsymbol{\Omega}_k)\, \partial \tau_i \right) \\
&\times \left(\int_{\kappa_{Ri}} \cdots \int_{\kappa_{1i}} f(\mathbf{m}_{1i} \,|\, \kappa_{1i}, \lambda_{1k})\, f(\kappa_{1i} \,|\, \boldsymbol{\Theta}_{1k}) \ldots f(\mathbf{m}_{Ri} \,|\, \kappa_{Ri}, \lambda_{Rk})\, f(\kappa_{Ri} \,|\, \boldsymbol{\Theta}_{Rk})\, \partial \kappa_{1i}, \ldots, \partial \kappa_{Ri} \right),
\end{aligned}
$$

where I can analytically compute only the integral for the longitudinal health outcomes $\mathbf{y}_i^*$. I estimate the integrals for the visit process $\mathbf{d}_i$ and response process given a clinic visit $\mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri}$ using numerical integration. I then define the BIC as

$$
\text{BIC} \;=\; \sum_{i=1}^{n} \log f(\mathbf{y}_i^*, \mathbf{d}_i, \mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri} \,|\, \hat{\pi};\, \hat{\beta}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\Psi}};\, \hat{\phi}, \hat{\boldsymbol{\Omega}};\, \hat{\lambda}, \hat{\boldsymbol{\Theta}}) + d \log N_{\text{eff}},
$$

where $\hat{\pi}$, $\hat{\beta}$, $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Psi}}$, $\hat{\phi}$, $\hat{\boldsymbol{\Omega}}$, $\hat{\lambda}$, $\hat{\boldsymbol{\Theta}}$ are the Bayesian estimators of the unknown parameters; d is the number of free parameters in the mixture model; and N_{eff} is the effective sample size from the model of longitudinal health outcomes $\mathbf{y}_i^*$ estimated by accounting for the correlations among the longitudinal measurements belonging to same patient [Jones, 2011]. The first term is a measure of goodness of fit, and the second term provides a penalty for model complexity.

In Bayesian hierarchical models, the number of free parameters may be unclear. As the Bayesian analogue to the BIC, [Spiegelhalter *et al.*, 2002] proposed the DIC in which the number of effective parameters is estimated. For some unknown parameter α, the DIC is computed as $\text{DIC} = \bar{D}(\alpha) + p_D$, where $\bar{D}(\alpha)$ is the posterior mean deviance estimated from MCMC samples, and p_D is the effective number of parameters taken as $p_D = \bar{D}(\alpha) - D(\hat{\alpha})$. The second term, $D(\hat{\alpha})$, is the point estimate for the deviance and is standardly evaluated

at the posterior mean estimator of α. However, according to [Celeux *et al.*, 2006], in finite mixture modeling, the posterior mean estimator often leads to a negative effective number of parameters. The authors recommend the DIC3, in which the posterior mean estimator is replaced by the estimator of the marginal density (3.11) obtained from MCMC samples. Analogous to the BIC, $\bar{D}(\alpha)$ is a measure of goodness of model fit, while p_D is a penalty for model complexity. Smaller values of BIC and DIC3 indicate a preferred model.

[Garrett and Zeger, 2000] propose using LCIDs to examine the extent to which the data are able to distinguish among the assumed number of latent classes. In LCIDs, plots of the prior versus posterior distributions for regression coefficients in the latent class membership model are examined. Largely overlapping prior and posterior distributions may suggest that the number of latent classes is too large given the data.

3.2.7 Model checking

For model checking under MAR or MNAR missing data mechanisms, previous research [Gelman *et al.*, 2005] has recommended conducting Bayesian posterior predictive checking [Gelman *et al.*, 1996] with completed datasets that include observed and imputed data, and replicates of the completed datasets drawn from the complete-data model in (3.1) - (3.5). At each MCMC iteration, a discrepancy measure is computed using the completed and replicated completed datasets. The Bayesian predictive p-value denotes the probability that the discrepancy measure under the replicated completed data is greater than that under the completed data, with p-values outside the range of 0.05 and 0.95 suggesting a lack of model fit. To examine overall model adequacy, I use the multivariate mean square error for my discrepancy measure [Daniels and Hogan, 2008], which for the complete-data

model in (3.1) - (3.5), I compute as

$$T = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{l=1}^{n_i}(\mathbf{y}_{iA(l)} - \mu_{iA(l)})\mathbf{\Sigma}_k^{-1}(\mathbf{y}_{iA(l)} - \mu_{iA(l)})^T \times \mathbf{1}_{c_i=k},$$

where $\mu_{iA(l)} = \mathbf{x}_{iA(l)}\beta_k^T + \mathbf{z}_{iA(l)}\mathbf{b}_i^T$. I generate the replicated completed dataset by first sampling replicate latent class indicators c_i^{rep} and then drawing $\mathbf{y}_{iA(l)}^{rep}$ conditional on c_i^{rep} [Fruhwirth-Schnatter, 2006].

In addition, for randomly selected datasets, I compare plots of the completed data with the replicated completed data to evaluate model fit and the reasonableness of the imputations [Gelman *et al.*, 2005].

3.3 Analysis of Early Childhood Weight and Height Measurements

I apply my proposed model to an illustrative dataset of EHR measurements on weight and height in a sample of US children followed from birth to age 4 years. These EHR measurements were linked to participants in the 1988 National Maternal and Infant Health Survey (NMIHS) and its 1991 Longitudinal Follow-Up, in which low birth weight infants (<2,500 g) were oversampled [Sanderson *et al.*, 1988]. In this dataset, clinic visit times are available in terms of a child's age in months. Clinical recommendation suggests that in early childhood, weight and height measurements should be collected at clinic visits classified as well-child checks [American Academy of Pediatrics, 2018]. The well-child check schedule prescribes clinic visits at age in months 1, 2, 4, 6, 9, 12, 15, 18, 24, 30, 36, and 48. To illustrate my proposed model, I used weight and height measurements from clinic visits classified as check-ups for a random sample of 500 children. I converted weight and height measurements to z-scores using a reference distribution from the Centers for Disease

Control and Prevention [Centers for Disease Control and Prevention, 2019]. Of the 500 children, I excluded one child whose available measurements were flagged as biologically implausible values. The patterns of missing values for the visit process and the response processes given a clinic visit for weight and height are shown in Figure E.1. Overall, of 5,988 well-child windows, 67% correspond to missed visits. Among 1,966 observed clinic visits, only 17 weight measurements are missing ($< 1\%$), whereas 207 height measurements are missing, corresponding to approximately 10%.

I compare three estimation methods, which I label as **MNAR**, **MAR**, and **Naïve**. For the **MNAR** method, I illustrate my proposed model: Assuming that the missing data mechanisms for the visit process and the response process for height are MNAR, I model them jointly with weight and height z-scores. On the other hand, since weight z-scores are rarely missing, I assume the response process for weight is MAR. For the **MAR** method, I assume each of the missing data mechanisms is ignorable. For the **Naïve** method, I fit the complete-data model using only time windows in which both weight and height z-scores are observed, herein "complete pairs". Whereas the **MNAR** and **MAR** methods are based on all 499 children, the **Naïve** method uses only 471 children who have at least one time window with a complete pair.

I consider models with $K = 1, 2, 3$ latent classes. I separately select the optimal number of latent classes for the **MNAR** and **MAR** methods that use all 499 children. Using the **MNAR** method, I fit models with $K = 2, 3$ latent classes, while in the **MAR** method, I assume $K = 1, 2, 3$, where the 1-class model is a multivariate normal model. For model selection, I do not include risk factors in $\mathbf{w}_i$ (3.2) to predict latent class membership. Given the selected number of latent classes for the **MAR** and **MNAR** methods, for a sensitivity analysis, I compare the **Naïve**, **MAR**, and **MNAR** methods based on a model

that includes a child's race, sex, and birth weight in $\mathbf{w}_i$. I model longitudinal trajectories as a cubic polynomial function of a child's age in months, and the patient-specific random effects are specified with a random intercept.

I ran the Gibbs sampler for 20,000 MCMC iterations discarding the first 10,000 as burn-in. To assess convergence, I calculated the Gelman-Rubin convergence diagnostic [Gelman *et al.*, 2014] based on three chains from dispersed initial values. The diagnostic indicated model convergence with values near 1 for all parameters. Trace plots did not show evidence of the label switching problem [Fruhwirth-Schnatter, 2006] that can occur in finite mixture modeling applications. I re-ordered MCMC samples so that latent classes are labeled in order of decreasing health status. For example, latent class 1 always represents the "healthy" trajectory, while the last latent class is for the "unhealthy" trajectory.

3.3.1 Model selection for the MNAR and MAR methods

Table E.1 presents the model information criteria using the **MAR** and **MNAR** methods. For the **MAR** method, the BIC and the DIC3 each chose the 2-class model. In contrast, using the **MNAR** method, the 3-class model was selected by both model information criteria. For $K = 2$ and $K = 3$ using the **MAR** and **MNAR** methods, respectively, the LCIDs show that the posterior distributions of the intercepts are narrow relative to the prior distributions (Figures E.2 and E.3). I therefore assess sensitivity of statistical inferences under the **Naïve**, **MAR**, and **MNAR** methods based on $K = 2, 3$ latent classes.

3.3.2 Sensitivity analysis for the 2 and 3-latent class models

Assuming 2 latent classes, I compared the latent class-specific trajectories of weight and height z-scores, the visit process, and the response process for height, and child latent class

assignment using the **Naïve**, **MAR**, and **MNAR** estimation methods. I conducted the same analysis for the 3-latent class models. Here, I describe the latent class-specific trajectories, and I explicate why some children were classified differently among the methods.

3.3.2.1 2-latent class models.

The **Naïve**, **MAR**, and **MNAR** methods each detected a Normal trajectory subgroup (purple) and a Low trajectory subgroup (orange) (Figure 3.1). Despite similar trajectory patterns across methods, the latent classes appear better separated in the **MNAR** method, particularly for height z-scores for which the response process was modeled. Based on the **MNAR** method, Figure 3.2 presents the latent class-specific visit process and response process for height. Compared to the Low subgroup, the Normal subgroup exhibits a higher probability of a clinic visit, except at the study end. Whereas in the Normal subgroup, the probability of a height response is invariably near 1, in the Low subgroup, the response process climbs sharply from probability below 0.75 at the start of follow-up. Risk factors associated with the probability of latent class membership are presented in Figure E.4.

I assigned children to the Normal or Low subgroup by the maximum of their mean posterior probabilities of belonging to each latent class (columns for $K = 2$, Table E.2). While the **Naïve** and **MAR** methods similarly placed about 67% and 33% of children in the Normal and Low subgroups, respectively, the **MNAR** method assigned 59% of children to the Normal subgroup, and 41% of children to the Low subgroup. Table E.3 cross-classifies the 499 children used in the **MAR** and **MNAR** methods according to posterior latent class assignment, and a covariate from the latent class membership model, birth weight. Since few children born low birth weight (LBW) were classified differently by the **MAR** and **MNAR** methods, I focus on children born non-LBW. Fifty non-LBW children were

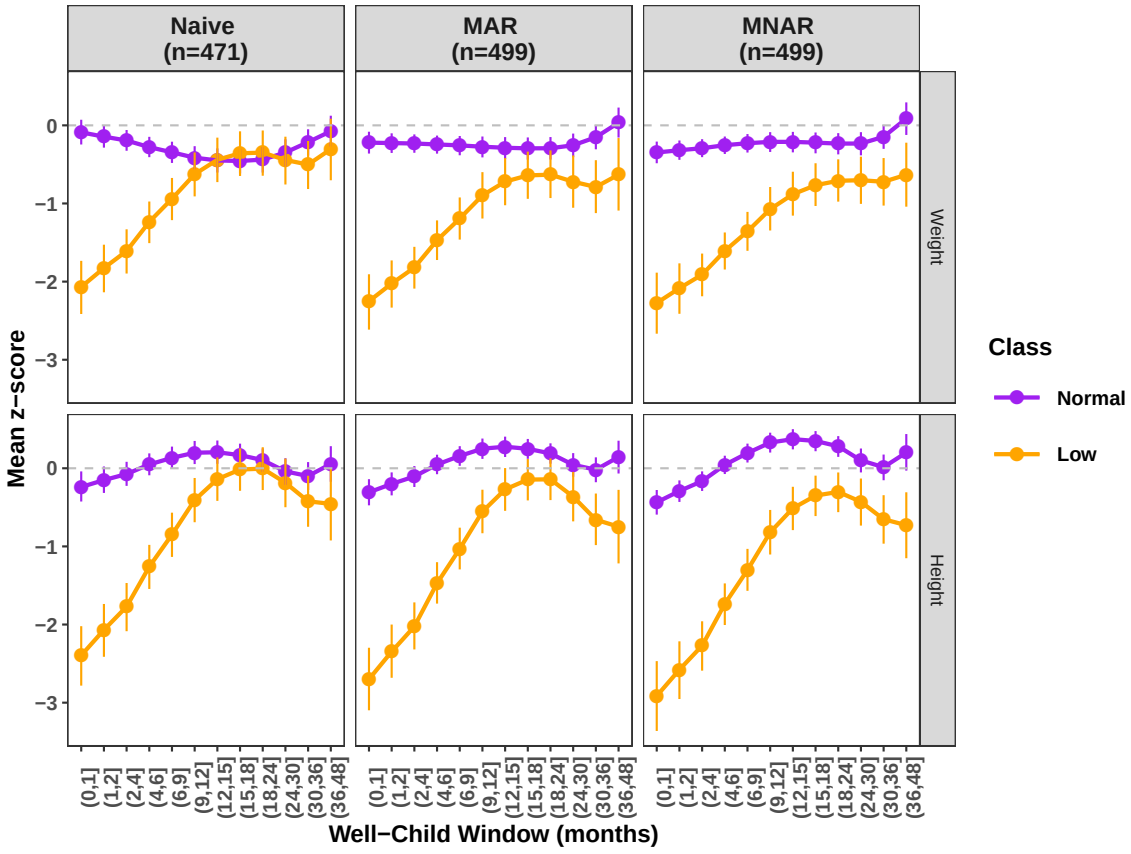

Figure 3.1: Latent class-specific average trajectories of weight and height z-scores estimated by the **Naïve**, **MAR**, and **MNAR** methods, assuming 2 latent classes. n refers to the number of children included by each method.

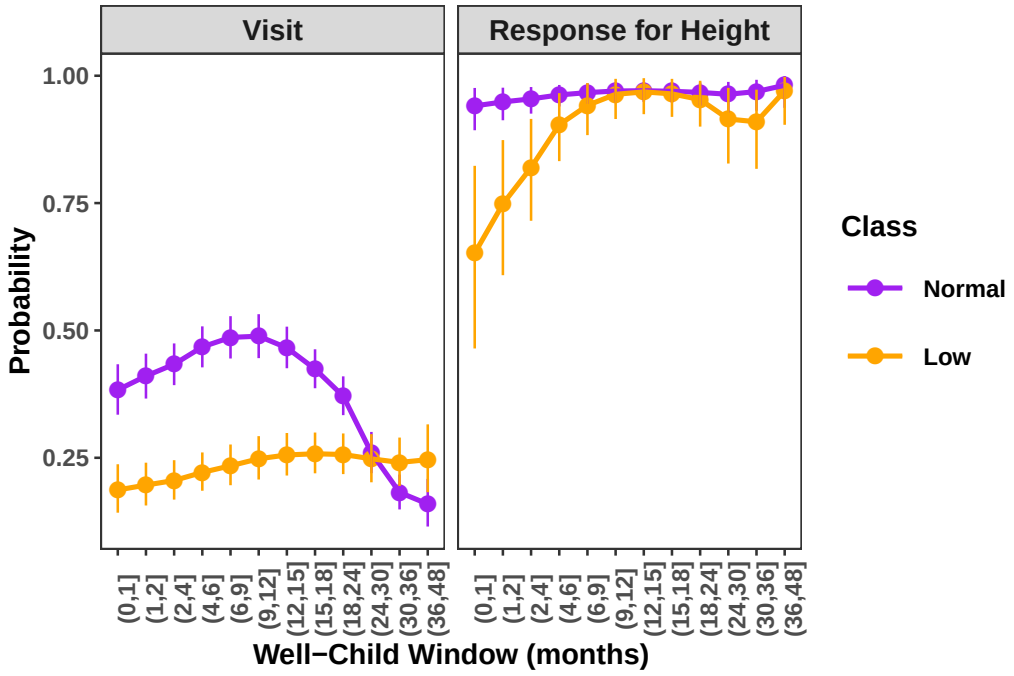

Figure 3.2: Latent class-specific trajectories of the probability of a clinic visit and the probability of a response for height z-scores using the **MNAR** method, assuming 2 latent classes.

placed in the Normal subgroup by the **MAR** method and the Low subgroup by the **MNAR** method. Based on the **MNAR** method, Figure 3.3 shows the sample means among the 50 children using their observed weight and height z-scores, overlaid on the average latent class-specific z-score trajectories. Larger points indicate sample means with more observed measurements. Sample means with more measurements appear in later follow-up when the latent class-specific trajectories are similar. In fact, especially for height z-scores, the 50 children have few observed measurements in early follow-up when the class trajectories are easily distinguished. Figure 3.4 shows the patterns of the proportions of observed visits and observed height responses in each time window among the 50 children, overlaid by the latent class-specific visit and response trajectories. Consistent with the **MNAR** method classifying the children in the Low subgroup, both the observed visit and response patterns resemble the corresponding Low trajectories.

Table E.3 also indicates that 18 non-LBW children were placed in the Low subgroup by the **MAR** method and the Normal subgroup by the **MNAR** method. In contrast to the scenario of the 50 children, these 18 children have more weight and height z-score measurements in early follow-up, when the sample means align to some extent with the Low trajectory (Figure E.5). However, the patterns of the proportions of observed visits and observed height responses among the 18 children correspond to the visit and response process trajectories in the Normal subgroup (Figure E.6), which is again consistent with the MNAR classification.

The comparison of the **Naïve** and **MNAR** methods for the 471 common children reveals patterns of classification similar to those heretofore explicated for the **MAR** and **MNAR** methods (data not shown).

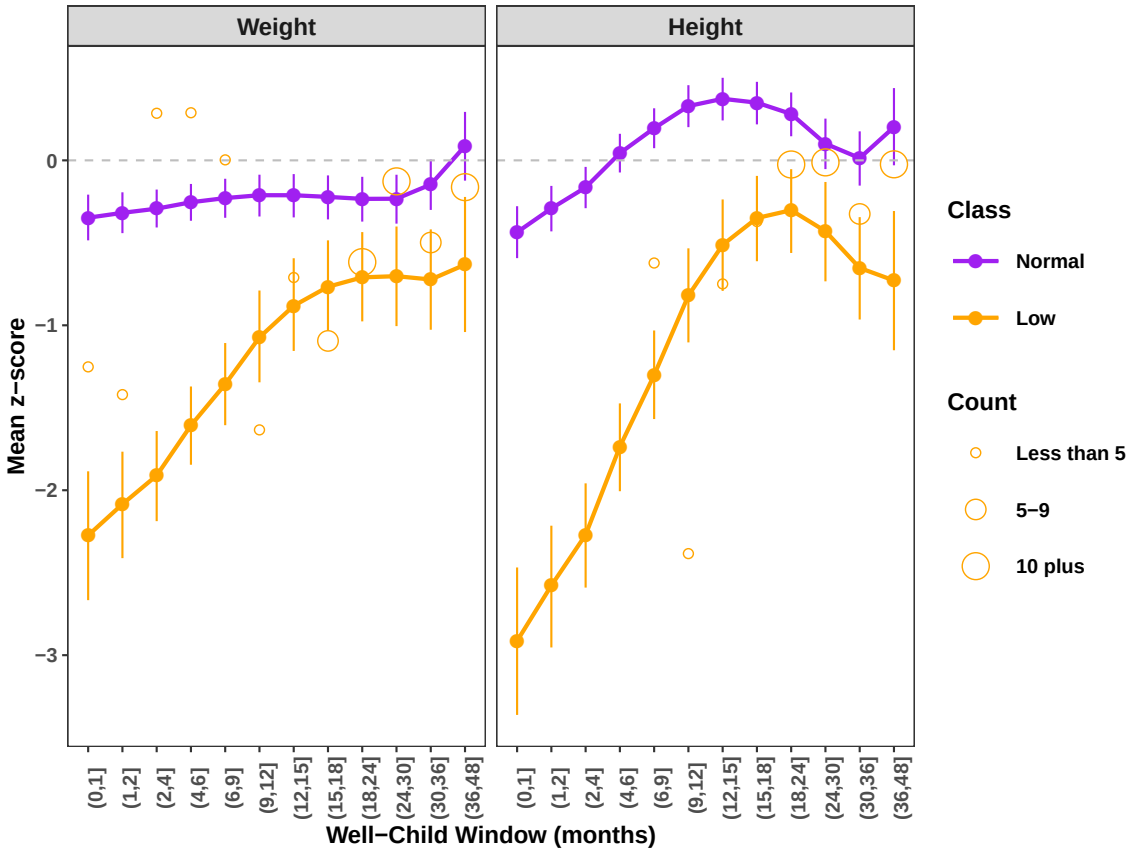

Figure 3.3: Sample means of observed weight and height z-scores (hollow circles) in each well-child window among the 50 non-low birth weight children moved from the Normal trajectory subgroup in the **MAR** method to the Low trajectory subgroup in the **MNAR** method, assuming 2 latent classes. The size of the point indicates the number of observations contributing to the sample mean. Overlaid are the average latent class-specific z-score trajectories estimated by the **MNAR** method.

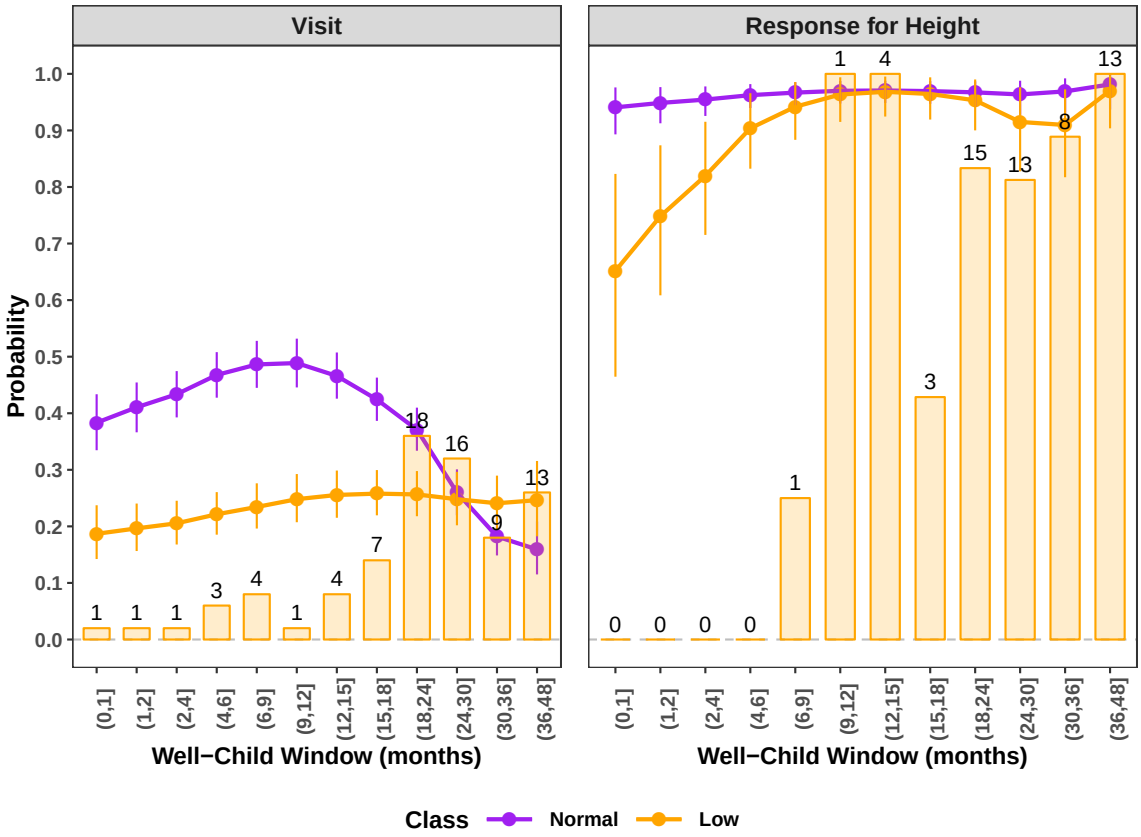

Figure 3.4: Bar plots of the observed proportions of children with a clinic visit, and the observed proportions of children with a height response, among the 50 non-low birth weight children moved from the Normal trajectory subgroup in the **MAR** method to the Low trajectory subgroup in the **MNAR** method. In the Visit panel, the number of children with a clinic visit in each window is provided. In the Response for Height panel, the number of children with a height response (given a clinic visit) is given. Overlaid are the latent class-specific visit and response trajectories estimated by the **MNAR** method assuming 2 latent classes.

3.3.2.2 3-latent class models.

In Figure 3.5, the **Naïve**, **MAR**, and **MNAR** methods each identified a Normal, increasing (purple); Normal, decreasing (orange); and Low (blue) subgroup. The latent class-specific average trajectories for weight and height z-scores appear similar across methods. In Figure 3.6, the visit process of the Normal, increasing subgroup decreases over follow-up, whereas for the Normal, decreasing subgroup, the probability of a clinic visit rises quickly until about 12 months before decreasing. The probability of a response for height is indistinguishable for these two subgroups. The Low subgroup exhibits a visit and response process similar to the Low subgroup in the 2-class model. Figure E.7 presents risk factors from the latent class membership model.

Using the **MAR** method, the percentage of children in the Normal, increasing; Normal, decreasing; and Low subgroups is 38, 37, and 24 (columns for $K = 3$, Table E.2). Posterior latent class assignment under **Naïve** method is comparable. In contrast, the **MNAR** method placed approximately one-third of children in each subgroup. For the **MAR** and **MNAR** methods in the 3-class analysis, Table E.4 presents the cross-classification of the 499 children according to their posterior latent class assignment and LBW status. To illustrate the patterns of classification between the **MAR** and **MNAR** methods, I focus on the two cells with the largest number of children.

Thirty children were placed in the Normal, increasing subgroup by the **MAR** method and the Normal, decreasing subgroup by the **MNAR** method. With measurements available over most of follow-up, the sample means of observed weight and height z-scores among the 30 children largely align with the Normal, increasing trajectory (Figure 3.7), which is consistent with the **MAR** classification. As evidenced by Figure 3.8, the **MNAR** method

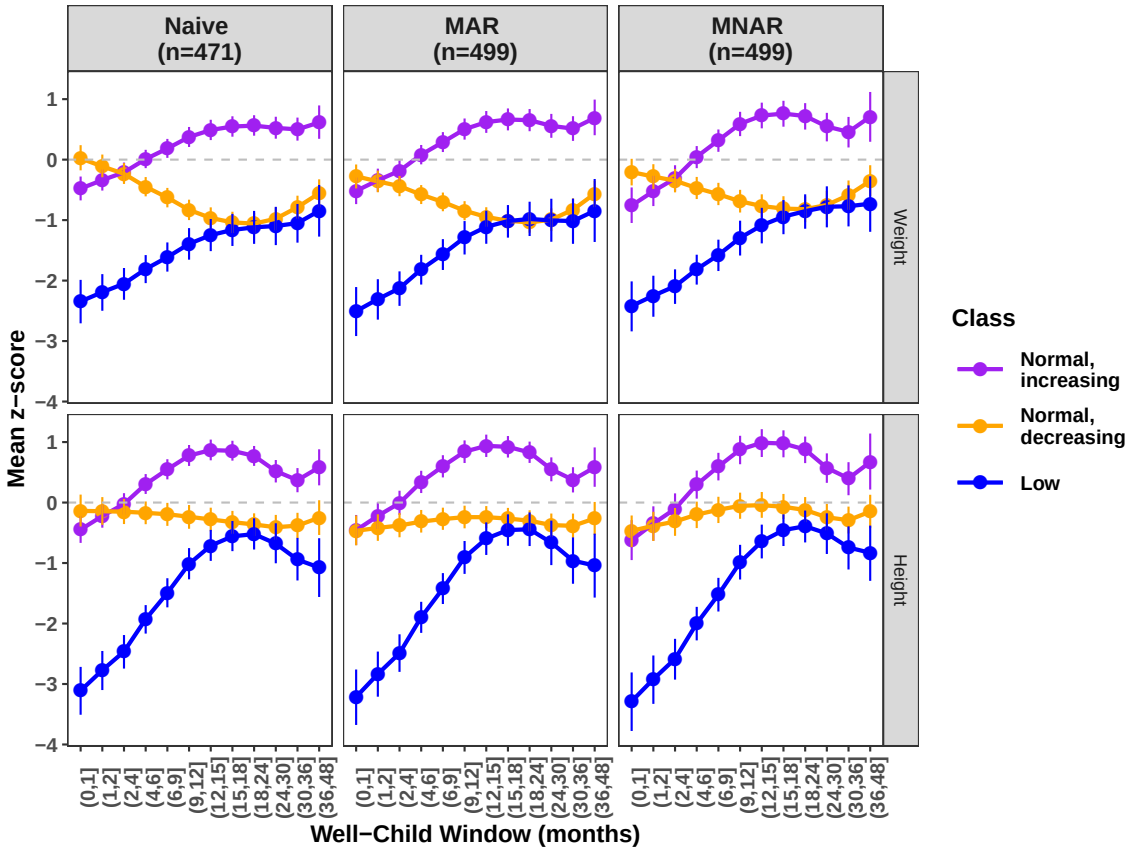

Figure 3.5: Latent class-specific average trajectories of weight and height z-scores estimated by the **Naïve**, **MAR**, and **MNAR** methods, assuming 3 latent classes. n refers to the number of children included in each analysis.

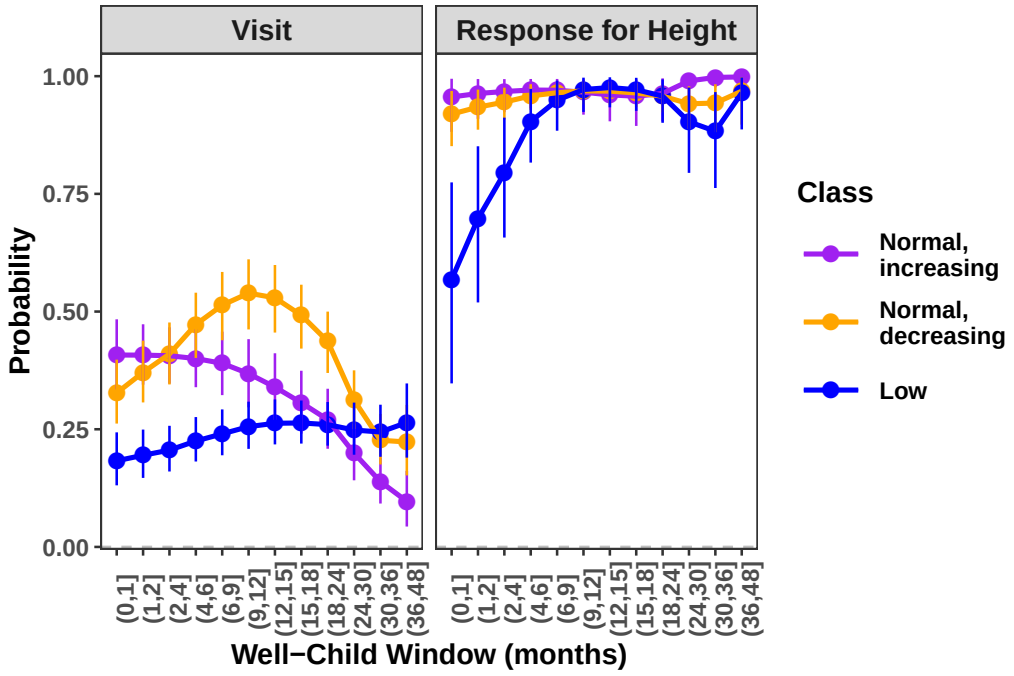

Figure 3.6: Latent class-specific trajectories of the probability of a clinic visit and the probability of a response for height z-scores in the **MNAR** method, assuming 3 latent classes.

classified these children in the Normal, decreasing subgroup on the basis of their pattern of proportion of observed visits.

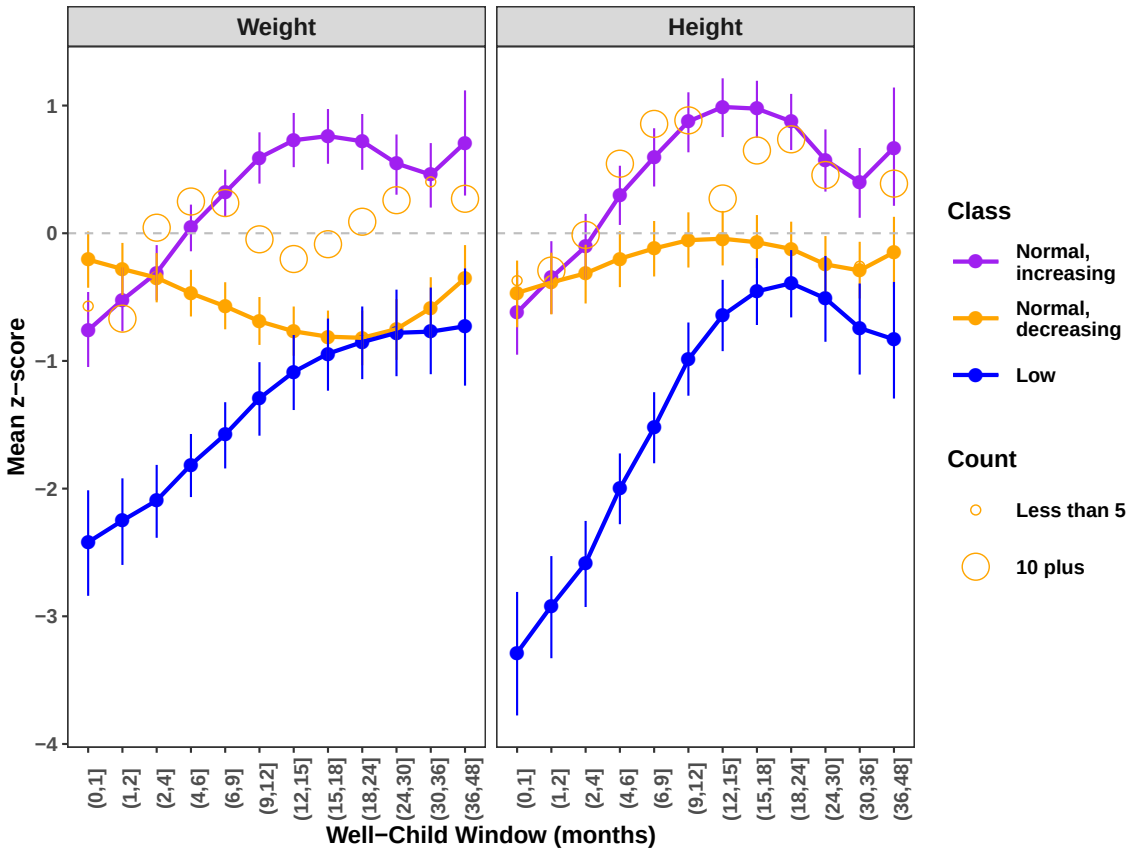

Figure 3.7: Sample means of observed weight and height z-scores (hollow circles) in each well-child window among the 30 non-low birth weight children moved from the Normal, increasing trajectory subgroup in the **MAR** method to the Normal, decreasing trajectory subgroup in the **MNAR** method, assuming 3 latent classes. The size of the point indicates the number of observations contributing to the sample mean. Overlaid are the average latent class-specific z-score trajectories estimated by the **MNAR** method.

For the second largest cell in Table E.4, 26 children were placed in the Normal, decreasing subgroup by the **MAR** method and the Low subgroup by the **MNAR** method. Corresponding to the scenario in Figure 3.3 from the 2-class sensitivity analysis, sample

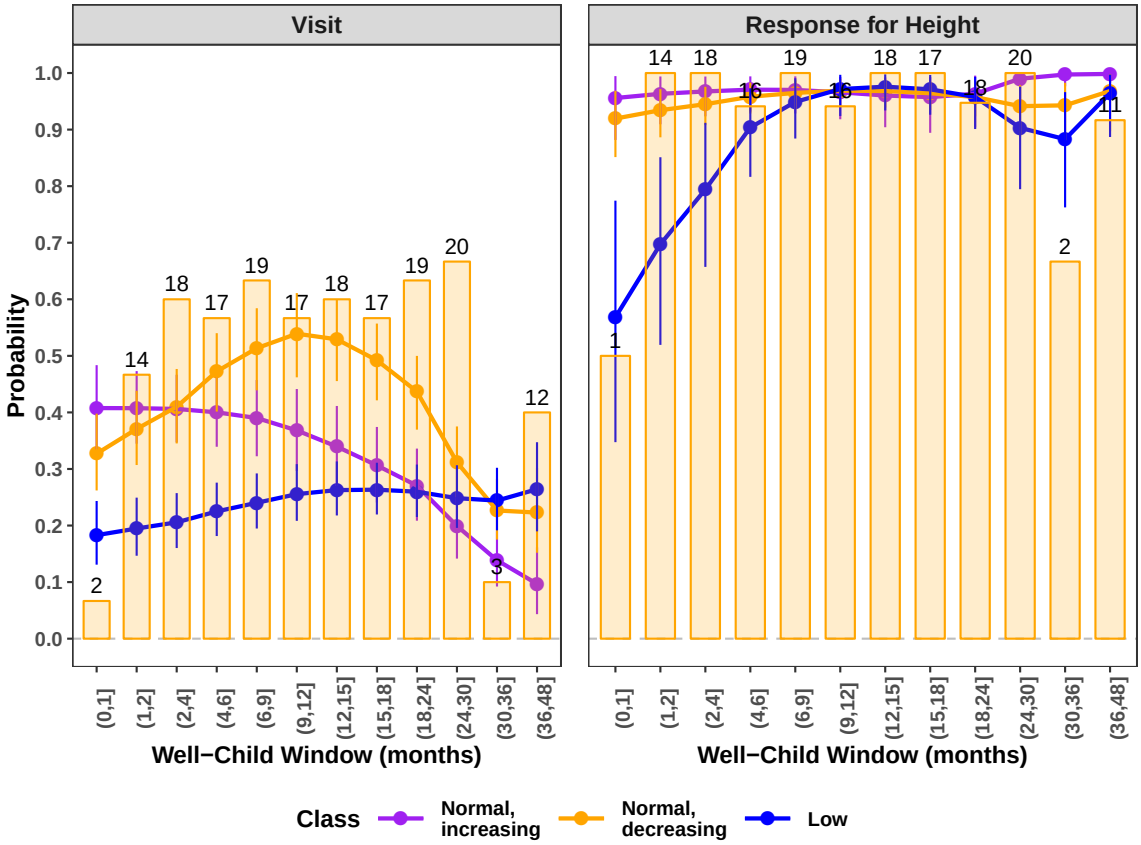

Figure 3.8: Bar plots of the observed proportions of children with a clinic visit, and the observed proportions of children with a height response, among the 30 non-low birth weight children moved from the Normal, increasing trajectory subgroup in the **MAR** method to the Normal, decreasing trajectory subgroup in the **MNAR** method. In the Visit panel, the number of children with a clinic visit in each window is provided. In the Response for Height panel, the number of children with a height response (given a clinic visit) is given. Overlaid are the latent class-specific visit and response trajectories estimated by the **MNAR** method assuming 3 latent classes.

means of observed weight and height z-scores based on more measurements appear in later follow-up when the Normal, decreasing and Low trajectories are similar, with very few measurements in early follow-up when the trajectories are different (Figure E.8). The **MNAR** method placed the 26 children in the Low subgroup because their patterns of observed proportions of visits and height responses resemble the Low visit and response trajectories (Figure E.9).

As in the 2-latent class sensitivity analysis, the analogous comparison based on 3-latent class models of the **Naïve** versus **MNAR** methods showed findings similar to those from comparing the **MAR** and **MNAR** methods (data not shown).

3.3.3 Model checking

Based on the sensitivity analysis of the 2 and 3-latent class models using the **Naïve**, **MAR**, and **MNAR** methods, I chose to conduct model checking for the 2-latent class model using the **MNAR** method. Figure E.10 presents a scatter plot of the replicated completed versus completed discrepancy measure T across MCMC samples. The Bayesian predictive p-value of 0.45 represents the proportion of samples above the diagonal, suggesting adequate overall model fit. For a randomly selected dataset, Figures E.11 and E.12 show histograms of completed weight and height z-scores overlaid by replicated completed weight and height z-scores by subgroup and well-child window. The model appears to fit the data well.

3.4 Simulation Study

3.4.1 Design

I conducted a simulation study to examine the effect of estimation method on estimating the regression coefficients from the longitudinal health outcomes model, β_{rk}, in (3.4); estimating the latent class-level weights calculated as $\pi_k = \frac{1}{n}\sum_{i=1}^{n}\pi_{ik}$ from (3.1); and, predicting a subject's true latent class from p_{ik} in (3.10). I designed the study based on the real data analysis with 2 latent classes estimated with the **MNAR** method. For 500 subjects, I generated longitudinal outcomes of interest y_{1ij} and y_{2ij} over 12 time windows, with about 60% and 40% of subjects in latent classes 1 and 2, respectively. I assumed the missing data mechanisms for the visit process and response process for y_{2ij} are MNAR, while y_{1ij} is fully observed given a clinic visit. In this setting, I considered four specific scenarios, which I describe briefly below, with details in the Appendix F.

In scenario 1 (S1), I mimicked the latent class-specific trajectories and missingness proportions in the real data analysis. I selected true parameter values for β_{rk} (3.4), ϕ_k (3.6), and λ_{2k} (3.8) in the models for the longitudinal health outcomes, visit process, and response process for y_{2ij}, respectively, to linearly represent the estimated trajectories in the 2-latent class model using the **MNAR** method. As in the real data analysis, in latent class 1, the percents of missed clinic visits and missed y_{2ij} responses are 55% and 10%, respectively. The corresponding values in latent class 2 are 70% and 20%. Figure 3.9 depicts S1 for y_{2ij}, in which the latent class-specific average trajectories are overlaid by points for observed measurements. In early follow-up when the class trajectories are better separated, missingness in y_{2ij} is high in latent class 2.

In S2 and S3, I examined whether the effect of estimation method varies by how different

the slopes are for the latent class-specific trajectories of y_{2ij}. In S2, I increase the difference by making the slope in latent class 2 steeper. In S3, I change the slope in latent class 2 to be nearly parallel to that in latent class 1. No other aspects of S1 were modified.

In S4 and S5, I modified S1 to examine whether the effect of estimation method varies by the extent of visit and response process missingness whilst maintaining the shapes of the class trajectories. In S4, I reduced the percent of missed clinic visits in latent classes 1 and 2 from 55% to 35%, and from 70% to 55%, respectively. In S5, I modified S1 by increasing the percent of missing y_{2ij} responses from 10% to 25% in latent class 1, and from 20% to 35% in latent class 2.

For each scenario, I compare estimation of the 2-latent class model using the **MNAR** method (under which the data are generated) to the **MAR** method and **Naïve** method – as in the real data analysis. In addition, for the benchmark, I include the **Full** method, in which the complete-data model is fit to the full data before introducing any missed visits or missed responses. I ran 500 data simulations. For parameter estimation of β_{rk}, I examined the performance measures including bias, mean squared error (MSE), 95% coverage probability, and the average length of the 95% credible interval. For the latent class-level weights, I compare the true weight to the average weight over the 500 simulations. For subject classification, I considered summary statistics of the proportion of misclassified subjects in each simulation. In the main text, I present the simulation results of S1 and summarize those from S2-S5, with the full details in Appendix F.

3.4.2 Results

Table 3.1 presents the simulation results from S1. As expected, estimation under the **Full** method presents the benchmark. The **MNAR** method generally shows better performance

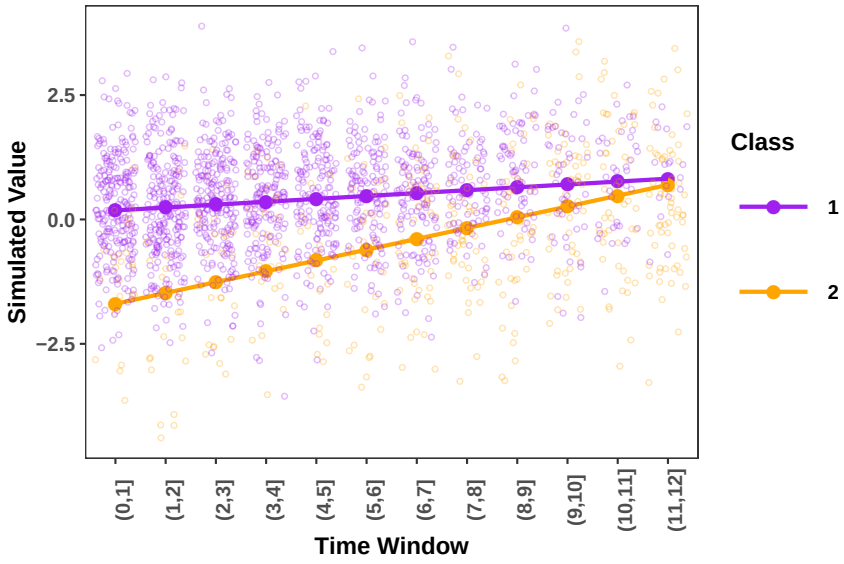

Figure 3.9: Average latent class-specific trajectories for y_{2ij} overlaid by points for observed measurements, under S1.

on all measures than the **Naïve** and **MAR** methods. Using the **Naïve** and **MAR** methods, parameter estimation in latent class 2 of the intercepts and slopes, β_{r21} and β_{r22}, respectively, is poor, particularly for β_{r22}. The positive bias in the intercepts and negative bias in the slopes suggest that poor estimation is driven by subjects from latent class 1 incorrectly classified into 2. On average, however, estimation of the latent class-level weights π_k reveals that more subjects from class 2 are misclassified into 1. In Table 3.2, under S1, the median subject misclassification rate for the **Naïve** and **MAR** methods is 0.15, while the distributional summaries for the **MNAR** and **Full** methods are similar.

Under S2 in which the latent class-specific slopes for y_{2ij} are more different (Table F.1), the **Full** method remained the benchmark, with the **MNAR** method outperforming the **Naïve** and **MAR** methods. However, compared to S1, performance using the **Naïve** and **MAR** methods improved: For y_{1ij}, estimation of the latent class-specific intercepts β_{1k1} and slopes β_{1k2} appears satisfactory, with the exception of β_{122} under the **Naïve** method.

For y_{2ij}, the intercept and slope in latent class 2, β_{221} and β_{222}, respectively, are biased in the same direction as in S1, but the magnitude of the bias is smaller. This suggests that the better performance of **Naïve** and **MAR** methods may be driven by fewer subjects from latent class 1 being misclassified into 2 – which is consistent with the worse estimation of π_k. Table F.2 shows that using the **Naïve** and **MAR** methods, the median subject misclassification rate decreased slightly in S2 relative to S1.

The performance of the **Full** and **MNAR** methods was robust to S3, in which the latent class-specific slopes for y_{2ij} are nearly parallel (Table F.3). However, estimation using the **Naïve** and **MAR** methods is worse than in S1. For y_{1ij}, in latent class 2, I observe positive bias in the intercept β_{121} and negative bias in the slope β_{122}, as in S1. In addition, however, in latent class 1, I observe positive bias in the intercept β_{111}, which is likely driven by the extent of subjects from latent class 1 with relatively low y_{1ij} values misclassified into class 2. Estimation of parameters for y_{2ij} reveals a similar phenomenon. Interestingly, estimation of the latent class-level weights has improved – suggesting comparable levels of misclassification between classes 1 and 2. Summary statistics of the subject misclassification rate are similar to those from S1 (Table F.4).

Table 3.1: Simulation results of S1 for parameter estimation of intercept β_{rk1} and slope β_{rk2} for longitudinal outcome r in latent class k, and latent class-level weights π_k under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

Parameter	Method	Truth	Mean	Bias	MSE	Coverage	Length
β_{111}	Full	-0.250	-0.252	-0.002	0.002	0.950	0.190
	Naïve		-0.221	0.029	0.005	0.904	0.224
	MAR		-0.231	0.019	0.004	0.908	0.219
	MNAR		-0.248	0.002	0.003	0.942	0.209
β_{121}	Full	-1.000	-1.000	0.000	0.003	0.956	0.230
	Naïve		-0.954	0.046	0.016	0.878	0.404
	MAR		-0.996	0.004	0.011	0.932	0.370
	MNAR		-0.995	0.005	0.007	0.936	0.312
β_{112}	Full	0.100	0.100	-0.000	0.000	0.930	0.048
	Naïve		0.089	-0.011	0.001	0.928	0.099
	MAR		0.092	-0.008	0.001	0.926	0.094
	MNAR		0.100	-0.000	0.001	0.954	0.091
β_{122}	Full	0.500	0.501	0.001	0.001	0.930	0.096
	Naïve		0.411	-0.089	0.013	0.720	0.266
	MAR		0.459	-0.041	0.007	0.850	0.238
	MNAR		0.499	-0.001	0.003	0.948	0.215
β_{211}	Full	0.500	0.500	-0.000	0.002	0.954	0.189
	Naïve		0.545	0.045	0.006	0.858	0.224
	MAR		0.536	0.036	0.005	0.886	0.221
	MNAR		0.505	0.005	0.003	0.938	0.210
β_{221}	Full	-0.500	-0.503	-0.003	0.003	0.940	0.196
	Naïve		-0.452	0.048	0.015	0.896	0.379
	MAR		-0.474	0.026	0.011	0.922	0.366
	MNAR		-0.500	0.000	0.007	0.956	0.310
β_{212}	Full	0.200	0.199	-0.001	0.000	0.918	0.048
	Naïve		0.185	-0.015	0.001	0.904	0.098
	MAR		0.186	-0.014	0.001	0.880	0.096
	MNAR		0.200	-0.000	0.001	0.950	0.093
β_{222}	Full	0.750	0.751	0.001	0.001	0.934	0.097
	Naïve		0.648	-0.102	0.017	0.646	0.270
	MAR		0.675	-0.075	0.012	0.738	0.262
	MNAR		0.747	-0.003	0.004	0.944	0.237
π_1	Full	0.557	0.557				
	Naïve	0.576	0.617				
	MAR	0.577	0.609				
	MNAR	0.576	0.575				

Table 3.2: Simulation results of S1 for subject misclassification under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

		Percentile			
Method	Min	25	50	75	Max
Full	0.00	0.01	0.02	0.02	0.04
Naïve	0.09	0.14	0.15	0.16	0.20
MAR	0.09	0.13	0.14	0.16	0.20
MNAR	0.01	0.03	0.03	0.04	0.06

Compared to S1, in S4, when I reduce the percent missed clinic visits in latent classes 1 and 2 to 35% and 55% respectively, estimation of β_{rk} and subject misclassification (Tables F.5 and F.6) using the **Naïve** and **MAR** methods improves. The slopes in latent class 2, β_{r22}, however, still show negative bias. The **MNAR** method often presents an efficiency gain. Conversely, in S5, with increased missed y_{2ij} responses, estimation of β_{rk} and subject misclassification using the **Naïve** and **MAR** methods worsens relative to S1 (Tables F.7 and F.8).

3.5 Discussion

In this study, I developed a Bayesian shared parameter model for multiple longitudinal health outcomes in EHRs to account for an MNAR visit process and response process given a clinic visit. My model targets longitudinal health outcomes collected according to a clinically prescribed visit schedule. By exploiting heterogeneity in EHR patient populations, I built a shared parameter model with a discrete latent class variable. Conducive to handling large numbers of missing values in EHRs, my model tractably summarizes missingness patterns into a pre-specified number of latent classes. My shared parameter model can be easily altered to conduct sensitivity analysis about missing data assumptions. I developed a user-friendly `R` package `EHRMiss` that can be used for model fitting, selection, and checking.

My study complements recent work on large clinical databases by [McCulloch *et al.*, 2016], who use a traditional shared parameter model in which patient-specific random effects link the longitudinal health outcomes and visit process. In contrast to my approach, [McCulloch *et al.*, 2016] define the visit process as a binary indicator for whether a response was observed, which corresponds to my definition of the response process given a clinic visit. Notwithstanding, the authors show analytically and via simulations that in the absence of accounting for an informative visit process, estimators of regression coefficients associated with the random effects can be badly biased. Using a discrete latent class variable to link the longitudinal health outcomes, visit process, response process given a clinic visit, I show empirically that failure to account for a nonignorable visit process and response process given a clinic visit may result in misleading statistical inferences. Estimated average latent class-specific health trajectories may be biased depending on whether the latent classes are well-identified, in addition to the shape of the class trajectories. Even when estimated class

trajectories are unbiased, the latent class-level weights may be poorly estimated, precluding unbiased population-averaged inferences. Finally, subject misclassification is sensitive to treatment of missing data.

In my data application, I found that given a selected number of latent classes, sensitivity analysis under different estimation methods is critical, particularly if the clinical interpretation of latent classes is of scientific interest. Comparing 2-latent class models estimated with the **MAR** and **MNAR** methods, I learned that the **MNAR** method used the visit process and response process for height z-scores given a clinic visit to reclassify children between the Low and Normal subgroups when few observed measurements were available. On the other hand, the corresponding sensitivity analysis based on the 3-latent class models showed that the **MNAR** method could reclassify children primarily on the basis of their observed visit process – contrary to the subgroup suggested by their observed weight and height z-scores. Carefully examining classification under different missing data assumptions can help ensure the interpretation of the latent classes is consistent with the scientific investigation.

I am primarily interested in two areas for future research. In developing the proposed model, I was motivated by longitudinal health outcomes with a clinically prescribed visit schedule, which I used to discretize time into observation windows during which to measure the visit process and response process given a clinic visit. However, when a prescribed visit schedule is unavailable, measuring the visit process in continuous time is consistent with the data generation in EHRs, since a patient can show up for a clinic visit at any time. I am currently modifying the proposed model for the continuous time setting. Second, Bayesian methods can be especially time intensive as the number of observations grows. To enhance the practicality of my proposed method for EHR-based research, I am interested in pursuing strategies for scaling MCMC algorithms to large datasets.

As EHRs are increasingly used in applied biomedical research, the use of statistical methods that account for the features of data generation process will heighten the credibility of the scientific findings. My proposed Bayesian shared parameter model exploits heterogeneity in EHR patient populations to account for an MNAR visit process and response process given a clinic visit.

Chapter 4

Software

I developed two R software packages to fit each of the proposed models in Chapters 2 and 3. In this Chapter, I explicate how to use each of the packages.

4.1 R Software Package Bsvygmm

The R package **Bsvygmm** can be used to fit the proposed Bayesian GMM for complex survey data in Chapter 2. In addition, the package can be used for model selection with three model information criteria, and model checking using Bayesian posterior predictive p-values. The proposed Bayesian GMM is fit using the `Bsvygmm` function. To predict each subject's latent class membership, `Bsvygmm` can model three different types of a cluster sample design, including:

1. correlations among subjects within the same area segment, referred to as "Unstr" in the package;

2. spatial correlations among neighboring area segments only, called "Str" in the package; and,

3. both types of correlations, called "Both" in the package.

In addition, in the multinomial model of latent class membership, `Bsvygmm` includes an option to use B-splines for flexibly modeling the relationship between one of the variables and the probability of belonging to a latent class. In Chapter 2, I used this option for the size variable used in probability proportional to size sampling.

I describe the functionality of **Bsvygmm**. Details on the package functions are accessible with the `R` help pages. For example, by typing `?Bsvygmm`, extensive information is provided on the `Bsvygmm` function. The **Bsvygmm** package contains an artificial dataset, called `data`. Using the package's `simdat` function, I generated the data from a 2-latent class model. In the latent class membership model, I used independent random effects to account for correlations among subjects in the same area segment ("Unstr"); and spatial random effects to account for correlations among neighboring area segments ("Str"). The dataset contains 600 subjects each of whom has 3 measurements. There are 50 clusters, each of which contains 12 subjects. There are 10 strata, each of which contains 5 clusters, and therefore 60 subjects. A preview of the data is

```
library(Bsvygmm)

data(data)

head(data, n = 3)
```

```
##   subjectID        Y time clusterID stratumID         x1 C
## 1         1 7.731119    1         1         1 -0.9258426 1
## 2         1 5.826467    2         1         1 -0.9258426 1
## 3         1 5.368823    3         1         1 -0.9258426 1
```

where *subjectID* is an integer-valued subject identifier for each longitudinal measurement; Y are longitudinal measurements; *time* is a categorical variable indicating interview wave; *clusterID* provides the integer-valued cluster identifier; *stratumID* provides the integer-valued stratum identifier; *x1* is a subject-level covariate generated from the standard normal distribution; and C is a discrete latent variable for each subject's latent class membership.

In addition, the package has a stored adjacency matrix that is used in modeling the spatial correlations among the area segments in the latent class membership model. Adjacency matrices can be easily created in `R` with the `readOGR` function in the **rgdal** package. A preview of the adjacency matrix is

```
data(ADJ)

ADJ[1:4, 1:4]
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    0    0    0
## [2,]    0    0    0    0
## [3,]    0    0    0    0
## [4,]    0    0    0    0,
```

where area segments 1 to 4 are evidently not neighbors with each other.

4.1.1 Analysis with "Both" types of correlations among area segments

I demonstrate fitting the model that generated the data with the function `Bsvygmm`. The model type is "Both", because both types of area segment correlations are included in the latent class membership model. In this model, I do not use a spline, and I use inverse gamma prior distributions on the hierarchical variances for the random effects and for the

observation-level data variance. A uniform prior distribution is also available by replacing "IG" with "Unif".

```
# Number of assumed latent classes
K <- 2
# Model type
modelType <- "Both"
# Include spline?
spline <- FALSE
# Priors on hierarchical variance of random effects
# and observation-level data variance
hierVar <- list("IG", "IG")
```

Before fitting the model, the design matrices for the latent class membership model, and the fixed and random effects in the model for Y must be specified. The columns in the design matrix for the random effects must be a subset of the columns of the fixed effects design matrix. In addition, the function call requires area segment (called "cluster" in the code chunk) and stratum identifiers at the observation-level and subject-level. For example, the stratum identifier at the observation-level indicates the stratum to which a given longitudinal measurement belongs.

```
# Aggregate to subject-level for the design matrix
# in the latent class membership model
dats <- aggregate(data[ , c("subjectID", "clusterID", "stratumID", "x1")],
by = list(data$subjectID), FUN = tail, n = 1)
W <- model.matrix( ~ x1, data = dats)
```

```
s <- ncol(W)

# In this analysis, model time using dummies

timedf <- data.frame(time = factor(data$time))

# Random effects design matrix

Vr <- data.matrix(dummy::dummy(timedf, int = TRUE))

colnames(Vr) <- c("time1", "time2", "time3")

q <- ncol(Vr)

# Fixed effects design matrix

Vf <- Vr

p <- ncol(Vf)

# Cluster and stratum identifiers at the subject and observation-level

clusterIDSub <- dats$clusterID

stratumIDSub <- dats$stratumID

clusterIDObs <- data$clusterID

stratumIDObs <- data$stratumID
```

The prior distributions and initial values must be specified as list objects in which the order of the elements matters. Note the use of `list(NULL)` when a parameter is not desired: I use `list(NULL)` for the position of the prior and initial values for the B-splines. The user cannot specify different prior distributions by latent class. However, initial values for each latent class are required.

```
# Prior distributions

priors <- list(list(rep(0, s), diag(1, s)),

# Latent class regression coefficients
```

```
list(.1, .1),
# Stratum-level independent random effects in latent class model
list(2, 1),
# Cluster-level spatial random effects in latent class model
list(.1, .1),
# Cluster-level independent random effects in latent class model
list(NULL),
# No spline
list(rep(0, p), diag(10, p)),
# Regression coefficients for Y
list(.1, .1),
# Stratum-level independent random effects for Y
list(.1, .1),
# Cluster-level independent random effects for Y
list((q + 2), diag(0.25, q)),
# Subject-level random effects for Y
list(.1, .1))
# Observation-level variance for Y
# Initial values following the same order as in priors
inits <- list(matrix(rep(0, s * (K - 1)), nrow = s, ncol = (K - 1)),
rep(0.2, K - 1),
rep(0.2, K - 1),
rep(0.2, K - 1),
```

```
NULL,
matrix(rnorm(p * K), nrow = p, ncol = K),
rep(0.1, K),
rep(0.1, K),
array(diag(0.5, q), dim = c(q, q, K)),
rep(1, K))
```

In the call to `Bsvygmm`, I run the MCMC sampler for 1000 iterations with a burn-in of 500. Since I set update = 500 with monitor = TRUE, the iteration number and predicted class size will be printed to the console every 500 iterations. In addition, a graphic with selected trace plots will be updated.

```
# Define subjectID and outcome
subjectID <- data$subjectID
Y <- data$Y
res <- Bsvygmm(K = K, W = W, B = NULL, ADJ = ADJ, Y = Y,
Vr = Vr, Vf = Vf, subjectID = subjectID,
clusterIDObs = clusterIDObs, stratumIDObs = stratumIDObs,
clusterIDSub = clusterIDSub, stratumIDSub = stratumIDSub,
spline = FALSE, modelType = modelType, priors = priors,
hierVar = hierVar, inits = inits, n.samples = 1000,
burn = 500, monitor = TRUE, update = 500,
writeSamples = TRUE)
```

```
## Iteration: 500
```

```
## Class size: 318 282

## Iteration: 1000

## Class size: 318 282
```

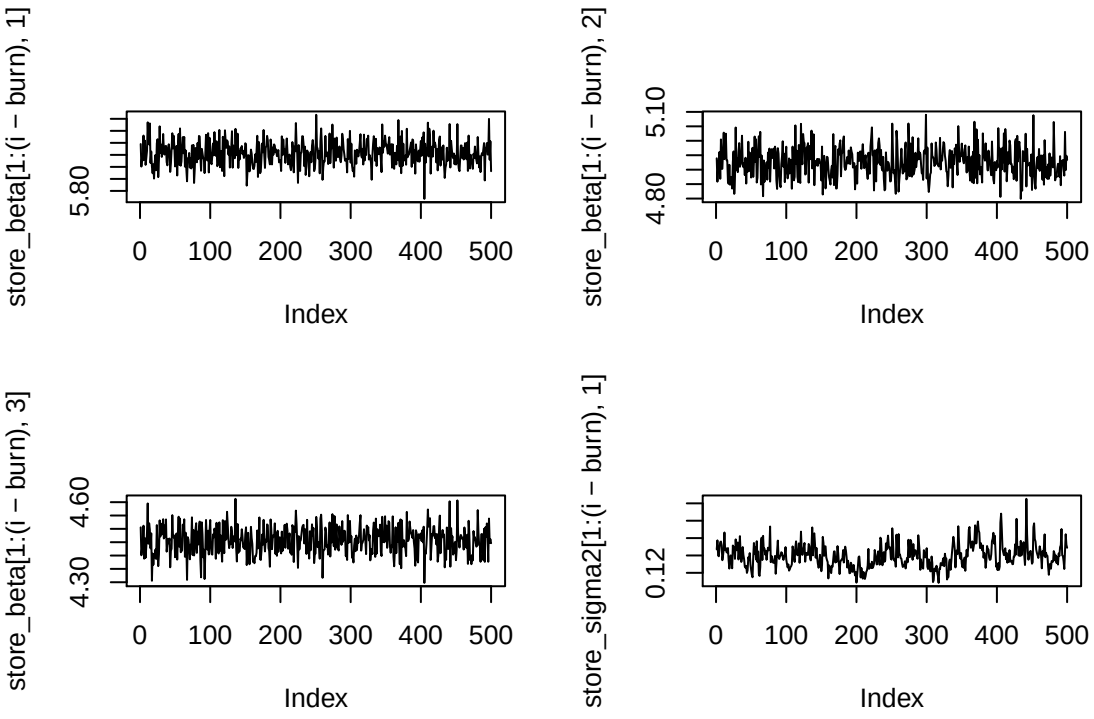

Figure 4.1: Trace plots of the first three regression coefficients in the longitudinal outcomes model, and the observation-level data variance in latent class 1.

```
## Total minutes elapsed: 25.16183 0.5193333 26.606 NA NA
##
## No evidence of label switching problem using Stephen's method
## from the label.switching package
##
## Background information
## Number of subjects: 600
## Number of observations: 1800
## Number of latent classes: 2
## Clusters in study area: 50
## Number of area segments (clusters): 50
##
## Posterior latent class assignment:
##                                          Class 1 Class 2
## Predicted class size                         318     282
## No. subjects with probability at least 0.95  318     281
## No. subjects with probability at least 0.90  318     281
## No. subjects with probability at least 0.80  318     282
## Mean probability                               1       1
## Median probability                             1       1
##
##  Reference class in latent class membership model: 1
## Posterior means and 95% credible intervals:
##                    Post. Mean   2.5 %  97.5%
```

```
## Class2_(Intercept)    -0.0749 -0.4534 0.3310
## Class2_x1              0.1407  0.0272 0.2693
## Class2_gamma2          0.3964  0.0689 1.1646
## Class2_xi2             1.1004  0.2266 2.8329
## Class2_tau2            0.6965  0.2060 1.4544
## Class1_time1           5.9567  5.8601 6.0602
## Class1_time2           4.9267  4.8321 5.0402
## Class1_time3           4.4560  4.3600 4.5514
## Class2_time1           7.1676  7.0751 7.2542
## Class2_time2           8.1733  8.0879 8.2545
## Class2_time3           9.2170  9.1294 9.3031
## Class1_psi2            0.5010  0.1832 1.1061
## Class2_psi2            0.1130  0.0314 0.2798
## Class1_omega2          0.2810  0.1783 0.4339
## Class2_omega2          0.2319  0.1427 0.3536
## Class1_phi11           0.0274  0.0160 0.0532
## Class1_phi21          -0.0006 -0.0175 0.0159
## Class1_phi31           0.0026 -0.0110 0.0161
## Class1_phi12          -0.0006 -0.0175 0.0159
## Class1_phi22           0.0396  0.0206 0.0625
## Class1_phi32           0.0036 -0.0132 0.0205
## Class1_phi13           0.0026 -0.0110 0.0161
## Class1_phi23           0.0036 -0.0132 0.0205
## Class1_phi33           0.0419  0.0221 0.0669
```

```
## Class2_phi11           0.0416  0.0196 0.0754
## Class2_phi21           0.0077 -0.0096 0.0249
## Class2_phi31          -0.0040 -0.0196 0.0097
## Class2_phi12           0.0077 -0.0096 0.0249
## Class2_phi22           0.0522  0.0239 0.0832
## Class2_phi32          -0.0093 -0.0238 0.0069
## Class2_phi13          -0.0040 -0.0196 0.0097
## Class2_phi23          -0.0093 -0.0238 0.0069
## Class2_phi33           0.0407  0.0225 0.0664
## Class1_sigma2          0.1399  0.1156 0.1718
## Class2_sigma2          0.1373  0.1119 0.1629
##
##  Key to table of posterior means and 95% credible intervals:
## gamma2: variance of stratum-level random effects in latent class membership model
## xi2: variance of cluster-level spatial random effects in latent class
## membership model
## tau2: variance of cluster-level independent random effects in latent class
## membership model
## psi2: variance of stratum-level random effects in longitudinal outcomes model
## omega2: variance of cluster-level random effects in longitudinal outcomes model
## phi: elements of variance-covariance of subject-level random effects in longitudinal
## outcomes model, indexed by row, then column
## sigma2: variance of observation-level in longitudinal outcomes model
##
```

```
##  Model comparison statistics:
##            BIC  ICL-BIC     DIC4
## value 3639.362 3640.102 3512.586
```

Model summaries are printed to the console, including posterior means and 95% credible intervals, posterior latent class assignment, and for model selection, the three model information criteria. A label switching diagnostic using Stephen's algorithm from the **label.switching** package in R is printed.

In the model fitting object, `Bsvygmm` provides a list of matrices of saved posterior samples after discarding the burn-in. For example, to access the samples of the regression coefficients in the longitudinal outcomes model,

```
head(res[["store_beta"]], n = 3)
```

```
##               Class1_time1 Class1_time2 Class1_time3 Class2_time1
## Iteration_501     5.995099     4.939528     4.505680     7.150514
## Iteration_502     5.901227     4.858443     4.401004     7.199024
## Iteration_503     5.941898     4.910269     4.428605     7.144700
##               Class2_time2 Class2_time3
## Iteration_501     8.149440     9.224054
## Iteration_502     8.216366     9.241020
## Iteration_503     8.124307     9.222587
```

The posterior samples can be used for post-estimation analysis.

If writeSamples = TRUE, in the working directory, `Bsvygmm` writes to individual comma-separated files samples for the random effects, draws of Y from the posterior predictive dis-

tribution, and the discrepancy measure in the form of the mean square error. Corresponding text files with the column names are also written.

The file `store_T.txt` contains samples of the discrepancy measure. The **Bsvygmm** function `get_discrepancy_plot` produces a scatter plot of the replicated versus observed discrepancy measure across MCMC samples. The plot is annotated with the Bayesian predictive p-value, which represents the proportion of samples above the diagonal.

```
store_T <- read.table("store_T.txt", header = FALSE, sep = ",")

get_discrepancy_plot(store_T)
```

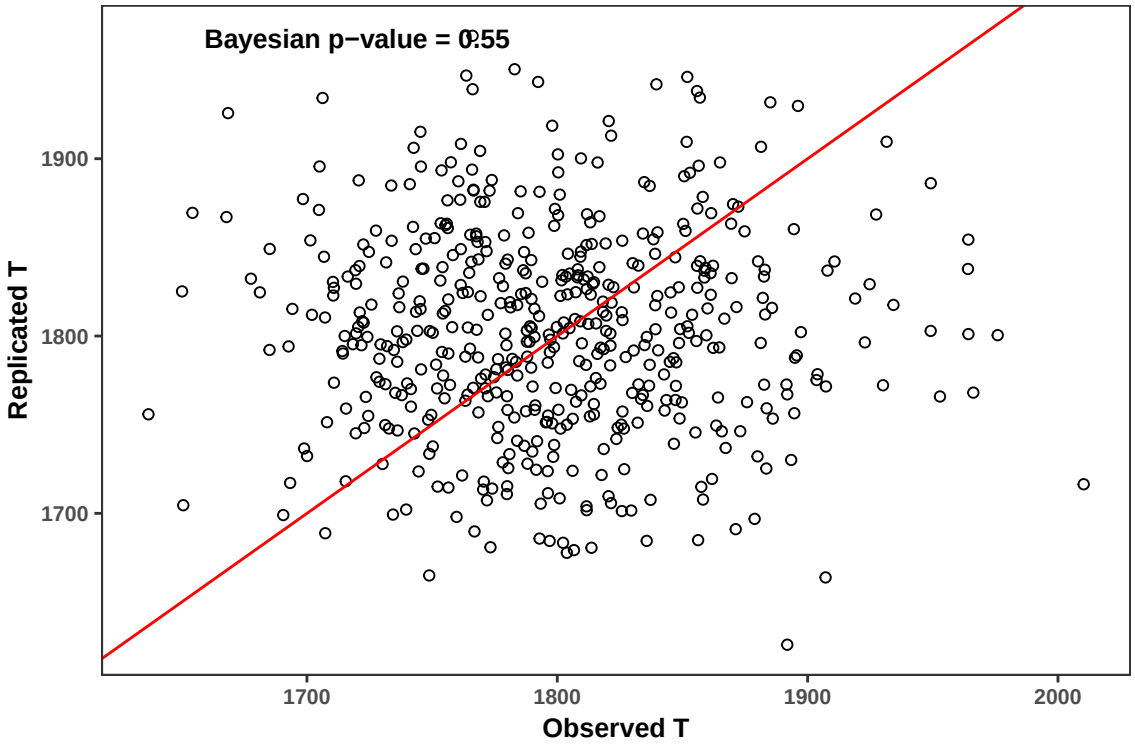

Figure 4.2: Posterior predictive checking for the 2-class model with both types of correlation in the latent class membership model. Observed T is computed using the observed Y. Replicated T is computed using the replicated Y from the posterior predictive distribution.

4.1.2 Analysis with other types of correlations among area segments

In the latent class membership model, instead of modeling both correlations among subjects within an area segment (independent random effects) and spatial correlations among area segments (spatial random effects), one or the other can be selected. If modelType = Str, then only spatial random effects will be modeled. If modelType = Unstr, then only independent random effects will be included.

4.2 R Software Package EHRMiss

The R package **EHRMiss** can be used to fit the proposed Bayesian shared parameter model in Chapter 3. **EHRMiss** is equipped to conduct analyses based on the following assumptions about the missing data mechanisms for the visit process and the response process given a clinic visit:

1. The visit process is MNAR, and one or more of the response processes given a clinic visit is MNAR, with remaining response processes assumed to be MAR;

2. The visit process is MNAR, and all of the response processes given a clinic visit are MAR; or,

3. The visit process is MAR, and all of the response processes given a clinic visit are MAR.

In addition, a naïve analysis that uses only time windows with observed measurements for all longitudinal outcomes may be conducted. In the naïve analysis, the visit process and response process given a clinic visit are not modeled.

EHRMiss can also be used for model selection based on model information criteria including the BIC and DIC3, and model checking using posterior predictive p-values.

To explicate **EHRMiss**, I use an artificial dataset named `growth` stored in the package. Documentation for the dataset can be accessed using ?`growth`. I generated the data using the `simdat` function within **EHRMiss** to reflect a real data analysis with longitudinal data from electronic health records on weight and height z-scores in early childhood. Based on a 2-latent class model, I assumed that the missing data mechanisms for the visit process and the response process for *Y2* given a clinic visit are MNAR. *Y1* is fully observed given a clinic visit. In the sub-models for the longitudinal outcomes, visit process, and response process for *Y2*, I included a random intercept.

The dataset contains longitudinal measurements for 173 subjects followed over 8 clinical time windows. Variables in the dataset include *subjectID*, an integer-valued subject identifier for each measurement time window; *time*, the measurement time window with original values 1,...,8 that is centered and scaled; *Y1* and *Y2*, the longitudinal outcomes of interest; *D*, a binary indicator for the visit process which equals 1 if a clinic visit is observed, and 0 otherwise; *M1* and *M2*, binary indicators for the response process of *Y1* and *Y2*, respectively, each of which equals 1 if a response is observed given a clinic visit, and 0 otherwise; and, *birthweight*, a simulated variable for each subject's birthweight that was centered and scaled. When *D* equals 0, the response indicators are NA. The variables *YC1* and *YC2* correspond to *Y1* and *Y2*, respectively, before the inserting any missed clinic visits or missed responses given a clinic visit. Finally, *Class* takes value 1 or 2 to indicate each subject's latent class membership. On installing **EHRMiss**, the data can be viewed as

```
library(EHRMiss)

data(growth)

head(growth, n = 3)
```

```
##   subjectID       time birthweight         Y1       Y2        YC1
## 1         1 -1.5270478     1.20245         NA       NA -0.8992924
## 2         1 -1.0907484     1.20245 -1.3989838 0.233472 -1.3989838
## 3         1 -0.6544491     1.20245 -0.1893407 0.753286 -0.1893407
##         YC2 D M1 M2 Class
## 1 0.5567838 0 NA NA     1
## 2 0.2334720 1  1  1     1
## 3 0.7532860 1  1  1     1
```

Each subject has 8 time windows of observation in which D measures the visit process, and *M1* and *M2* measure the response process given a clinic visit.

4.2.1 Analysis under an MNAR visit process and response process for *Y2*

I demonstrate fitting the model that generated the data with the function `MVNYMissBinary`. Before fitting the model, a named list with formulas for each of the design matrices must be specified. `MVNYMissBinary` parameterizes the sub-model for the longitudinal outcomes using hierarchical centering. This means that the design matrices for the random effects ("YRe") and the observation-level covariates ("YObs") must not have overlapping columns. "YSub" is a subject-level design matrix for covariates that will enter the random effects equations. Unlike the longitudinal outcomes model, the visit process and response process models do

not use hierarchical centering. Therefore, their design matrices for the fixed and random effects will have overlapping columns.

```
# Named list of formulas for design matrices
regf <- list(LatentClass = ~ 1 + birthweight, # Latent class membership
YRe = ~ 1, # Random effects for Y1, Y2
YObs = ~ -1 + time, # Observation-level fixed effects for Y1, Y2
YSub = ~ 1, # Subject-level fixed effects for Y1, Y2
DObs = ~ 1 + time, # Fixed effects for D
DRe = ~ 1, # Random effects for D
MObs = ~ 1 + time, # Fixed effects for M2
MRe = ~ 1) # Random effects for M2
```

`MVNYBinaryMiss` also requires specifying the parameters for the prior distributions and the initial values. The prior distributions and initial values are supplied to `MVNYBinaryMiss` as lists in which the order of the elements matters. While the prior distributions are not allowed to vary by latent class, initial values must be specified for each latent class. `?MVNYBinaryMiss` provides extensive detail.

```
# Number of outcomes
J <- 2


# Number of latent classes
K <- 2


# Number of covariates for each design matrix
```

```
m <- length(all.vars(regf[["LatentClass"]])) + 1

s <- length(all.vars(regf[["YObs"]]))

p <- length(all.vars(regf[["YSub"]])) + 1

e <- length(all.vars(regf[["DObs"]])) + 1

f <- length(all.vars(regf[["MObs"]])) + 1


# Number of random effects, assumed the same for all models

q <- length(all.vars(regf[["YRe"]])) + 1


# Prior distributions

priors <- list(list(rep(0, m), diag(1, m)),

# Latent class membership

list(rep(0, s), diag(100, s)),

# Observation-level design matrix for Y1, Y2

list(rep(0, p), diag(10000, p)),

# Subject-level design matrix for Y1, Y2

list(1, 1),

# Variance of random intercept for for Y1, Y2

list(diag(c(0.5, 0.5), J), (J + 2)),

# Variance-covariance of Y1, Y2

list(rep(0, e), diag(100, e)),

# Observation-level design matrix for D

list(1, 1),
```

```
# Variance of random intercept for D
list(rep(0, f), diag(100, f)),
# Observation-level design matrix for M2
list(1, 1))  # Variance of random intercept for M2


# Initial values following the same order as in priors
inits <- list(matrix(rep(0, m*(K - 1)), nrow = m, ncol = (K - 1)),


list(matrix(rnorm(s*K), ncol = K, nrow = s),
matrix(rnorm(s*K), ncol = K, nrow = s)),


list(array(rnorm(p*q*K), dim = c(p, q, K)),
array(rnorm(p*q*K), dim = c(p, q, K))),


list(array(rep(0.4, K), dim = c(q, q, K)),
array(rep(0.4, K), dim = c(q, q, K))),


array(c(1, 0, 0, 1, 0.5, 0, 0, 0.5), dim = c(J, J, K)),


matrix(rnorm(e*K), ncol = K),


array(rep(0.5, K), dim = c(q, q, K)),
```

```
list(matrix(rnorm(f*K), ncol = K)),


list(array(rep(0.5, K), dim = c(q, q, K))))
```

To account for the MNAR visit process, I set modelVisit = TRUE. By modelResponse = TRUE, `MVNYMissBinary` understands that one or more of the response processes given a clinic visit will be assumed to be MNAR, thus requiring modeling. I set Mvec = 2 to indicate that *M2* (the response process for *Y2* given a clinic visit) will be modeled. Except for a naïve analysis, all analyses require that imputeResponse = TRUE.

```
# Set interval update to 500 and monitor = TRUE
res <- MVNYBinaryMiss(K = K, J = J, data = growth, regf = regf,
imputeResponse = TRUE, Mvec = 2,
modelVisit = TRUE, modelResponse = TRUE,
priors = priors, inits = inits, n.samples = 1000, burn = 500, monitor =
TRUE, update = 500, modelComparison = TRUE, sims = FALSE)
```

`MVNYBinaryMiss` processes the dataset for an analysis with indicated missing data assumptions. The function prints the number of observations that will be used in the model for *Y1* and *Y2*. When imputeResponse = TRUE, this is the number of observed clinic visits. If the number of unique subjects in the sub-models for latent class membership, the longitudinal outcomes of interest, the visit process, and the response process given a clinic visit are not equal, `MVNYBinaryMiss` will produce an error.

In the call to `MVNYBinaryMiss`, I run the MCMC sampler for 1000 iterations with a burn-in of 500. Since I set update = 500 with monitor = TRUE, the iteration number

and predicted class size will be printed to the console every 500 iterations. In addition, a graphic with selected trace plots will be updated.

```
## Number of obs. after restricting to observed visits: 532

## Iteration: 500

## Class size: 69 104

## Iteration: 1000

## Class size: 70 103
```

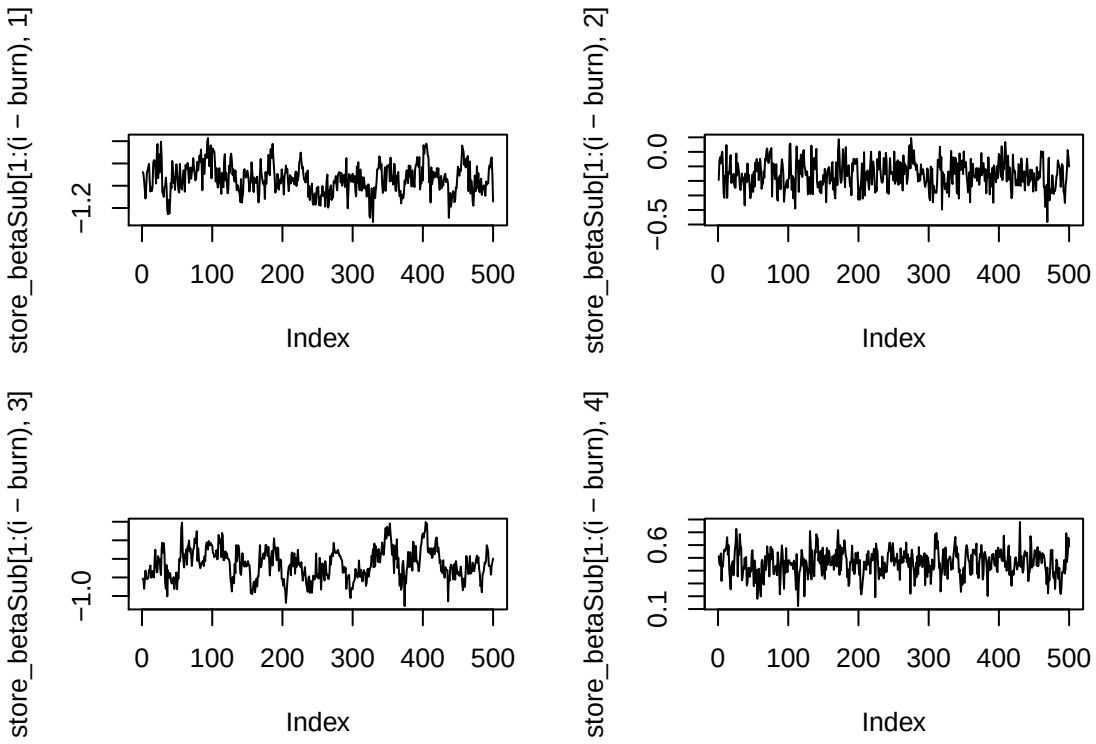

Figure 4.3: Trace plots of the first four regression coefficients in the design matrix for "YSub". In this analysis, these are the latent-class specific intercepts for *Y1* and *Y2*.

```
## Total minutes elapsed: 47.1395 0.019 48.09517 NA NA
## No evidence of label switching problem using Stephen's method
## from the label.switching package
##
## Background information
## Number of subjects: 173
## Number of observations: 532
## Number of latent classes: 2
##
## Posterior latent class assignment:
##                                          Class 1 Class 2
## Predicted class size                       72.00  101.00
## No. subjects with probability at least 0.95   58.00   86.00
## No. subjects with probability at least 0.90   62.00   90.00
## No. subjects with probability at least 0.80   66.00   94.00
## Mean probability                            0.95    0.96
## Median probability                          1.00    1.00
##
##  Reference class in latent class membership model: 1
## Posterior means and 95% credible intervals:
##                          Post. Mean   2.5 %   97.5%
## Class2_(Intercept)           0.1954 -0.0398  0.4053
## Class2_birthweight           0.8025  0.5056  1.0628
## Y1_Class1_time               0.6117  0.4118  0.8500
```

```
## Y1_Class2_time              -0.0025 -0.1028  0.0942
## Y2_Class1_time               0.7881  0.5219  1.0440
## Y2_Class2_time               0.0734 -0.0147  0.1591
## Class1_Sigma11               1.4582  1.0945  1.8805
## Class1_Sigma21               0.7045  0.4274  1.0198
## Class1_Sigma12               0.7045  0.4274  1.0198
## Class1_Sigma22               1.3905  1.0320  1.9122
## Class2_Sigma11               0.5109  0.4282  0.6086
## Class2_Sigma21               0.1671  0.1081  0.2315
## Class2_Sigma12               0.1671  0.1081  0.2315
## Class2_Sigma22               0.4943  0.4143  0.5903
## Y1_Class1_RE1_(Intercept)   -0.9347 -1.1772 -0.6664
## Y1_Class2_RE1_(Intercept)   -0.1551 -0.3355  0.0302
## Y2_Class1_RE1_(Intercept)   -0.6544 -0.9569 -0.3228
## Y2_Class2_RE1_(Intercept)    0.4597  0.2413  0.6624
## Y1_Class1_Psi11              0.4925  0.1549  0.9929
## Y1_Class2_Psi11              0.6266  0.4266  0.9158
## Y2_Class1_Psi11              0.4810  0.1610  1.0044
## Y2_Class2_Psi11              0.6586  0.4509  0.9102
## AME_Y1_RE1_(Intercept)      -0.4963 -0.6502 -0.3409
## AME_Y1_time                  0.2661  0.1576  0.3727
## AME_Y2_RE1_(Intercept)      -0.0270 -0.1893  0.1430
## AME_Y2_time                  0.3855  0.2697  0.5165
## D_Class1_(Intercept)        -0.6300 -0.7620 -0.5014
```

```
## D_Class1_time                0.1143 -0.0036  0.2473
## D_Class2_(Intercept)        -0.1444 -0.2962  0.0080
## D_Class2_time               -0.8636 -0.9996 -0.7369
## D_Class1_Omega11             0.1167  0.0589  0.2002
## D_Class2_Omega11             0.3809  0.2176  0.5942
## M2_Class1_(Intercept)        1.0102  0.5457  1.5592
## M2_Class1_time               0.3650  0.0519  0.6729
## M2_Class2_(Intercept)        1.8812  1.4256  2.3099
## M2_Class2_time              -0.0210 -0.3242  0.2650
## M2_Class1_Theta11            1.7606  0.5193  3.9284
## M2_Class2_Theta11            0.6784  0.2894  1.3450
##
##  Footnotes for posterior means and 95% credible intervals:
## Elements of variance-covariances are indexed by row and then column.
## RE indexes the random effects equations. If there is only a
## random intercept, then this will be RE1.
## AME indicates the population-averaged regression coefficients.
##
##  Model comparison statistics:
##           BIC1   BIC2     DIC3      LPML
## value 5194.167 5183.4 5330.418 -2924.723
##
##  Footnotes for model comparison:
## BIC1: Computed using number of observations equal to the number of
```

```
## observed clinic visits
## BIC2: Computed using number of observations equal to the effective sample size
## from the longitudinal health outcomes model
## DIC calculation details
##             DIC3
## Dbar   5168.0095
## Dtilde 5005.6009
## pD      162.4086
```

`MVNYMissBinary` prints model output to the `R` console, saves output in the model fitting object, and writes output to comma-separated text files in the working directory. First, I explain the model output printed to the console. Before conducting post-estimation analysis with the posterior samples, `MVNYMissBinary` uses Stephen's algorithm in the `R` package **label.switching** to diagnose the presence of the label switching phenomenon which can occur in finite mixture modeling. After background information, a summary of posterior latent class assignment is provided, followed by the posterior mean estimators and associated 95% credible intervals. If modelComparison = TRUE, three model information criteria are computed, in addition to the log pseudo-marginal likelihood (LPML).

In the model fitting object, `MVNYBinaryMiss` provides a list of matrices of posterior samples saved after burn-in. For example, to access the samples of the observation-level design matrix for *Y1* and *Y2*,

```
head(res[["store_betaObs"]], n = 3)
```

```
##               Y1_Class1_time Y1_Class2_time Y2_Class1_time Y2_Class2_time
## Iteration_501      0.5183279    0.018091410      0.7843445     0.12130878
```

```
## Iteration_502      0.5152602    0.002508136      0.7778971     0.08035319
## Iteration_503      0.7144048    0.007618712      0.8257602     0.05694670
```

or the posterior probabilities of latent class membership,

```
head(res[["store_pi"]][ , 1:4], n = 3)
```

```
##                Class1_Subject_1 Class1_Subject_2 Class1_Subject_3
## Iteration_501     4.500746e-05        0.9999993        0.9999827
## Iteration_502     3.112051e-07        0.9999398        0.9999995
## Iteration_503     5.682738e-04        0.9999890        0.9999999
##                Class1_Subject_4
## Iteration_501        0.9987603
## Iteration_502        0.9998808
## Iteration_503        0.9999902
```

The posterior samples can be used for post-estimation analysis.

In the working directory, `MVNYBinaryMiss` writes to separate files to store imputations of each longitudinal outcome given an observed clinic visit (e.g., `store_miss_Y2.txt`). To form a completed dataset, the imputations can be inserted into the data, for example, as

```
Y <- subset(growth, subset = D == 1, select = paste("Y", 1:J, sep = ""))
M <- subset(growth, subset = D == 1, select = paste("M", 1:J, sep = ""))

store_miss_Y2 <- data.matrix(read.table("store_miss_Y2.txt", sep = ","))
```

```
Ycomplete <- Y

# e.g., use the iteration 10 (after burn-in)

Ycomplete[M[ , 2] == 0 , 2] <- store_miss_Y2[10, ]
```

The file `store_T_completed.txt` of the stored discrepancy measure, the multivariate mean square error, is written. The **EHRMiss** function `get_discrepancy_plot` produces a scatter plot of the replicated completed versus completed discrepancy measure across MCMC samples. The plot is annotated with the Bayesian predictive p-value, which represents the proportion of samples above the diagonal.

```
store_T_completed <- read.table("store_T_completed.txt", header = FALSE,

sep = ",")

get_discrepancy_plot(store_T_completed)
```

Samples of replicated completed longitudinal outcomes are written to `store_Ydraw.txt`. These samples can be used to diagnose model fit.

4.2.2 Analysis under different missing data assumptions

To conduct an analysis assuming that the visit process is MNAR, and all of the response processes given a clinic visit are MAR, I change the function call in 4.2.1 with modelResponse = FALSE and Mvec = NULL. For the assumptions of an MAR visit process, and all MAR response processes given a clinic visit, I also set modelVisit = FALSE. For the different assumed missing data mechanisms, imputeResponse = TRUE is required.

A naïve analysis, in which only time windows with observed measurements for all longitudinal outcomes are used, is conducted by setting imputeResponse = FALSE. Modeling of the visit process or response process given a clinic visit is not permitted.

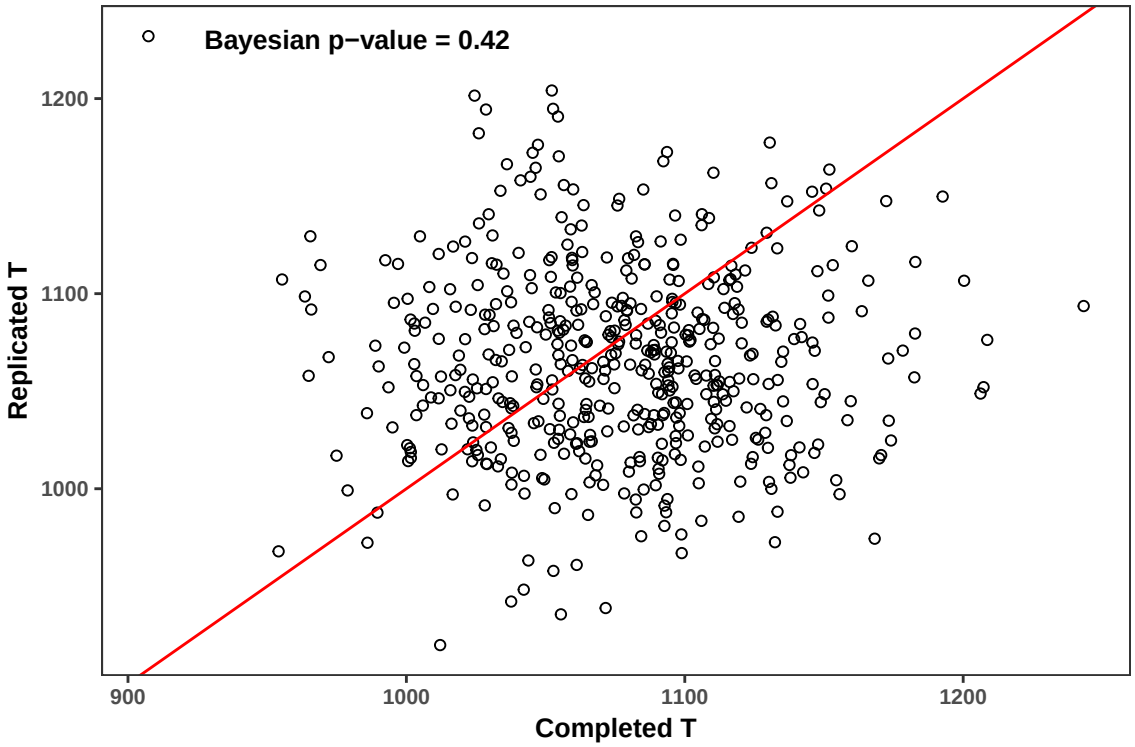

Figure 4.4: Posterior predictive checking for the 2-class model estimated assuming an MNAR visit process and response process for *Y2* given a clinic visit. Completed T is computed using the completed data. Replicated T is computed using the replicated completed datasets from the posterior predictive distribution.

Chapter 5

Conclusion

In this dissertation, I proposed statistical methods for modeling latent heterogeneity in complex survey data and electronic health records, and developed corresponding software to make these methods widely accessible. Each of the methods addresses a gap in the existing literature. For complex survey data, the proposed Bayesian growth mixture model complements existing pseudo-likelihood methods. By flexibly incorporating the hierarchical structure of the data and the different features of the complex sample design, my method can easily be applied to diverse survey data applications. For electronic health records, the proposed Bayesian shared parameter model extends a growth mixture model of multiple longitudinal health outcomes to account for different missing data assumptions. As routinely collected data sources are increasingly used for scientific research, my method provides a necessary tool for validating statistical findings.

Part I

Appendices

Appendix A

MCMC Algorithm for Bayesian GMM in Complex Survey Data

Appendix A explicates the MCMC algorithm for fitting the Bayesian GMM in complex survey data.

A.0.1 Update parameters in the latent class membership model

The Gibbs steps are given for the latent class membership model with both u_{sjk} and ν_{sjk}.

1. Update $\mathbf{z}^*_{sji}$. Recall that $\mathbf{z}^*_{sji} = (z^*_{sji1}, \ldots, z^*_{sjiK-1})^T$. Per [McCulloch and Rossi, 1994], for $i = 1, \ldots, n$, the distribution of $\mathbf{z}^*_{sji} \mid \delta, c^*_{sji}$ is a $(K-1)$-variate normal distribution truncated over the appropriate cone in $\mathbf{R}^{K-1}$. Let $\mathbf{d}_{sji}$ be a multinomial vector with entries $\mathbf{d}_{sji} = (d_{sji1}, \ldots, d_{sjiK})$ equal to 1 if the i^{th} subject is in latent class k and 0 otherwise. If $d_{sjik} = 1$, then $z^*_{sjik} > max(\mathbf{z}^*_{sji,-k}, 0)$. If $d_{sjik} = 0$, then $z^*_{sjik} < max(\mathbf{z}^*_{sji,-k}, 0)$. $\mathbf{z}^*_{sji,-k}$ is a $K-2$ dimensional vector of all components of $\mathbf{z}^*_{sji}$ excluding z^*_{sjik}. This algorithm avoids the problem of drawing from a truncated multivariate normal. Instead each draw is a truncated univariate normal because I am using the conditional distribution $z^*_{sjik} \mid \mathbf{z}^*_{sji,-k}, \delta_k, c^*_{sji}$, where $c^*_{sji} = K$ if $\max(\mathbf{z}^*_{sji}) <$

0, or else $c^*_{sji} =$ index of $max(\mathbf{z}^*_{sji})$ for $k = 1, \ldots, K-1$.

2. Update δ_k. For $k = 1, \ldots, K-1$, I assume the prior $\delta_k \sim N_m(0, \mathbf{\Sigma}_0)$. The full conditional is $N_m(\mu_{\delta_k}, \mathbf{V}_\delta)$, where

$$\mathbf{V}_\delta = \left(\sum_{s=1}^{S} \sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} \mathbf{w}_{sji} \mathbf{w}_{sji}^T + \mathbf{\Sigma}_0^{-1} \right)^{-1}$$

$$\mu_{\delta_k} = \mathbf{V}_{\delta_k} \times \left(\sum_{s=1}^{S} \sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} \mathbf{w}_{sji} (z^*_{sjik} - \lambda_{sk} - u_{sjk} - \nu_{sjk}) \right),$$

with $\mathbf{w}_{sji}$ being an m-length column vector of covariates.

3. Update λ_{sk}. For $k = 1, \ldots, K-1$, $s = 1, \ldots, S$, I assume the prior $\lambda_{sk} \sim N(0, \gamma_k^2)$. The full conditional is $N(\mu_{\lambda_{sk}}, V_{\lambda_{sk}})$, where

$$V_{\lambda_{sk}} = \left(n_s + \frac{1}{\gamma_k^2} \right)^{-1}$$

$$\mu_{\lambda_{sk}} = V_{\lambda_{sk}} \times \left(\sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} (z^*_{sjik} - \mathbf{w}_{sji}^T \delta_k - u_{sjk} - \nu_{sjk}) \right).$$

$n_s = \sum_{j=1}^{J_s} n_{sj}$ is the number of subjects in stratum s.

4. Update γ_k^2. For $k = 1, \ldots, K-1$, the prior is $\propto 1$. Per [Gelman, 2006], this is $\propto \frac{1}{\gamma_k}$. Therefore, the full conditional is $IG(a_{\gamma_k}, b_{\gamma_k})$, where

$$a_{\gamma_k} = \frac{S-1}{2}$$

$$b_{\gamma_k} = \frac{\sum_{s=1}^{S} \lambda_{sk}^2}{2}.$$

5. Update u_{sjk}. For $k = 1, \ldots, K-1$, $j = 1, \ldots, J_s$, and $s = 1, \ldots, S$, I assume the prior $u_{sjk} \sim N(0, \tau_k^2)$. The full conditional is $N(\mu_{u_{sjk}}, V_{u_{sjk}})$, where

$$V_{u_{sjk}} = \left(n_{sj} + \frac{1}{\tau_k^2} \right)^{-1}$$

$$\mu_{u_{sjk}} = V_{u_{sjk}} \times \left(\sum_{i=1}^{n_{sj}} (z^*_{sjik} - \mathbf{w}_{sji}^T \delta_k - \lambda_{sk} - \nu_{sjk}) \right).$$

n_{sj} is the number of subjects in stratum s and area segment j.

6. Update τ_k^2. For $k = 1, \ldots, K-1$, the prior is $\propto 1$. Per [Gelman, 2006], this is $\propto \frac{1}{\tau_k}$. Therefore, the full conditional is $IG(a_{\tau_k}, b_{\tau_k})$, where

$$a_{\tau_k} = \frac{J-1}{2}$$
$$b_{\tau_k} = \frac{\sum_{s=1}^{S}\sum_{j=1}^{J_s} u_{sjk}^2}{2},$$

with $J = \sum_{s=1}^{S} J_s$ is the number of area segments.

7. Update ν_{sjk}. For $k = 1, \ldots, K-1$, $j = 1, \ldots, J_s$, and $s = 1, \ldots, S$, the prior distribution is an intrinsic conditional autoregressive (ICAR) [Besag, 1974; Besag and Kooperberg, 1995]:

$$\nu_{sjk}|\nu_{(-sjk)} \sim N\left(\bar{\nu}_{sjk}, \frac{\xi_k^2}{m_{sj}}\right),$$

where m_{sj} is the number of neighbors for area segment j of stratum s, and ξ_k^2 is the latent class-specific spatial variance scaled by the number of neighbors. The conditional mean is defined according to neighboring area segments of area segment j in stratum s, indicated by ∂_{sj}. I write $\bar{\nu}_{sjk} = \sum_{l \in \partial_{sj}} \frac{\nu_{sjk,l}}{m_{sj}}$. The full conditional is $N(\mu_{\nu_{sjk}}, V_{\nu_{sjk}})$, where

$$V_{\nu_{sjk}} = \left(n_{sj} + \frac{m_{sj}}{\xi_k^2}\right)^{-1}$$
$$\mu_{\nu_{sjk}} = V_{\nu_{sjk}} \times \left(\sum_{i=1}^{n_{sj}} (z_{sjik}^* - \mathbf{w}_{sji}^T \delta_k - \lambda_{sk} - u_{sjk}) + \frac{m_{sj}\bar{\nu}_{sjk}}{\xi_k^2}\right).$$

8. Update ξ_k^2. For $k = 1, \ldots, K-1$, the prior is $\propto 1$. Per [Gelman, 2006], this is $\propto \frac{1}{\xi_k}$. Therefore, the full conditional is $IG(a_{\xi_k}, b_{\xi_k})$, where

$$a_{\xi_k} = \frac{J-1}{2}$$
$$b_{\xi_k} = \frac{\sum_{s=1}^{S}\sum_{j=1}^{J_s}(m_{sj}(\nu_{sjk}^2 - 2\nu_{sjk}\bar{\nu}_{sjk} + \bar{\nu}_{sjk}^2))}{2}.$$

$J = \sum_{s=1}^{S} J_s$ is the total number of area segments, and m_{sj} is the number of neighbors for the j^{th} area segment of stratum s.

A.0.2 Update parameters in the longitudinal outcomes model

I provide the MCMC algorithm for a general version of the longitudinal model of PTSD severity scores. Specifically, I rewrite $\mathbf{b}_{sji}$ as $\mathbf{b}_{sji} = \mathbf{v}^f_{sji}\beta_k + \mathbf{v}^r_{sji}\eta_{sji}$, where $\mathbf{v}^f_{sji}$ is an $n_{sji} \times p$ design matrix for fixed effects with corresponding latent class-specific regression coefficients in β_k, and $\mathbf{v}^r_{sji}$ is an $n_{sji} \times q$ design matrix for random effects that is a subset of $\mathbf{v}^f_{sji}$. n_{sji} is the number of longitudinal measurements for the i^{th} subject in area segment j of stratum s. I then assume $\eta_{sji} \sim N_q(0, \mathbf{\Phi}_k)$.

Following [Frühwirth-Schnatter *et al.*, 2004] and [Fruhwirth-Schnatter, 2006], in the partially marginalized Gibbs sampler, the updates for β_k and c^*_{sji} are based on the marginal distribution of $\mathbf{y}_{sji} \mid c^*_{sji}$ that is obtained by integrating out the random effects. Specifically, the marginal mean (conditioning on latent class) is

$$E[\mathbf{y}_{sji} \mid c^*_{sji} = k] = \mathbf{v}^f_{sji}\beta_k. \tag{A.1}$$

The marginal variance (conditioning on latent class) is

$$\mathbf{R}_{sjik} = Var[\mathbf{y}_{sji} \mid c^*_{sji} = k] = \mathbf{I}_{n_{sji}}\sigma^2_k + \mathbf{v}^r_{sji}\mathbf{\Phi}_k\mathbf{v}^{r^T}_{sji} + \omega^2_k + \psi^2_k. \tag{A.2}$$

1. Update $\eta_{sji} \mid c^*_{sji} = k$. For $i = 1, \ldots, n_{sj}$, $j = 1, \ldots, J_s$, and $s = 1, \ldots, S$, I assume the prior distribution $\eta_{sji} \mid c^*_{sji} = k \sim MVN_q(0, \mathbf{\Phi}_k)$. The full conditional is $N_q(\mu_\eta, \mathbf{V}_\eta)$, where

$$\mathbf{V}_\eta = \left(\frac{(\mathbf{v}^r_{sji})^T(\mathbf{v}^r_{sji})}{\sigma^2_k} + \Phi^{-1}_k \right)^{-1}$$

$$\mu_\eta = \mathbf{V}_\eta \times \left(\frac{(\mathbf{v}^r_{sji})^T(\mathbf{y}_{sji} - \mathbf{v}^f_{sji}\beta_k - \rho_{sjk} - \zeta_{sk})}{\sigma^2_k} \right)$$

2. Update ρ_{sjk}. For $k = 1, \ldots, K$, $j = 1, \ldots, J_s$, and $s = 1, \ldots, S$, I assume the prior $\rho_{sjk} \sim N(0, \omega_k^2)$. The full conditional is $N(\mu_{\rho_{sjk}}, V_{\rho_{sjk}})$, where

$$V_{\rho_{sjk}} = \left(\frac{N_{sjk}}{\sigma_k^2} + \frac{1}{\omega_k^2} \right)^{-1}$$
$$\mu_{\rho_{sjk}} = V_{\rho_{sjk}} \times \left(\frac{\sum_{i=1}^{n_{sj}} \mathbf{1}_{c^*_{sji}=k} \times \sum_{t=1}^{n_{sji}} (y_{sjit} - \mathbf{v}_{sjit}^{f^T} \beta_k - \mathbf{v}_{sjit}^{r^T} \eta_{sji} - \zeta_{sk})}{\sigma_k^2} \right),$$

with $N_{sjk} = \sum_{i=1}^{n_{sj}} \mathbf{1}_{c^*_{sji}=k} \times n_{sji}$. N_{sjk} is the number of observations in stratum s, area segment j, and latent class k.

3. Update ζ_{sk}. For $k = 1, \ldots, K$, and $s = 1, \ldots, S$, I assume the prior $\zeta_{sk} \sim N(0, \psi_k^2)$. The full conditional is $N(\mu_{\zeta_{sk}}, V_{\zeta_{sk}})$, where

$$V_{\zeta_{sk}} = \left(\frac{N_{sk}}{\sigma_k^2} + \frac{1}{\psi_k^2} \right)^{-1}$$
$$\mu_{\zeta_{sk}} = V_{\zeta_{sk}} \times \left(\frac{\sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} \mathbf{1}_{c^*_{sji}=k} \times \sum_{t=1}^{n_{sji}} (y_{sjit} - \mathbf{v}_{sjit}^{f^T} \beta_k - \mathbf{v}_{sjit}^{r^T} \eta_{sji} - \rho_{sjk})}{\sigma_k^2} \right).$$

with $N_{sk} = \sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} \mathbf{1}_{c^*_{sji}=k} \times n_{sji}$. N_{sk} is the number of observations in stratum s and latent class k.

4. Update $\mathbf{\Phi}_k$. For $k = 1, \ldots, K$, I assume $\mathbf{\Phi}_k \sim IW(\nu_0, S_0^{-1})$. The full conditional is $IW(a_{\Phi_k}, b_{\Phi_k})$, where

$$a_{\Phi_k} = \nu_0 + n_k$$
$$b_{\Phi_k} = S_0 + \sum_{s=1}^{S} \sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} \left(\mathbf{1}_{c^*_{sji}=k} \times \eta_{sji} \eta_{sji}^T \right),$$

with $n_k = \sum_{s=1}^{S} \sum_{j=1}^{J_s} \sum_{i=1}^{n_{sj}} \mathbf{1}_{c^*_{sji}=k}$. n_k is the number of subjects in latent class k.

5. Update σ_k^2. For $k = 1, \ldots, K$, I assume $\sigma_k^2 \sim IG(0.1, 0.1)$. The full conditional is

$IG(a_{\sigma_k^2}, b_{\sigma_k^2})$, where

$$a_{\sigma_k^2} = 0.1 + \frac{N_k}{2}$$

$$b_{\sigma_k^2} = 0.1 + \frac{\sum_{s=1}^{S}\sum_{j=1}^{J_s}\sum_{i=1}^{n_{sj}}\left(\mathbf{1}_{c^*_{sji}=k} \times \sum_{t=1}^{n_{sji}}\left(y_{sjit} - \mathbf{v}_{sjit}^{f^T}\beta_k + \mathbf{v}_{sjit}^{f^T}\eta_{sji} + \rho_{sjk} + \zeta_{sk}\right)\right)^2}{2}.$$

$N_k = \sum_{s=1}^{S}\sum_{j=1}^{J_s}\sum_{i=1}^{n_{sj}}(\mathbf{1}_{c^*_{sji}=k} \times n_{sji})$, the number of observations in latent class k.

6. Update ω_k^2. For $k = 1, \ldots, K$, the prior is $\propto 1$. Per [Gelman, 2006], this is $\propto \frac{1}{\omega_k}$. Therefore, the full conditional is $IG(a_{\omega_k^2}, b_{\omega_k^2})$, where

$$a_{\omega_k^2} = \frac{\sum_{s=1}^{S} J_s - 1}{2}$$
$$b_{\omega_k^2} = \frac{\sum_{s=1}^{S}\sum_{j=1}^{J_s}\rho_{sjk}^2}{2}.$$

7. Update ψ_k^2. For $k = 1, \ldots, K$, the prior is $\propto 1$. Per [Gelman, 2006], this is $\propto \frac{1}{\psi_k}$. Therefore, the full conditional is $IG(a_{\psi_k^2}, b_{\psi_k^2})$, where

$$a_{\psi_k^2} = \frac{S-1}{2}$$
$$b_{\psi_k^2} = \frac{\sum_{s=1}^{S}\zeta_{sk}^2}{2}.$$

8. Update β_k. In the partially marginalized Gibbs sampler, the update for β_k uses the marginal distribution of $\mathbf{y}_{sji}$ with ζ_k, ρ_k, and $\mathbf{b}_{sji}$ integrated out. The mean and variance of this distribution are shown in equations A.1 and A.2, respectively.

   For $k = 1, \ldots, K$, assuming the prior $\beta_k \sim N_p(0, \mathbf{\Sigma}_0)$, the full conditional is $N_p(\mu_{\beta_k}, \mathbf{V}_{\beta_k})$ where

$$\mathbf{V}_{\beta_k} = \left(\sum_{s=1}^{S}\sum_{j=1}^{J_s}\sum_{i=1}^{n_{sj}}\left(\mathbf{1}_{c^*_{sji}=k} \times \mathbf{v}_{sji}^{f^T}\mathbf{R}_{sjik}^{-1}\mathbf{v}_{sji}^{f^T}\right) + \mathbf{\Sigma}_0^{-1}\right)^{-1}$$

$$\mu_{\beta_k} = \mathbf{V}_{\beta_k} \times \left(\sum_{s=1}^{S}\sum_{j=1}^{J_s}\sum_{i=1}^{n_{sj}}\mathbf{1}_{c^*_{sji}=k} \times \mathbf{v}_{sji}^{f^T}\mathbf{R}_{sjik}^{-1}\mathbf{y}_{sji}\right).$$

9. Update c^*_{sji}. In the partially marginalized Gibbs sampler, the update for c^*_{sji} uses the marginal distribution of $\mathbf{y}_{sji} \mid c^*_{sji}$ with ζ_k, ρ_k, and $\mathbf{b}_{sji}$ integrated out. The mean and variance of this distribution are shown in equations A.1 and A.2, respectively.

   Using Bayes' theorem, the posterior probability that subject i belongs to latent class k $(k = 1, \ldots, K)$ is

$$\begin{aligned} p_{sjik} =& Pr(c^*_{sji} = k \mid \mathbf{y}_{sji}, \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k, \pi_{sjik}; \mathbf{v}^f_{sji}, \mathbf{v}^r_{sji}) \\ =& \frac{\pi_{sjik}\, f(\mathbf{y}_{sji} \mid \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k; \mathbf{v}^f_{sji}, \mathbf{v}^r_{sji})}{\sum_{k=1}^{K} \pi_{sjik}\, f(\mathbf{y}_{sji} \mid \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k; \mathbf{v}^f_{sji}, \mathbf{v}^r_{sji})}, \end{aligned}$$

   where π_{sjik} is the probability of latent class membership obtained from the latent class membership model, and the likelihood contribution to latent class k is obtained from the partially marginalized density $f(\mathbf{y}_{sji} \mid \beta_k, \sigma^2_k, \mathbf{\Phi}_k, \omega^2_k, \psi^2_k; \mathbf{v}^f_{sji}, \mathbf{v}^r_{sji})$ with mean and variance given in equations (A.1) and (A.2), respectively.

Appendix B

Model Information Criteria for Bayesian GMM in Complex Survey Data

In Appendix B, I explicate the different model information criteria used in model selection.

The Bayesian Information Criterion (BIC) [Schwarz, 1978] is derived as an approximation to the marginal likelihood using the Laplace method. In mixture models, however, the necessary regularity conditions do not hold for assessing the number of components K. Notwithstanding, the BIC has been shown to be consistent for choosing the number of components if the distribution family of component densities is correctly specified [Keribin, 2000]. According to simulation studies in Biernacki et al. [Biernacki *et al.*, 2000], the BIC exhibits superior performance in selecting the true number of components if the modeling objective is non-parametric density estimation. However, if the modeling objective is a clustering analysis, the BIC tends to overestimate the number of clusters K when the quality

of model fit is poor. I compute the BIC as

$$\text{BIC} = -2\log\prod_{s=1}^{S}\prod_{j=1}^{J_s}\prod_{i=1}^{n_{sj}}\sum_{k=1}^{K}\pi_{sjik}f(\mathbf{y}_{sji}\,|\,\hat{\beta}_k,\hat{\sigma}_k^2,\hat{\mathbf{\Phi}}_k,\hat{\omega}_k^2,\hat{\psi}_k^2;\mathbf{v}_{sji}^f,\mathbf{v}_{sji}^r) + d_K\log N,$$

where $f(\mathbf{y}_{sji}\,|\,\hat{\beta}_k,\hat{\sigma}_k^2,\hat{\mathbf{\Phi}}_k,\hat{\omega}_k^2,\hat{\psi}_k^2;\mathbf{v}_{sji}^f,\mathbf{v}_{sji}^r)$ is the partially marginalized density after integrating out the random effects (as in A.1 and A.2 and called the observed data likelihood) evaluated at the maximum likelihood estimates of the parameters. d_K is the number of free parameters, and N is the sample size. I approximated the maximum likelihood estimator by maximizing the log of the observed data likelihood over MCMC samples.

The integrated classification likelihood (ICL) [Biernacki *et al.*, 2000] extends the BIC to account for the clustering structure of the data. The ICL has been shown to detect the correct number of clusters even under model misspecification. When the number of observations is large in a component, the ICL can be approximated using the BIC as

$$\text{ICL-BIC} = \text{BIC} + 2\sum_{s=1}^{S}\sum_{j=1}^{J_s}\sum_{i=1}^{n_{sj}}\sum_{k=1}^{K}\hat{p}_{sjik}\log\hat{p}_{sjik}$$

where the second term is a measure of entropy using p_{sjik} evaluated at the maximum likelihood estimator of the observed data likelihood. Entropy quantifies the degree to which the fitted K component model fails to partition the data. Under well-separated clusters, entropy will be near 0. As the degree of separation worsens, the value of entropy will become very large [Fruhwirth-Schnatter, 2006]. Therefore, ICL-BIC penalizes not only model complexity but also poorly separated clusters.

Outside of latent variable modeling, the Deviance Information Criterion (DIC) is based on the effective number – as opposed to the actual number – of model parameters. The DIC is calculated by subtracting the deviance evaluated at the posterior means of model parameters from the expected deviance averaged over MCMC iterations [Spiegelhalter *et al.*, 2002]. In mixture models, however, the DIC often results in a negative number of effective

parameters [Celeux *et al.*, 2006]. For an analogous criterion in latent variable modeling, Celeux et al. [Celeux *et al.*, 2006] recommend the DIC4. Below, I define the DIC4.

Let $\mathbf{\Theta}_K$ be a container for parameters in a K component model. Define the observed data likelihood $f(\mathbf{y} \mid \mathbf{\Theta}_K)$ as

$$f(\mathbf{y} \mid \mathbf{\Theta}_K) = \prod_{s=1}^{S} \prod_{j=1}^{J_s} \prod_{i=1}^{n_{sj}} \sum_{k=1}^{K} \pi_{sjik} f(\mathbf{y}_{sji} \,|\, \beta_k, \sigma_k^2, \mathbf{\Phi}_k, \omega_k^2, \psi_k^2; \mathbf{v}_{sji}^f, \mathbf{v}_{sji}^r).$$

The DIC4 can be approximated by [Celeux *et al.*, 2006]

$$\text{DIC4} = -4E_{\mathbf{\Theta}_K}[\log f(\mathbf{y} \mid \mathbf{\Theta}_K) \mid \mathbf{y}] + 2(\log f(\mathbf{y} \mid \hat{\mathbf{\Theta}}_K^M, \mathbf{y}) + E_{\mathbf{\Theta}_K}[EN(\mathbf{\Theta}_K \mid \mathbf{y})]),$$

where $\hat{\mathbf{\Theta}}_K^M$ is the posterior mode estimator obtained from the observed data posterior, and $EN(\mathbf{\Theta}_K \mid \mathbf{y})$ is the measure of entropy used in the ICL-BIC. The expectations are calculated by averaging over MCMC samples. The DIC4 penalizes poorly separated clusters in addition to model complexity.

Appendix C

Sensitivity Analysis Removing Complex Sample Design

In Appendix C, I present the findings from the sensitivity analysis in which I fit Bayesian GMMs assuming $K = 2, 3, 4$ latent classes that removed all information about the complex sample design.

Table C.1: Comparison of information criteria among models without accounting for complex sample design, assuming $K = 2, 3, 4$ latent classes.

	K		
Criterion	2	3	4
BIC	-628.27	-726.79	-649.25
ICL - BIC	-566.91	-584.32	-426.56
DIC4	-760.28	-946.04	-982.62

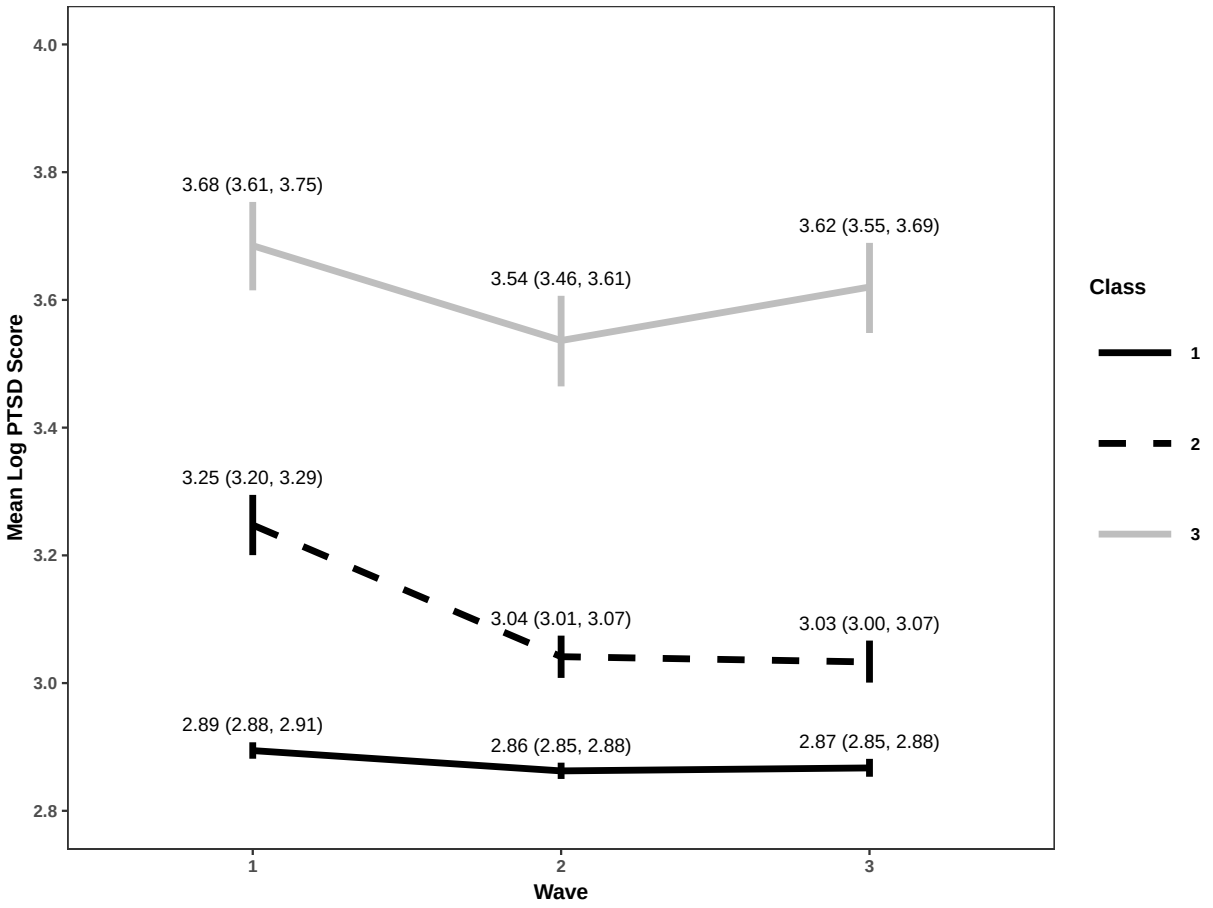

Figure C.1: Mean log PTSD severity score trajectory in each latent class based on the posterior mean and 95% credible interval of β_k in the longitudinal model of PTSD that did not include information on the complex sample design.

Table C.2: Variance components in the longitudinal model for PTSD severity score trajectories that did not include information on the complex sample design.

	Resilience	Recovery	Chronic
Variance	Posterior Mean (95% CrI)	Posterior Mean (95% CrI)	Posterior Mean (95% CrI)
Observation-level:			
σ_k^2	0.003 (0.002, 0.003)	0.015 (0.011, 0.021)	0.054 (0.035, 0.076)
Subject-level:			
ϕ_{11k}	0.006 (0.004, 0.007)	0.049 (0.034, 0.066)	0.082 (0.048, 0.124)
ϕ_{12k}	0 (-0.001, 0.001)	-0.005 (-0.013, 0.003)	0.046 (0.023, 0.075)
ϕ_{13k}	0 (-0.001, 0.001)	-0.003 (-0.01, 0.004)	0.01 (-0.013, 0.034)
ϕ_{22k}	0.004 (0.003, 0.005)	0.018 (0.011, 0.027)	0.065 (0.036, 0.1)
ϕ_{23k}	0 (0, 0.001)	0.001 (-0.003, 0.006)	0.019 (-0.002, 0.043)
ϕ_{33k}	0.005 (0.004, 0.006)	0.014 (0.009, 0.021)	0.053 (0.025, 0.088)

Appendix D

MCMC Algorithm for the Bayesian Shared Parameter Model in Electronic Health Records

I explicate the MCMC algorithm for fitting the proposed shared parameter model to EHRs. I provide the MCMC algorithm using a parametrization based on hierarchical centering in the longitudinal health outcomes model [Gelfand *et al.*, 1995; Gelfand *et al.*, 1996], in contrast to the parameterization in the main text. The hierarchically-centered parameterization is used in the `R` package `EHRMiss`. This parameterization is given as

$$\left[\begin{array}{c|c} \mathbf{y}_{1i} & \\ \vdots & c_i = k \\ \mathbf{y}_{Ri} & \end{array}\right] \sim MVN_{RJ}\left(\begin{bmatrix} \beta_{1k}\mathbf{x}_i^{h,T} + \mathbf{b}_{1i}\mathbf{z}_i^T \\ \vdots \\ \beta_{Rk}\mathbf{x}_i^{h,T} + \mathbf{b}_{Ri}\mathbf{z}_i^T \end{bmatrix}, diag(\boldsymbol{\Sigma}_k)\right) \tag{D.1}$$

$$\left[\begin{array}{c|c} \mathbf{b}_{1i} & \\ \vdots & c_i = k \\ \mathbf{b}_{Ri} & \end{array}\right] \sim MVN_{Rq}\left(\begin{bmatrix} \mathbf{u_i}\eta_{1k}^T \\ \vdots \\ \mathbf{u_i}\eta_{Rk}^T \end{bmatrix}, \boldsymbol{\Psi}_k\right) \tag{D.2}$$

where I use a superscript h for the fixed effects design matrix $\mathbf{x}_i^h$ $(J \times p^h)$ to indicate the change in parameterization. Unlike the main text, in (D.1), the columns in the random

effects design matrix $\mathbf{z}_i$ are no longer a subset of the columns in $\mathbf{x}_i^h$. For example, in a random intercept model, only $\mathbf{z}_i$ will include a column of ones for an intercept. In (D.2), the random effects $\mathbf{b}_{1i}, \ldots, \mathbf{b}_{Ri}$ are distributed with mean as a function of patient-level risk factors in $\mathbf{u}_i$ $(1 \times e)$ and corresponding regression coefficients in $\eta_{1k}, \ldots, \eta_{Rk}$ $(q \times e)$. $diag(\mathbf{\Sigma}_k)$ is an $RJ \times RJ$ block diagonal matrix with elements $\mathbf{\Sigma}_k$ for the variance-covariance among $y_{1ij}, \ldots, y_{Rij}$ in each time window j $(j = 1, \ldots, J)$.

D.0.1 Update parameters in the latent class membership model

The Gibbs steps are given for the latent class membership model.

1. Update ξ_{ik}. Let $\xi_i^T = (\xi_{i1}, \ldots, \xi_{iK-1})$ be a $(K-1)$-length column vector. Per [McCulloch and Rossi, 1994], for $i = 1, \ldots, n$, the distribution of $\xi_i \mid \delta, c_i$ is a $(K-1)$-variate normal distribution truncated over the appropriate cone in $\mathbf{R}^{K-1}$. Let $\mathbf{c}_i^*$ be a multinomial vector with entries $\mathbf{c}_i^* = (c_{i1}^*, \ldots, c_{iK}^*)$ equal to 1 if the i^{th} subject is in latent class k and 0 otherwise. If $c_{ik}^* = 1$, then $\xi_{ik} > max(\xi_{i,-k}, 0)$. If $c_{ik}^* = 0$, then $\xi_{ik} < max(\xi_{i,-k}, 0)$. $\xi_{i,-k}$ is a $K-2$ dimensional vector of all components of ξ_i excluding ξ_{ik}. This algorithm avoids the problem of drawing from a truncated multivariate normal. Instead each draw is a truncated univariate normal because I am using the conditional distribution $\xi_{ik} \mid \xi_{i,-k}, \delta_k, c_i$, where $c_i = K$ if $max(\xi_i) < 0$, or else $c_i =$ index of $max(\xi_i)$ for $k = 1, \ldots, K-1$.

2. Update δ_k. For $k = 1, \ldots, K-1$, I assume the prior $\delta_k \sim MVN_s(0, \mathbf{\Sigma}_\delta)$. The full

conditional is $MVN_s(\mu_{\delta_k}, \mathbf{V}_\delta)$, where

$$\mathbf{V}_\delta = \left(\sum_{i=1}^{n} \mathbf{w}_i^T \mathbf{w}_i + \mathbf{\Sigma}_\delta^{-1} \right)^{-1}$$
$$\mu_{\delta_k} = \mathbf{V}_\delta \times \left(\sum_{i=1}^{n} \mathbf{w}_i^T \xi_{ik} \right),$$

with $\mathbf{w}_i$ being an s-length row vector of patient-level risk factors, including a column of ones for an intercept.

D.0.2 Update parameters in the longitudinal outcomes model

1. Update β_{rk}.

   To update β_{rk}, based on the properties of the multivariate normal distribution, I use the conditional distribution of longitudinal health outcome r given health outcomes $r^{'}$ for all $r^{'} \neq r$. Let $\mathbf{y}^*_{ri} = (y_{riA(1)}, \ldots, y_{riA(n_i)})^T$. Let $\mathbf{Q}$ be a matrix of conditional coefficients defined as $\mathbf{Q} = \mathbf{I} - [diag(\mathbf{\Sigma}_k^{-1})]^{-1}\mathbf{\Sigma}_k^{-1}$, with elements $q_{rr'}$ $(r = 1, \ldots, R,$ $r^{'} = 1, \ldots, R)$ [Gelman *et al.*, 2014]. For longitudinal health outcome r of patient i in window j, the conditional distribution of $\mathbf{y}^*_{ri}$ given $\mathbf{y}^*_{r^{'}i}$ for all $r^{'} \neq r$ and latent class c_i is

   $$[\mathbf{y}^*_{ri} \,|\, \mathbf{y}^*_{r^{'}i} \text{ all } r^{'} \neq r,\, c_i = k] \sim \tag{D.3}$$
   $$MVN_{n_i}\left(\beta_{rk}\mathbf{x}_i^{h*,T} + \mathbf{b}_{ri}\mathbf{z}_i^{*,T} + \sum_{r' \neq r} q_{rr'}(\mathbf{y}^*_{r'i} - \beta_{r'k}\mathbf{x}_i^{h*,T} - \mathbf{b}_{r'i}\mathbf{z}_i^{*,T}),\, diag\left([\mathbf{\Sigma}^{-1}_{krr}]^{-1}\right) \right),$$

   where $\mathbf{x}_i^{h*}$ $(n_i \times p^h)$ is the fixed effects design matrix for time windows $A(l)$ for $l = 1, \ldots, n_i$. $\mathbf{z}_i^*$ is the corresponding random effects design matrix.

   For latent classes $k = 1, \ldots, K$, assuming the prior distribution $\beta_{rk} \sim MVN_{p^h}(\mathbf{0}, \mathbf{\Sigma}_\beta)$, the full conditional is $MVN_{p^h}(\mu_{\beta_{rk}}, \mathbf{V}_{\beta_{rk}})$, where

   $$\mathbf{V}_{\beta_{rk}} = \left(\sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times \frac{\mathbf{x}_i^{h*,T}\mathbf{x}_i^{h*}}{[\mathbf{\Sigma}^{-1}_{krr}]^{-1}} + \mathbf{\Sigma}_\beta^{-1} \right)^{-1}$$

$\mu_{\beta_{rk}}$

$$= \mathbf{V}_{\beta_{rk}} \times \left(\sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times \frac{\mathbf{x}_i^{h*,T} \left(\mathbf{y}_{ri}^{*,T} - \mathbf{z}_i^* \mathbf{b}_{ri}^T - (\sum_{r' \neq r} q_{rr'} (\mathbf{y}_{r'i}^* - \beta_{r'k} \mathbf{x}_i^{h*,T} - \mathbf{b}_{r'i} \mathbf{z}_i^{*,T}))^T \right)}{[\mathbf{\Sigma}_{krr}^{-1}]^{-1}} \right)$$

2. Update $\mathbf{b}_{ri}$. Using the conditional distribution in (D.3), the full conditional is $MVN_q(\mu_{b_{ri}}, \mathbf{V}_{b_{ri}})$, where

$$\mathbf{V}_{b_{ri}} = \sum_{k=1}^{K} \mathbf{1}_{c_i=k} \left(\frac{\mathbf{z}_i^{*,T} \mathbf{z}_i^*}{[\mathbf{\Sigma}_{krr}^{-1}]^{-1}} + \mathbf{\Psi}_{kr}^{-1} \right)^{-1}$$

$\mu_{b_{ri}}$

$$= \mathbf{V}_{b_{ri}} \times \sum_{k=1}^{K} \mathbf{1}_{c_i=k} \times \left(\frac{\mathbf{z}_i^{*,T} \left(\mathbf{y}_{ri}^{*,T} - \mathbf{x}_i^{h*} \beta_{rk}^T - (\sum_{r' \neq r} q_{rr'} (\mathbf{y}_{r'i}^* - \beta_{r'k} \mathbf{x}_i^{h*,T} - \mathbf{b}_{r'i} \mathbf{z}_i^{*,T}))^T \right)}{[\mathbf{\Sigma}_{krr}^{-1}]^{-1}} + \mathbf{\Psi}_{kr}^{-1} \eta_{rk} \mathbf{u}_i^T \right)$$

3. Update η_{rk}. Let the elements of $\mathbf{b}_{ri}$ be indexed as b_{rig} for $g = 1, \ldots, q$. For the g^{th} random effect, let $\eta_{rkg} = (\eta_{rkg1}, \ldots, \eta_{rkge})^T$ $(1 \times e)$. Then, $b_{rig} \sim N(\mathbf{u}_i \eta_{rkg}^T, \psi_{krgg})$. Assuming the prior distribution $MVN_e(\mathbf{0}, \mathbf{\Sigma}_\eta)$, the full conditional of η_{rkg} is $MVN_e(\mu_{\eta_{rkg}}, \mathbf{V}_{\eta_{rkg}})$, where

$$\mathbf{V}_{\eta_{rkg}} = \left(\sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times \frac{\mathbf{u}_i^T \mathbf{u}_i}{\psi_{krgg}} + \mathbf{\Sigma}_\eta^{-1} \right)^{-1}$$

$$\mu_{\eta_{rkg}} = \mathbf{V}_{\eta_{rkg}} \times \left(\sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times \frac{\mathbf{u}_i^T b_{rig}}{\psi_{krgg}} \right)$$

4. Update $\mathbf{\Sigma}_k$. Recall the R-length row vectors $\mathbf{y}_{iA(l)} = (y_{1iA(l)}, \ldots, y_{RiA(l)})^T$, and $\mu_{iA(l)} = \mathbf{x}_{iA(l)} \beta_k^T + \mathbf{z}_{iA(l)} \mathbf{b}_i^T$. Assuming an inverse-Wishart prior distribution $\mathbf{\Sigma}_k \sim$

$IW(\nu_\Sigma, S_\Sigma^{-1})$, the full conditional is $IW(a_{\Sigma_k}, b_{\Sigma_k})$, where

$$a_{\Sigma_k} = \nu_\Sigma + \sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times n_i$$

$$b_{\Sigma_k} = S_\Sigma + \sum_{i=1}^{n} \mathbf{1}_{c_i=k} \sum_{l=1}^{n_i} (\mathbf{y}_{iA(l)} - \mu_{iA(l)})^T (\mathbf{y}_{iA(l)} - \mu_{iA(l)})$$

5. Update $\mathbf{\Psi}_k$. The block diagonal matrix $\mathbf{\Psi}_k$ $(Rq \times Rq)$ contains elements $\mathbf{\Psi}_{kr}$ $(q \times q)$. Assuming $\mathbf{\Psi}_{kr} \sim IW(\nu_\Psi, S_\Psi^{-1})$, the full conditional is $IW(a_{\Psi_{kr}}, b_{\Psi_{kr}})$, where

$$a_{\Psi_{kr}} = \nu_\Psi + \sum_{i=1}^{n} \mathbf{1}_{c_i=k}$$

$$b_{\Psi_{kr}} = S_\Psi + \sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times (\mathbf{b}_{ri} - \mathbf{u}_i \eta_{rk}^T)^T (\mathbf{b}_{ri} - \mathbf{u}_i \eta_{rk}^T)$$

D.0.3 Update parameters in the visit process model

Following [Albert and Chib, 1993], I use a data augmentation approach [Tanner and Wong, 1987] to model the probability of a clinic visit using Bayesian probit regression. Corresponding to the visit process for patient i in clinical window j, I introduce latent variables ξ_{ij}^d $(i = 1, \ldots, n,\ j = 1, \ldots, J)$. The latent variables ξ_{ij}^d are assumed to be distributed as $N(\mathbf{x}_{ij}\phi_k^T + \mathbf{z}_{ij}\tau_i^T, 1)$, where the observation-level error variance is fixed to 1. To connect latent ξ_{ij}^d to the visit process d_{ij}, define $d_{ij} = 1$ if $\xi_{ij}^d > 0$ and $d_{ij} = 0$ if $\xi_{ij}^d \leq 0$. With the introduction of the latent variables, the Gibbs sampling steps are as follows.

1. Update ξ_{ij}^d. The full conditional is $\xi_{ij}^d \,|\, d_{ij}, \phi_k, \tau_i, c_i = k \sim N(\sum_{k=1}^{K} \mathbf{1}_{c_i=k} \times (\mathbf{x}_{ij}\phi_k^T + \mathbf{z}_{ij}\tau_i^T), 1)$, truncated at the left by 0 if $d_{ij} = 1$. Otherwise, $\xi_{ij}^d \,|\, d_{ij}, \phi_k, \tau_i, c_i = k \sim N(\sum_{k=1}^{K} \mathbf{1}_{c_i=k} \times (\mathbf{x}_{ij}\phi_k^T + \mathbf{z}_{ij}\tau_i^T), 1)$, truncated at the right by 0 if $d_{ij} = 0$.

2. Update ϕ_k. For latent classes $k = 1, \ldots, K$, assuming the prior distribution $\phi_k \sim$

$MVN_p(\mu_\phi, \mathbf{\Sigma}_\phi)$, the full conditional is $MVN_p(\mu_{\phi_k}, \mathbf{V}_{\phi_k})$, where

$$\mathbf{V}_{\phi_k} = \left(\sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times \mathbf{x}_i^T \mathbf{x}_i + \mathbf{\Sigma}_\phi^{-1} \right)^{-1}$$

$$\mu_{\phi_k} = \mathbf{V}_{\phi_k} \times \left(\sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times \mathbf{x}_i^T \left(\xi_i^{d,T} - \mathbf{z}_i \tau_i^T \right) + \mathbf{\Sigma}_\phi^{-1} \mu_\phi^T \right),$$

where the random effects design matrix $\mathbf{z}_i$ $(J \times q)$ contains a subset of the columns in the fixed effects design matrix $\mathbf{x}_i$ $(J \times p)$.

3. Update τ_i. The full conditional is $MVN_q(\mu_{\tau_i}, \mathbf{V}_{\tau_i})$, where

$$\mathbf{V}_{\tau_i} = \sum_{k=1}^{K} \mathbf{1}_{c_i=k} \times \left(\mathbf{z}_i^T \mathbf{z}_i + \mathbf{\Omega}_k^{-1} \right)^{-1}$$

$$\mu_{\tau_i} = \mathbf{V}_{\tau_i} \times \sum_{k=1}^{K} \mathbf{1}_{c_i=k} \times \left(\mathbf{z}_i^T \left(\xi_i^{d,T} - \mathbf{x}_i \phi_k^T \right) \right)$$

4. Update $\mathbf{\Omega}_k$. Assuming an inverse-Wishart prior distribution $\mathbf{\Omega}_k \sim IW(\nu_\Omega, S_\Omega^{-1})$, the full conditional is $IW(a_{\Omega_k}, b_{\Omega_k})$, where

$$a_{\Omega_k} = \nu_\Omega + \sum_{i=1}^{n} \mathbf{1}_{c_i=k}$$

$$b_{\Omega_k} = S_\Omega + \sum_{i=1}^{n} \mathbf{1}_{c_i=k} \times \tau_i^T \tau_i$$

D.0.4 Update parameters in the response process given a clinic visit model

The Gibbs steps to update the parameters in the model for the response process given a clinic visit are analogous to the steps in the visit process model, except that I use observed clinic visits.

For patient i in clinical window l with an observed visit ($l = 1, \ldots, n_i$), I introduce latent variables $\xi_{riA(l)}^m$ $(i = 1, \ldots, n,\ l = 1, \ldots, n_i)$. The latent variables $\xi_{riA(l)}^m$ are assumed to be distributed as $N(\mathbf{x}_{iA(l)} \lambda_{rk}^T + \mathbf{z}_{iA(l)} \kappa_{ri}^T, 1)$, where the observation-level error variance is

fixed to 1. To connect latent $\xi^m_{riA(l)}$ to the response process $m_{riA(l)}$, define $m_{riA(l)} = 1$ if $\xi^m_{riA(l)} > 0$ and $m_{riA(l)} = 0$ if $\xi^m_{riA(l)} \leq 0$. Upon introducing the latent variables, the Gibbs sampling steps for λ_{rk}, κ_{ri}, and $\mathbf{\Theta}_{rk}$ proceed as in the visit process model.

D.0.5 Update latent class membership

Sample latent class indicators c_i for $i = 1, \ldots, n$ from $Multinomial(1; p_{i1}, \ldots, p_{iK})$, where $p_{i1}, \ldots, p_{iK}$ are the posterior probabilities of latent class assignment. For $k = 1, \ldots, K$,

$$
\begin{aligned}
& p_{ik} \\
&= Pr(c_i = k \,|\, \pi_{ik};\, \mathbf{y}^*_i, \mathbf{b}_i;\, \mathbf{d}_i, \tau_i;\, \mathbf{m}_{1i}, \ldots, \mathbf{m}_{Ri}, \kappa_{1i}, \ldots, \kappa_{Ri};\, rest) \\
&\propto \pi_{ik}\, f(\mathbf{y}^*_i \,|\, \mathbf{b}_i, \beta_k, \mathbf{\Sigma}^*_k)\, f(\mathbf{b}_i \,|\, \mathbf{\Psi}_k) \\
&\times f(\mathbf{d}_i \,|\, \tau_i, \phi_k)\, f(\tau_i \,|\, \mathbf{\Omega}_k) \\
&\times \prod_{r=1}^{R} f(\mathbf{m}_{ri} \,|\, \kappa_{ri}, \lambda_{rk})\, f(\kappa_{ri} \,|\, \mathbf{\Theta}_{rk}),
\end{aligned}
$$

where $\mathbf{y}^*_i = (\mathbf{y}^T_{iA(1)}, \ldots, \mathbf{y}^T_{iA(n_i)})$, and $\mathbf{\Sigma}^*_k$ is an $n_i R \times n_i R$ block diagonal matrix with elements $\mathbf{\Sigma}_k$ $(R \times R)$ for each $\mathbf{y}_{iA(l)}$ $(l = 1, \ldots, n_i)$.

Appendix E

Addendum to the Analysis of Weight and Height Z-scores

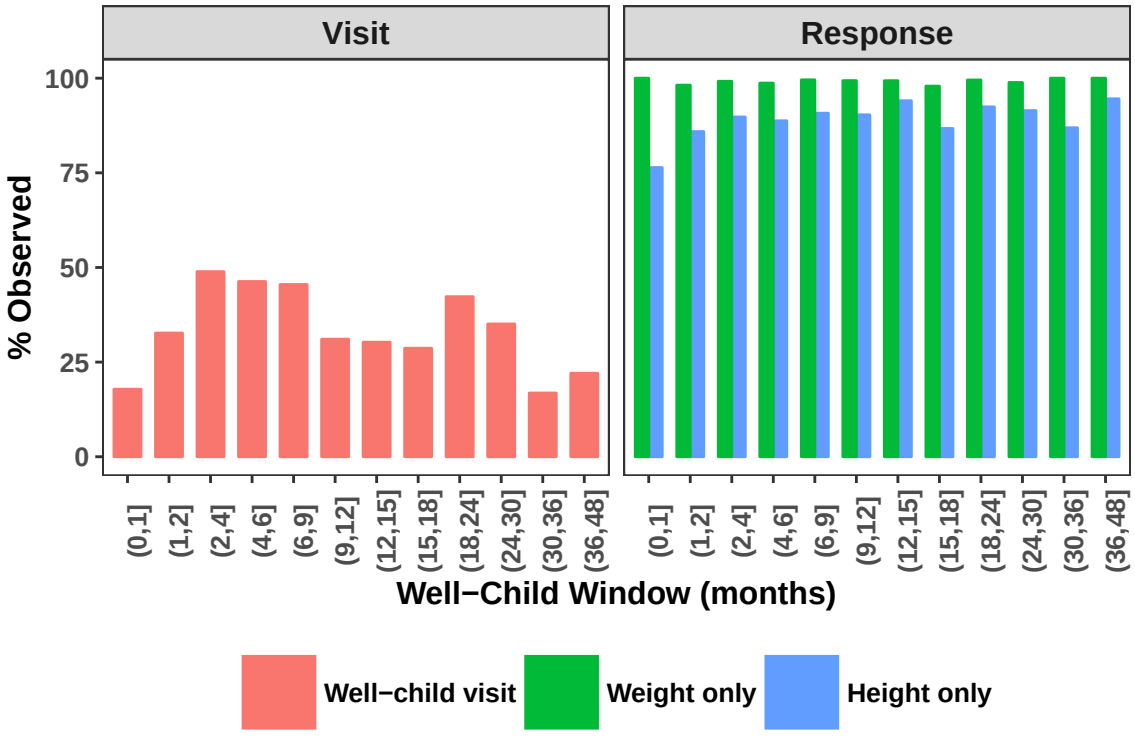

Figure E.1: Patterns of missed visits and missed responses in weight and height z-scores given a clinic visit.

E.0.1 Model selection for the MNAR and MAR methods

Table E.1: Comparison of model information criteria among models with up to $K = 3$ latent classes using the **MAR** and **MNAR** methods.

	MAR			**MNAR**	
	K			K	
Criterion	1	2	3	2	3
BIC	11854	**10978**	12114	21469	**21104**
DIC3	12093	**11384**	13673	23087	**22483**

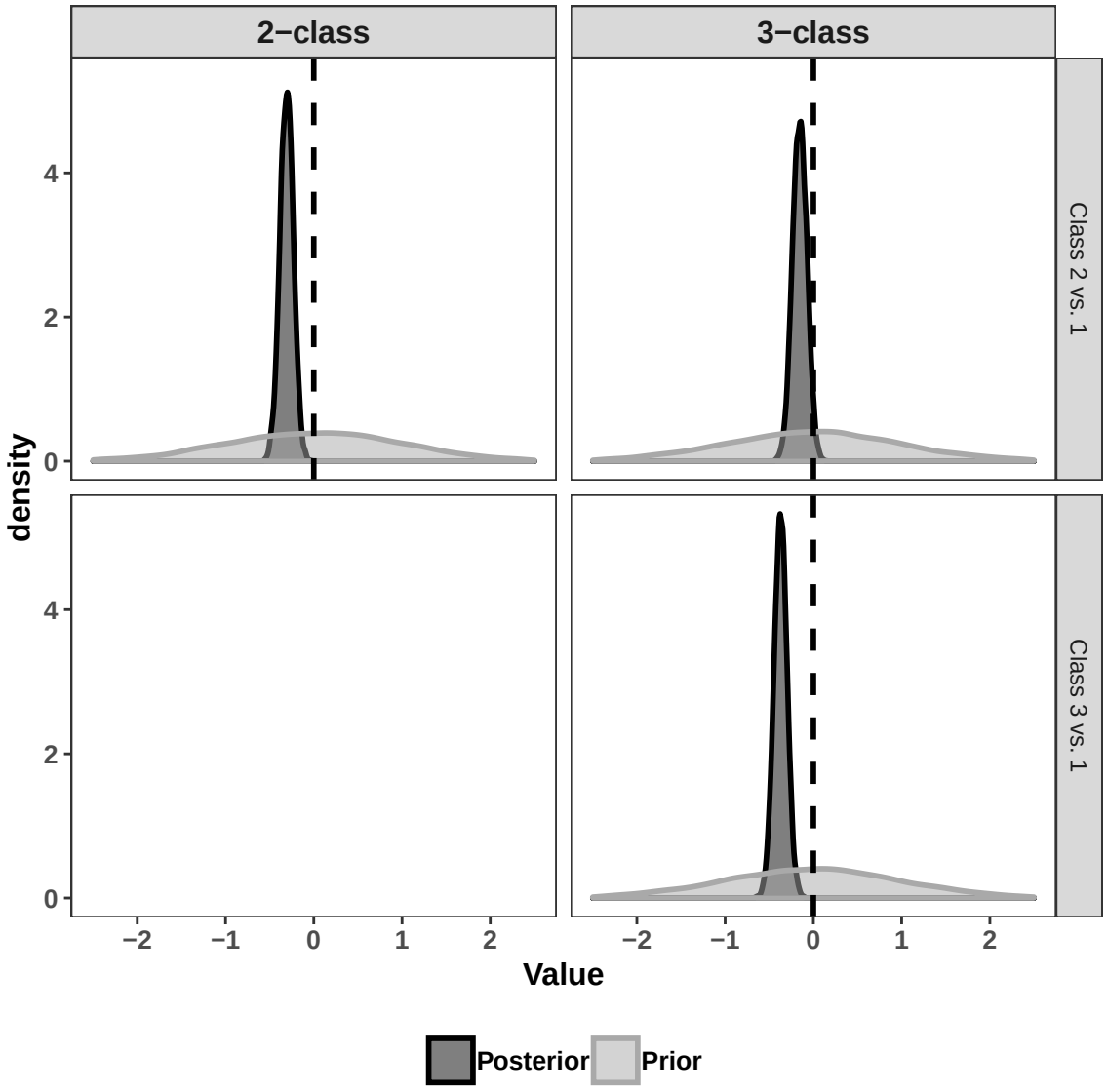

Figure E.2: Posterior versus prior distributions for the intercepts in the multinomial probit model of latent class membership using the **MAR** method, $K = 2, 3$.

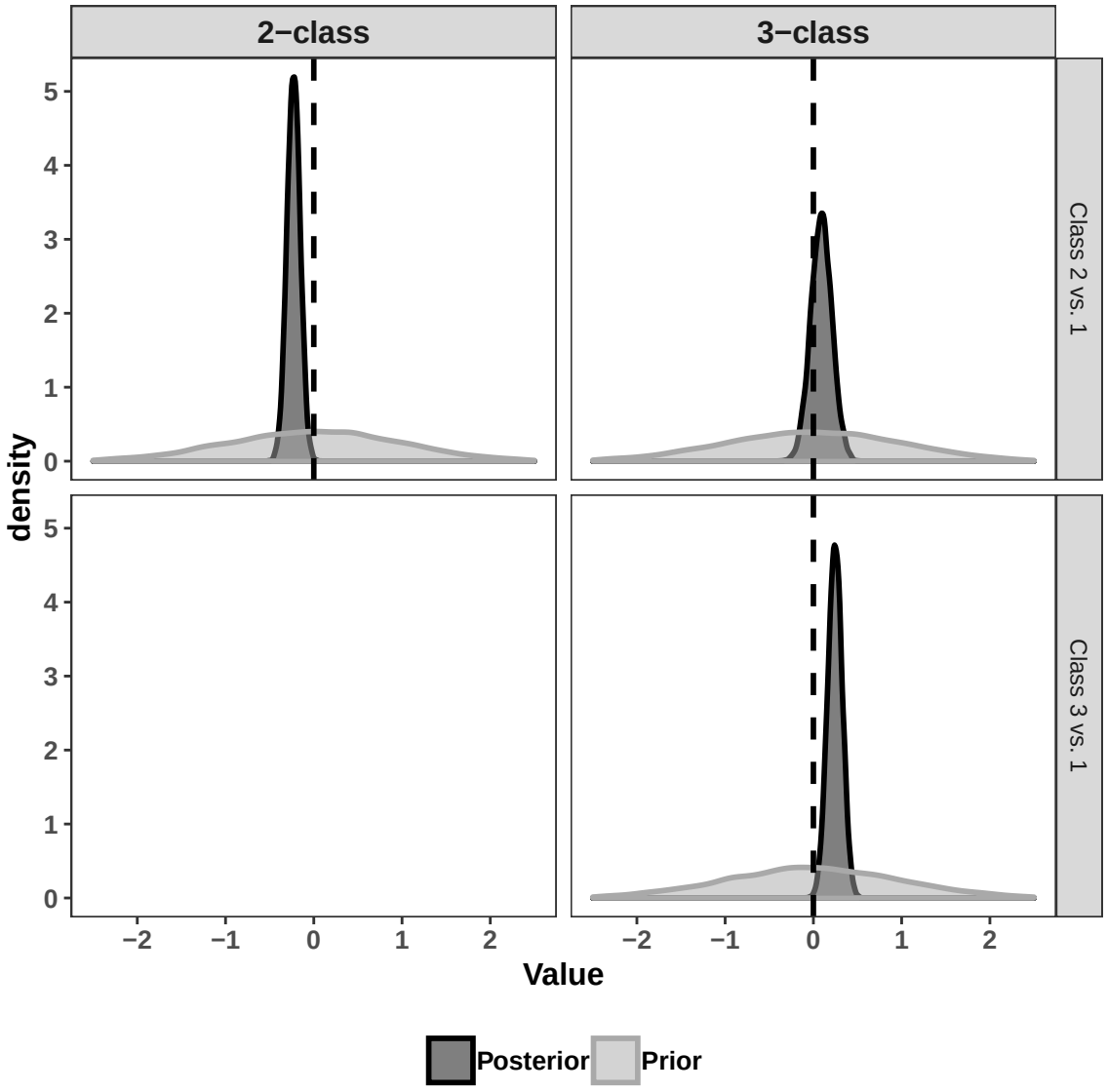

Figure E.3: Posterior versus prior distributions for the intercepts in the multinomial probit model of latent class membership using the **MNAR**, $K = 2, 3$.

E.0.2 Sensitivity analysis for the 2 and 3-latent class models

E.0.2.1 2-latent class models.

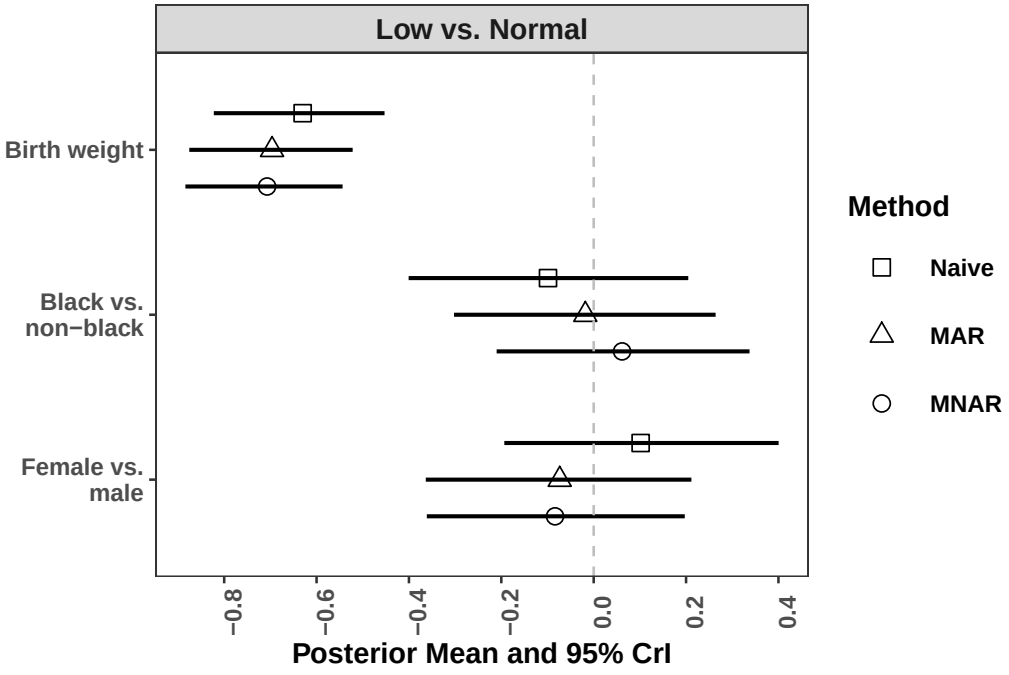

Figure E.4: Regression coefficients for predictors in the multinomial probit model of latent class membership in the **Naïve**, **MAR**, and **MNAR** methods, assuming 2 latent classes. Birth weight was inversely associated with probability of belonging to the Low versus Normal subgroup, while race and sex were not related to probability of latent class membership.

Table E.2: Posterior latent class assignment in the $K = 2, 3$-class models based on assigning children to a trajectory subgroup according to the maximum of the mean posterior probabilities of class assignment. The **Naïve**, **MAR**, and **MNAR** methods are shown.

	$K = 2$		$K = 3$		
	Normal	Low	Normal, increasing	Normal, decreasing	Low
Naïve ($n = 471$):					
Predicted class size (%)	307 (65)	164 (35)	197 (42)	163 (35)	111 (24)
Mean probability	0.87	0.88	0.83	0.79	0.91
Median probability	0.92	0.98	0.89	0.82	0.99
MAR ($n = 499$):					
Predicted class size (%)	335 (67)	164 (33)	192 (38)	185 (37)	122 (24)
Mean probability	0.90	0.91	0.82	0.82	0.91
Median probability	0.96	0.99	0.87	0.88	1
MNAR ($n = 499$):					
Predicted class size (%)	295 (59)	204 (41)	159 (32)	165 (33)	175 (35)
Mean probability	0.93	0.87	0.82	0.85	0.84
Median probability	0.99	0.95	0.84	0.94	0.95

Table E.3: Cross-classification of 499 children assigned to the Normal and Low trajectory subgroups by the **MAR** and **MNAR** methods, according to latent class assignment and low birth weight (LBW) status.

	MNAR			
	Non-LBW		LBW	
	Normal	Low	Normal	Low
MAR				
Normal	258	50	16	11
Low	18	57	3	86

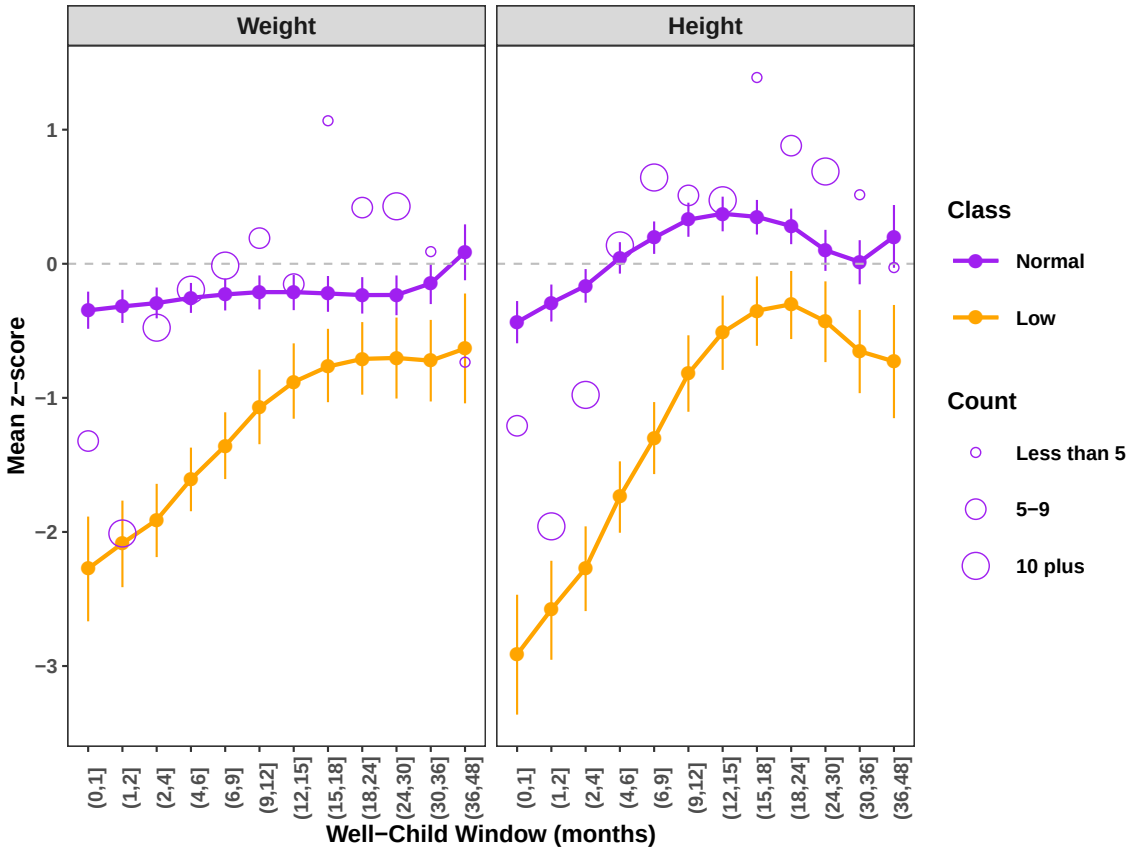

Figure E.5: Sample means of observed weight and height z-scores (hollow circles) in each well-child window among the 18 non-low birth weight children moved from the Low trajectory subgroup in the **MAR** method to the Normal trajectory subgroup in the **MNAR** method, assuming 2 latent classes. The size of the point indicates the number of observations contributing to the sample mean. Overlaid are the average latent class-specific z-score trajectories estimated by the MNAR method.

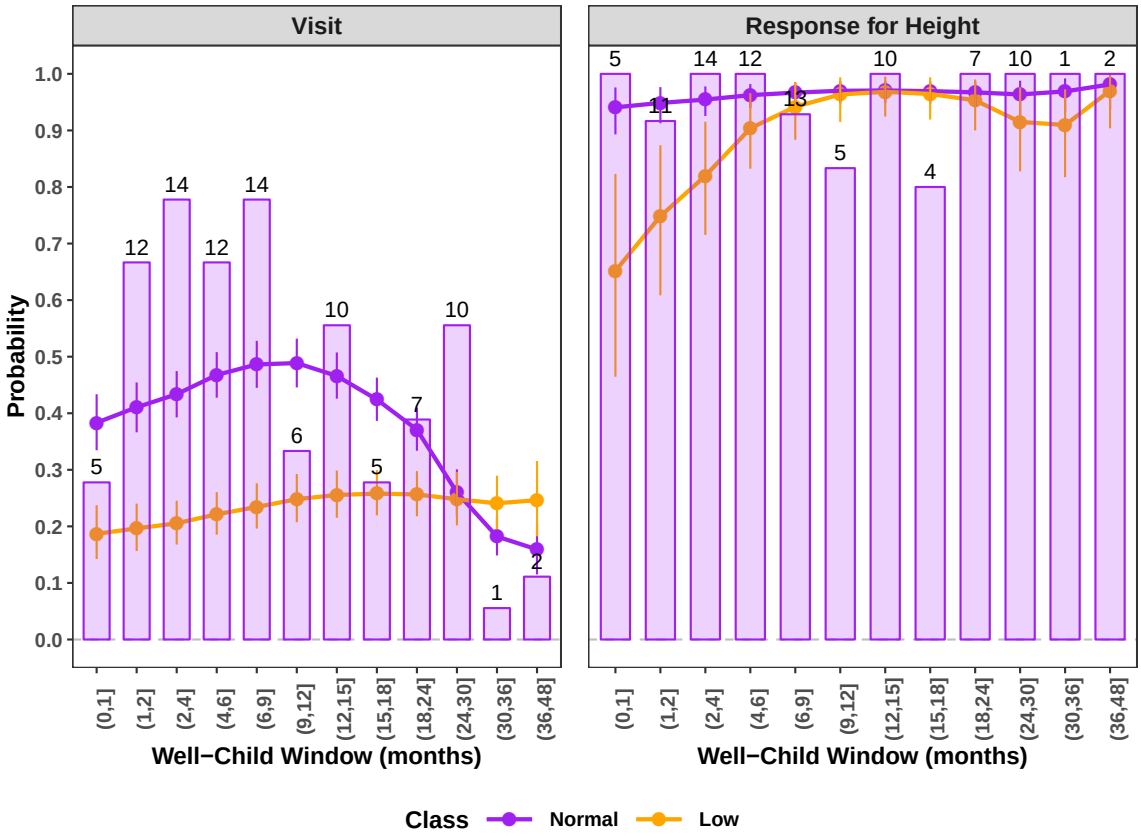

Figure E.6: Bar plots of the observed proportions of children with a clinic visit, and the observed proportions of children with a height response, among the 18 non-low birth weight children moved from the Low trajectory subgroup in the **MAR** method to the Normal trajectory subgroup in the **MNAR** method. In the Visit panel, the number of children with a clinic visit in each window is provided. In the Response for Height panel, the number of children with a height response (given a clinic visit) is given. Overlaid are the latent class-specific visit and response trajectories estimated by the **MNAR** method assuming 2 latent classes.

E.0.2.2 3-latent class models.

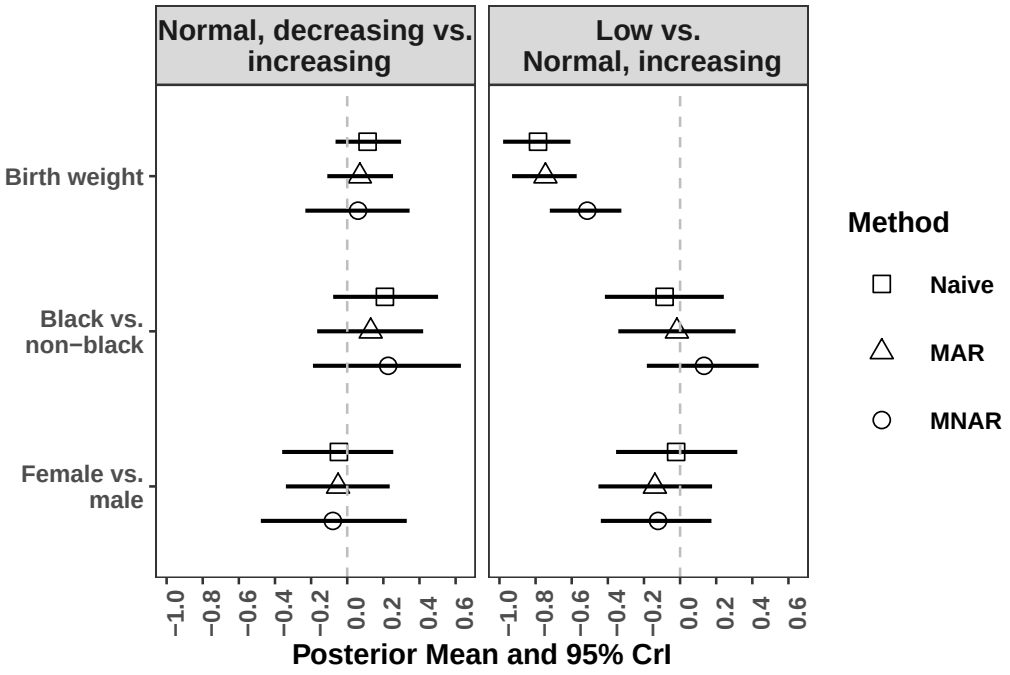

Figure E.7: Regression coefficients for predictors in the multinomial probit model of latent class membership in the **Naïve**, **MAR**, and **MNAR** methods, assuming 3 latent classes. Birth weight is inversely associated with probabilty of belonging to the Low versus Normal, increasing subgroup.

Table E.4: Cross-classification of 499 children assigned to the Normal, increasing; Normal, decreasing, and Low trajectory subgroups by the **MAR** and **MNAR** methods, according to latent class assignment and low birth weight (LBW) status.

	MNAR					
	Non-LBW			LBW		
	Normal, increasing	Normal, decreasing	Low	Normal, increasing	Normal, decreasing	Low
MAR						
Normal, increasing	120	30	19	14	1	8
Normal, decreasing	18	121	26	1	11	8
Low	5	2	42	1	0	72

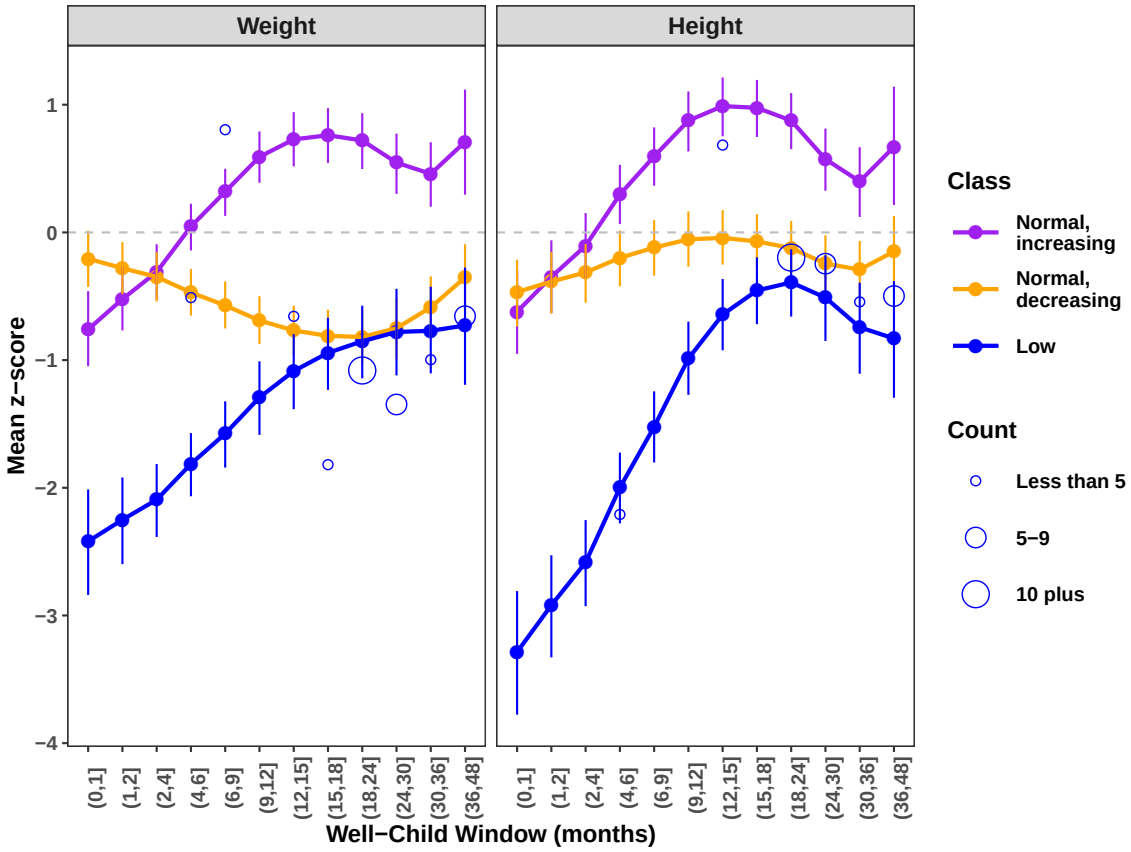

Figure E.8: Sample means of observed weight and height z-scores (hollow circles) in each well-child window among the 26 non-low birth weight children moved from the Normal, decreasing trajectory subgroup in the **MAR** method to the Low trajectory subgroup in the **MNAR** method, assuming 3 latent classes. The size of the point indicates the number of observations contributing to the sample mean. Overlaid are the average latent class-specific z-score trajectories estimated by the **MNAR** method.

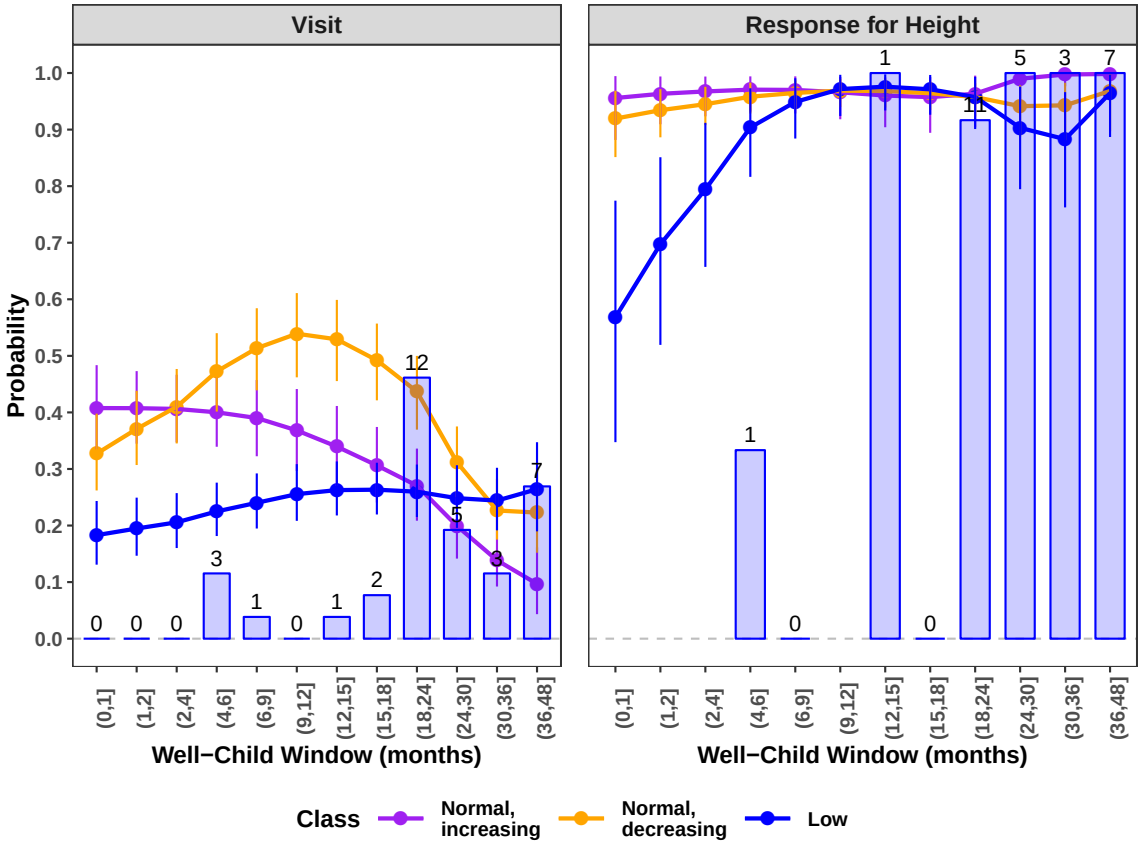

Figure E.9: Bar plots of the observed proportions of children with a clinic visit, and the observed proportions of children with a height response, among the 26 non-low birth weight children moved from the Normal, decreasing trajectory subgroup in the **MAR** method to the Low trajectory subgroup in the **MNAR** method. In the Visit panel, the number of children with a clinic visit in each window is provided. In the Response for Height panel, the number of children with a height response (given a clinic visit) is given. Overlaid are the latent class-specific visit and response trajectories estimated by the **MNAR** method assuming 3 latent classes.

E.0.3 Model checking

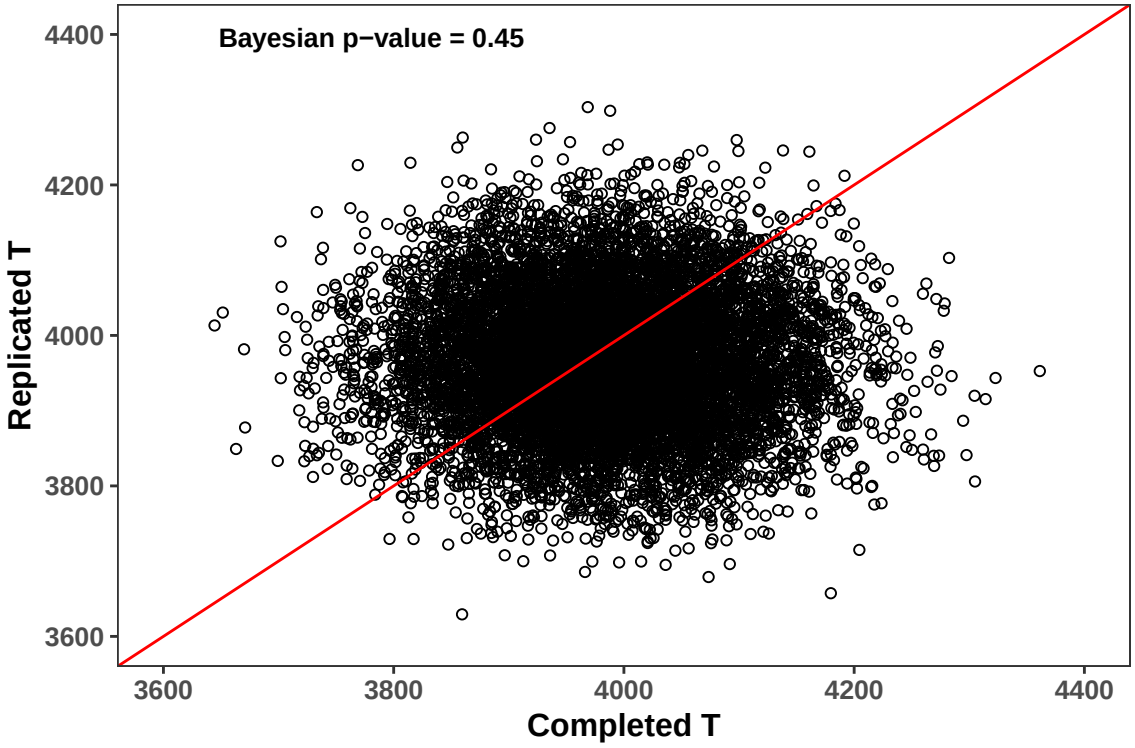

Figure E.10: Posterior predictive checking for the 2-class model estimated using the **MNAR** method. Completed T is computed using the completed data. Replicated T is computed using the replicated completed datasets from the posterior predictive distribution.

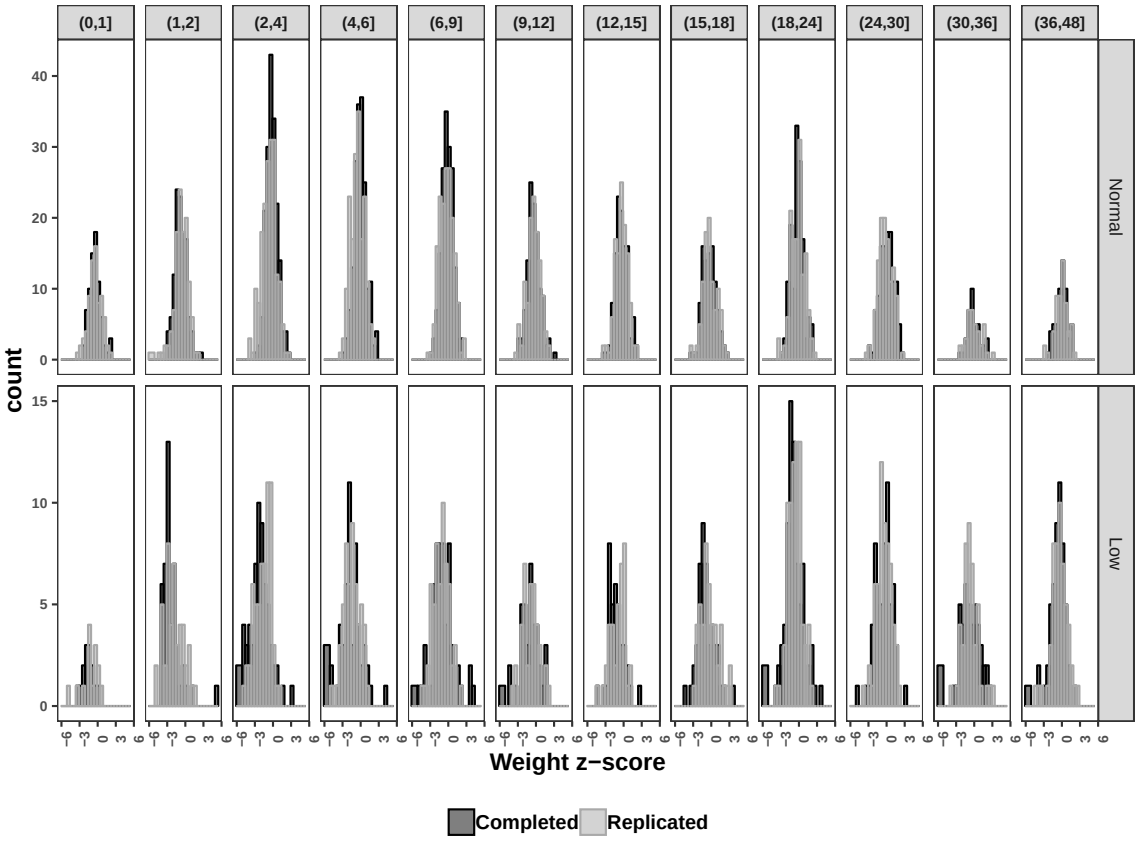

Figure E.11: Histograms of completed and replicated completed weight z-scores from the posterior predictive distribution, by subgroup and well-child window, assuming 2 latent classes and using the **MNAR** method.

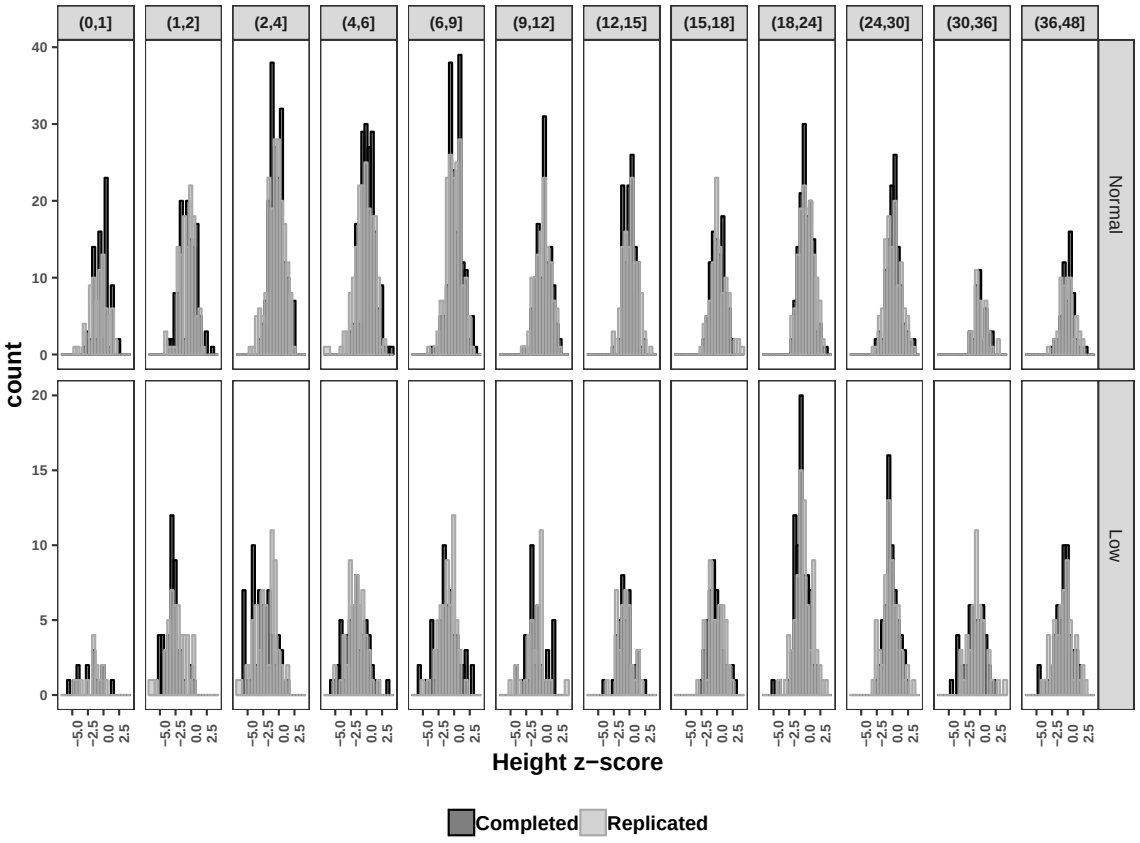

Figure E.12: Histograms of completed and replicated completed height z-scores from the posterior predictive distribution, by subgroup and well-child window, assuming 2 latent classes and using the **MNAR** method.

Appendix F

Addendum to the Simulation Study

F.1 Design

I designed the study based on the real data analysis with 2 latent classes estimated with the **MNAR** method. For 500 subjects, I generated longitudinal outcomes of interest y_{1ij} and y_{2ij} over 12 time windows, with about 60% and 40% of subjects in latent classes 1 and 2, respectively. I assumed the missing data mechanisms for the visit process and response process for y_{2ij} are MNAR, while y_{1ij} is fully observed given a clinic visit. In this setting, I considered the following five specific scenarios (S1-S5):

1. Under S1, I mimicked the latent class-specific trajectories and missingness proportions in the real data analysis. True parameter values for variance components were selected according to the real data analysis.

   First, I generated the latent class membership of subject i as

$$\left[\; c_i \;\middle|\; w_{1i} \; \right] \sim Bernoulli\left(\pi_{i2}\right),$$

where using a probit link function, $\pi_{i2} = \Phi\{-0.25 - w_i\}$. π_{i2} is the probability that subject i belongs to latent class 2, and w_i is a scaled and centered simulated variable for a subject's birth weight.

Then, I generated the longitudinal outcomes, visit process, and response process given a clinic visit conditional on a subject's latent class membership as

$$\left[\begin{array}{c} y_{1ij} \\ y_{2ij} \end{array} \middle| \; c_i = k \right] \sim MVN_2 \left(\begin{bmatrix} \beta_{1k1} + \beta_{1k2} x_{ij} + b_{1i1} \\ \beta_{2k1} + \beta_{2k2} x_{ij} + b_{2i1} \end{bmatrix}, \boldsymbol{\Sigma}_k \right) \tag{F.1}$$

$$\left[\begin{array}{c} b_{1i1} \\ b_{2i1} \end{array} \middle| \; c_i = k \right] \sim MVN_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Psi_{k1} & 0 \\ 0 & \Psi_{k2} \end{bmatrix} \right) \tag{F.2}$$

$$\left[\; d_{ij} \; \middle| \; c_i = k \; \right] \sim Bernoulli \Big(\Phi\{\phi_{k1} + \phi_{k2} x_{ij} + \tau_{i1}\} \Big) \tag{F.3}$$

$$\left[\; \tau_{i1} \; \middle| \; c_i = k \; \right] \sim Normal \Big(0, \, 0.25 \Big)$$

$$\left[\; m_{2i,A(l)} \; \middle| \; c_i = k \; \right] \sim Bernoulli \Big(\Phi\{\lambda_{2k1} + \lambda_{2k2} x_{iA(l)} + \kappa_{2i1}\} \Big) \tag{F.4}$$

$$\left[\; \kappa_{2i1} \; \middle| \; c_i = k \; \right] \sim Normal \Big(0, \, \Theta_{2k} \Big) \tag{F.5}$$

In (F.1), for y_{1ij}, in latent class 1, $\beta_{11} = (\beta_{111}, \, \beta_{112})^T = (-0.25, \, 0.1)$, and in latent class 2, $\beta_{12} = (\beta_{121}, \, \beta_{122})^T = (-1, \, 0.5)$. For y_{2ij}, $\beta_{21} = (\beta_{211}, \, \beta_{212})^T = (0.5, \, 0.2)$, and $\beta_{22} = (\beta_{221}, \, \beta_{222})^T = (-0.5, \, 0.75)$. The latent class-specific variance-covariances of y_{1ij}, y_{2ij} are $\boldsymbol{\Sigma}_1 = \left[\begin{smallmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{smallmatrix} \right]$ and $\boldsymbol{\Sigma}_2 = \left[\begin{smallmatrix} 1.5 & 1 \\ 1 & 1.5 \end{smallmatrix} \right]$. In (F.2), for the random intercept of y_{1ij}, the latent class-specific variances are $\Psi_{11} = \Psi_{21} = 0.6$. For y_{2ij}, $\Psi_{12} = 0.6$ and $\Psi_{22} = 0.4$.

For the visit process, in (F.3), $\phi_k = (\phi_{k1}, \, \phi_{k2})^T$, with $\phi_1 = (-0.2, \, -0.8)$ and $\phi_2 = (-0.8, \, 0.2)$.

For the response process of y_{2ij} given a clinic visit, in (F.4), $\lambda_{2k} = (\lambda_{2k1}, \, \lambda_{2k2})^T$,

with $\lambda_{21} = (1.9, 0.1)$ and $\lambda_{22} = (1.1, 0.25)$. The latent class-specific random intercept variances (F.5) are $\Theta_{21} = 0.5$ and $\Theta_{22} = 1.5$.

2. In S2, I modified S1 by increasing the difference in the slopes for the latent class-specific trajectories of y_{2ij} by making the slope in latent class 2 steeper. Specifically, in (F.1), $\beta_{222} = 1$. No other changes to S1 were made.

3. In S3, I modified S1 by decreasing the difference in the slopes for the latent class-specific trajectories of y_{2ij} by making the slope in latent class 2 nearly parallel to the latent class 1 slope. Specifically, in (F.1), $\beta_{222} = 0.3$. No other changes to S1 were made.

4. In S4, I altered the visit process of S1 to reduce the percent of missed clinic visits in each latent class whilst maintaining the general visit process trajectory. In (F.3), I set $\phi_1 = (0.4, -0.2)$ and $\phi_2 = (-0.1, 0.9)$. These changes resulted in 35% missed clinic visits in latent class 1, and 55% missed clinic visits in latent class 2. No other changes to S1 were made.

5. In S5, I modified S1 by increasing the percent of missed responses of y_{2ij} in each latent class whilst maintaining the general response process trajectory. In (F.4), I set $\lambda_{21} = (0.8, 0.1)$ and $\lambda_{22} = (0.5, 0.2)$. These changes resulted in 25% missed responses in y_{2ij} in latent class 1, and 35% missed responses in y_{2ij} in latent class 2. No other changes to S1 were made.

F.2 Results from the simulation study

Table F.1: Simulation results of S2 for parameter estimation of intercept β_{rk1} and slope β_{rk2} for longitudinal outcome r in latent class k, and latent class-level weights π_k under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

Parameter	Method	Truth	Mean	Bias	MSE	Coverage	Length
β_{111}	Full	-0.250	-0.248	0.002	0.002	0.960	0.188
	Naïve		-0.232	0.018	0.004	0.930	0.217
	MAR		-0.244	0.006	0.003	0.948	0.212
	MNAR		-0.245	0.005	0.003	0.946	0.209
β_{121}	Full	-1.000	-1.003	-0.003	0.003	0.952	0.228
	Naïve		-0.995	0.005	0.014	0.922	0.404
	MAR		-1.010	-0.010	0.009	0.946	0.369
	MNAR		-0.994	0.006	0.007	0.932	0.314
β_{112}	Full	0.100	0.100	-0.000	0.000	0.940	0.048
	Naïve		0.092	-0.008	0.001	0.936	0.096
	MAR		0.094	-0.006	0.001	0.950	0.092
	MNAR		0.100	0.000	0.001	0.946	0.091
β_{122}	Full	0.500	0.499	-0.001	0.001	0.946	0.096
	Naïve		0.445	-0.055	0.009	0.854	0.273
	MAR		0.480	-0.020	0.005	0.928	0.242
	MNAR		0.495	-0.005	0.003	0.940	0.214
β_{211}	Full	0.500	0.500	0.000	0.002	0.938	0.188
	Naïve		0.529	0.029	0.004	0.894	0.214
	MAR		0.529	0.029	0.004	0.914	0.212
	MNAR		0.509	0.009	0.003	0.932	0.209
β_{221}	Full	-0.500	-0.504	-0.004	0.003	0.942	0.194
	Naïve		-0.461	0.039	0.014	0.890	0.380
	MAR		-0.475	0.025	0.011	0.916	0.364
	MNAR		-0.491	0.009	0.008	0.934	0.311
β_{212}	Full	0.200	0.201	0.001	0.000	0.946	0.048
	Naïve		0.187	-0.013	0.001	0.904	0.095
	MAR		0.189	-0.011	0.001	0.902	0.094
	MNAR		0.202	0.002	0.001	0.934	0.093
β_{222}	Full	1.000	1.001	0.001	0.001	0.944	0.097
	Naïve		0.938	-0.062	0.012	0.814	0.287
	MAR		0.966	-0.034	0.007	0.896	0.275
	MNAR		0.992	-0.008	0.004	0.954	0.239
π_1	Full	0.558	0.557				
	Naïve	0.576	0.632				
	MAR	0.576	0.623				
	MNAR	0.576	0.576				

Table F.2: Simulation results of S2 for subject misclassification under the **Full**, **Naïve**, **MAR**, and **MNAR** methods

		Percentile			
Method	Min	25	50	75	Max
Full	0.00	0.01	0.01	0.01	0.03
Naïve	0.09	0.12	0.13	0.14	0.19
MAR	0.08	0.12	0.13	0.14	0.18
MNAR	0.01	0.02	0.03	0.03	0.06

Table F.3: Simulation results of S3 for parameter estimation of intercept β_{rk1} and slope β_{rk2} for longitudinal outcome r in latent class k, and latent class-level weights π_k under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

Parameter	Method	Truth	Mean	Bias	MSE	Coverage	Length
β_{111}	Full	-0.250	-0.247	0.003	0.002	0.942	0.190
	Naïve		-0.196	0.054	0.007	0.832	0.231
	MAR		-0.214	0.036	0.005	0.890	0.225
	MNAR		-0.249	0.001	0.003	0.956	0.209
β_{121}	Full	-1.000	-1.003	-0.003	0.003	0.950	0.231
	Naïve		-0.943	0.057	0.016	0.885	0.397
	MAR		-0.975	0.025	0.010	0.928	0.366
	MNAR		-0.992	0.008	0.006	0.962	0.314
β_{112}	Full	0.100	0.100	-0.000	0.000	0.948	0.048
	Naïve		0.086	-0.014	0.001	0.913	0.101
	MAR		0.088	-0.012	0.001	0.934	0.096
	MNAR		0.098	-0.002	0.001	0.960	0.091
β_{122}	Full	0.500	0.500	0.000	0.001	0.938	0.097
	Naïve		0.393	-0.107	0.017	0.593	0.252
	MAR		0.438	-0.062	0.008	0.790	0.227
	MNAR		0.497	-0.003	0.003	0.958	0.216
β_{211}	Full	0.500	0.503	0.003	0.003	0.926	0.191
	Naïve		0.538	0.038	0.006	0.866	0.236
	MAR		0.534	0.034	0.005	0.866	0.231
	MNAR		0.509	0.009	0.003	0.936	0.210
β_{221}	Full	-0.500	-0.503	-0.003	0.002	0.950	0.197
	Naïve		-0.477	0.023	0.010	0.929	0.363
	MAR		-0.492	0.008	0.009	0.932	0.349
	MNAR		-0.486	0.014	0.007	0.932	0.310
β_{212}	Full	0.200	0.200	-0.000	0.000	0.940	0.048
	Naïve		0.192	-0.008	0.001	0.917	0.099
	MAR		0.192	-0.008	0.001	0.942	0.098
	MNAR		0.201	0.001	0.001	0.932	0.093
β_{222}	Full	0.300	0.299	-0.001	0.001	0.942	0.096
	Naïve		0.198	-0.102	0.014	0.597	0.236
	MAR		0.223	-0.077	0.010	0.726	0.233
	MNAR		0.300	-0.000	0.003	0.944	0.235
π_1	Full	0.557	0.556				
	Naïve	0.577	0.598				
	MAR	0.576	0.592				
	MNAR	0.577	0.574				

Table F.4: Simulation results of S3 for subject misclassification under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

		Percentile			
Method	Min	25	50	75	Max
Full	0.00	0.02	0.02	0.03	0.05
Naïve	0.08	0.14	0.15	0.17	0.22
MAR	0.10	0.13	0.14	0.16	0.20
MNAR	0.01	0.03	0.03	0.04	0.06

Table F.5: Simulation results of S4 for parameter estimation of intercept β_{rk1} and slope β_{rk2} for longitudinal outcome r in latent class k, and latent class-level weights π_k under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

Parameter	Method	Truth	Mean	Bias	MSE	Coverage	Length
β_{111}	Full	-0.250	-0.252	-0.002	0.002	0.950	0.190
	Naïve		-0.242	0.008	0.003	0.954	0.199
	MAR		-0.251	-0.001	0.003	0.946	0.197
	MNAR		-0.251	-0.001	0.002	0.958	0.195
β_{121}	Full	-1.000	-1.000	0.000	0.003	0.956	0.230
	Naïve		-0.986	0.014	0.009	0.922	0.338
	MAR		-0.999	0.001	0.007	0.944	0.307
	MNAR		-0.992	0.008	0.005	0.940	0.274
β_{112}	Full	0.100	0.100	-0.000	0.000	0.930	0.048
	Naïve		0.096	-0.004	0.000	0.944	0.066
	MAR		0.096	-0.004	0.000	0.930	0.064
	MNAR		0.101	0.001	0.000	0.938	0.062
β_{122}	Full	0.500	0.501	0.001	0.001	0.930	0.096
	Naïve		0.460	-0.040	0.005	0.878	0.222
	MAR		0.483	-0.017	0.003	0.924	0.194
	MNAR		0.496	-0.004	0.002	0.968	0.176
β_{211}	Full	0.500	0.500	-0.000	0.002	0.954	0.189
	Naïve		0.508	0.008	0.003	0.946	0.199
	MAR		0.506	0.006	0.003	0.936	0.198
	MNAR		0.503	0.003	0.003	0.952	0.195
β_{221}	Full	-0.500	-0.503	-0.003	0.003	0.940	0.196
	Naïve		-0.489	0.011	0.007	0.938	0.311
	MAR		-0.490	0.010	0.007	0.918	0.296
	MNAR		-0.490	0.010	0.005	0.924	0.263
β_{212}	Full	0.200	0.199	-0.001	0.000	0.918	0.048
	Naïve		0.193	-0.007	0.000	0.916	0.066
	MAR		0.192	-0.008	0.000	0.918	0.065
	MNAR		0.200	0.000	0.000	0.944	0.064
β_{222}	Full	0.750	0.751	0.001	0.001	0.934	0.097
	Naïve		0.706	-0.044	0.006	0.872	0.223
	MAR		0.719	-0.031	0.004	0.885	0.212
	MNAR		0.743	-0.007	0.003	0.928	0.190
π_1	Full	0.557	0.557				
	Naïve	0.557	0.601				
	MAR	0.558	0.590				
	MNAR	0.557	0.553				

Table F.6: Simulation results of S4 for subject misclassification under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

		Percentile			
Method	Min	25	50	75	Max
Full	0.00	0.01	0.02	0.02	0.04
Naïve	0.07	0.10	0.11	0.12	0.16
MAR	0.06	0.09	0.10	0.11	0.14
MNAR	0.00	0.02	0.02	0.02	0.04

Table F.7: Simulation results of S5 for parameter estimation of intercept β_{rk1} and slope β_{rk2} for longitudinal outcome r in latent class k, and latent class-level weights π_k under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

Parameter	Method	Truth	Mean	Bias	MSE	Coverage	Length
β_{111}	Full	-0.250	-0.252	-0.002	0.002	0.950	0.190
	Naïve		-0.204	0.046	0.007	0.870	0.249
	MAR		-0.216	0.034	0.004	0.900	0.222
	MNAR		-0.245	0.005	0.003	0.964	0.210
β_{121}	Full	-1.000	-1.000	0.000	0.003	0.956	0.230
	Naïve		-0.909	0.091	0.024	0.830	0.424
	MAR		-0.985	0.015	0.011	0.928	0.374
	MNAR		-0.990	0.010	0.006	0.942	0.316
β_{112}	Full	0.100	0.100	-0.000	0.000	0.930	0.048
	Naïve		0.088	-0.012	0.001	0.920	0.117
	MAR		0.091	-0.009	0.001	0.930	0.095
	MNAR		0.098	-0.002	0.001	0.950	0.091
β_{122}	Full	0.500	0.501	0.001	0.001	0.930	0.096
	Naïve		0.378	-0.122	0.022	0.590	0.282
	MAR		0.448	-0.052	0.007	0.846	0.235
	MNAR		0.497	-0.003	0.003	0.952	0.216
β_{211}	Full	0.500	0.500	-0.000	0.002	0.954	0.189
	Naïve		0.576	0.076	0.011	0.754	0.250
	MAR		0.545	0.045	0.006	0.888	0.233
	MNAR		0.509	0.009	0.004	0.922	0.220
β_{221}	Full	-0.500	-0.503	-0.003	0.003	0.940	0.196
	Naïve		-0.404	0.096	0.025	0.778	0.405
	MAR		-0.451	0.049	0.015	0.874	0.388
	MNAR		-0.497	0.003	0.007	0.950	0.334
β_{212}	Full	0.200	0.199	-0.001	0.000	0.918	0.048
	Naïve		0.180	-0.020	0.001	0.882	0.116
	MAR		0.184	-0.016	0.001	0.898	0.109
	MNAR		0.197	-0.003	0.001	0.930	0.105
β_{222}	Full	0.750	0.751	0.001	0.001	0.934	0.097
	Naïve		0.602	-0.148	0.030	0.478	0.285
	MAR		0.666	-0.084	0.014	0.740	0.280
	MNAR		0.746	-0.004	0.004	0.976	0.258
π_1	Full	0.557	0.557				
	Naïve	0.575	0.598				
	MAR	0.577	0.604				
	MNAR	0.576	0.573				

Table F.8: Simulation results of S5 for subject misclassification under the **Full**, **Naïve**, **MAR**, and **MNAR** methods.

		Percentile			
Method	Min	25	50	75	Max
Full	0.00	0.01	0.02	0.02	0.04
Naïve	0.11	0.15	0.17	0.18	0.25
MAR	0.11	0.14	0.15	0.17	0.21
MNAR	0.01	0.03	0.04	0.04	0.07

Bibliography

[Albert and Chib, 1993] J. Albert and S. Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[American Academy of Pediatrics, 2018] American Academy of Pediatrics. AAP Schedule of Well-Child Care Visits, 2018.

[Asparouhov, 2005] T. Asparouhov. Sampling Weights in Latent Variable Modeling. *Structural Equation Modeling*, 12(3):411–434, 2005.

[Besag and Kooperberg, 1995] J. Besag and C. Kooperberg. On Conditional and Intrinsic Autoregression. *Biometrika*, 82(4):733–746, 1995.

[Besag, 1974] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

[Biernacki *et al.*, 2000] C. Biernacki, G. Celeux, and G. Govaert. Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725, 2000.

[Bonanno and Diminich, 2013] G. A. Bonanno and E. D. Diminich. Annual Research Review: Positive Adjustment to Adversity - Trajectories of Minimal-Impact Resilience and Emergent Resilience. *Journal of Child Psychology and Psychiatry*, 54(4):378–401, 2013.

[Bunch, 1991] D. S. Bunch. Estimability in the Multinomial Probit Model. *Transportation Research Part B: Methodological*, 25B(1):1–12, 1991.

[Celeux *et al.*, 2006] G. Celeux, F. Forbes, C. P. Robert, and D. M. Titterington. Deviance Information Criteria for Missing Data Models. *Bayesian Analysis*, 1(4):651–674, 2006.

[Centers for Disease Control and Prevention, 2019] Centers for Disease Control and Prevention. A SAS Program for the 2000 CDC Growth Charts (ages 0 to <20 years), 2019.

[Chen *et al.*, 2010] Q. Chen, M. R. Elliott, and R. J. A. Little. Bayesian Penalized Spline Model-based Inference for Finite Population Proportion in Unequal Probability Sampling. *Survey Methodology*, 36(1):23–34, 2010.

[Daganzo, 1979] C. Daganzo. *Multinomial Probit: The Theory and its Application to Demand Forecasting*. Academic Press, New York, 1979.

[Daniels and Hogan, 2008] M. J. Daniels and J. W. Hogan. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall/CRC, Boca Raton, 2008.

[Dansie, 1985] B. R. Dansie. Parameter Estimability in the Multinomial Probit Model. *Transportation Research Part B: Methodological*, 19B(6):526–528, 1985.

[Elliott *et al.*, 2005] M. R. Elliott, J. J. Gallo, T. R. Ten Have, H. R. Bogner, and I. R. Katz. Using a Bayesian Latent Growth Curve Model to Identify Trajectories of Positive

Affect and Negative Events Following Myocardial Infarction. *Biostatistics*, 6(1):119–143, 2005.

[Follmann and Wu, 1995] D. Follmann and M. Wu. An Approximate Generalized Linear Model with Random Effects for Informative Missing Data. *Biometrics*, 51(1):151–168, 1995.

[Frühwirth-Schnatter *et al.*, 2004] S. Frühwirth-Schnatter, R. Tüchler, and T. Otter. Bayesian Analysis of the Heterogeneity. *Journal of Business and Economic Statistics*, 22(1):2–15, 2004.

[Fruhwirth-Schnatter, 2006] S. Fruhwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Science & Business Media, New York, 2006.

[Galea, 2017] S. Galea. Health Haves, Health Have Nots, and Heterogeneity in Population Health. *The Lancet Public Health*, 2:e388–e389, 2017.

[Garrett and Zeger, 2000] E. S. Garrett and S. L. Zeger. Latent Class Model Diagnosis. *Biometrics*, 56:1055–1067, 2000.

[Gelfand *et al.*, 1995] A. E. Gelfand, S. K. Sahu, and B. P. Carlin. Efficient Parametrisations for Normal Linear Mixed Models. *Biometrika*, 82(3):479–488, 1995.

[Gelfand *et al.*, 1996] A. E. Gelfand, S. K. Sahu, and B. P. Carlin. Efficient Parametrizations for Generalized Linear Mixed Models, (with discussion). In J. M. Bernardo, J. O. Berger, and A. P. Dawid, editors, *Bayesian Statistics 5*, pages 165–180. Clarendon Press, 1996.

[Gelman *et al.*, 1996] A. Gelman, X. Meng, and H. Stern. Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

[Gelman *et al.*, 2005] A. Gelman, I. V. Mechelen, G. Verbeke, D. F. Heitjan, and M. Meulders. Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data. *Biometrics*, 61:74–85, 2005.

[Gelman *et al.*, 2014] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. Taylor & Francis, 2014.

[Gelman, 2006] A. Gelman. Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis*, 1(3):515–534, 2006.

[Gruebner *et al.*, 2016] O. Gruebner, S. R. Lowe, M. Tracy, M. Cerdá, S. Joshi, F. H. Norris, and S. Galea. The Geography of Mental Health and General Wellness in Galveston Bay after Hurricane Ike: A Spatial Epidemiologic Study with Longitudinal Data. *Disaster Medicine and Public Health Preparedness*, 10(2):261–273, 2016.

[Hripcsak and Albers, 2013] G. Hripcsak and D. J. Albers. Next-Generation Phenotyping of Electronic Health Records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[Imai and van Dyk, 2005] K. Imai and D. A. van Dyk. MNP: R Package for Fitting the Multinomial Probit Model. *Journal of Statistical Software*, 14(3):1–32, 2005.

[Jones and Nagin, 2007] B. L. Jones and D. L. Nagin. Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them. *Sociological Methods and Research*, 35(4):542–571, 2007.

[Jones *et al.*, 2001] B. L. Jones, D. S. Nagin, and K. Roeder. A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. *Sociological Methods and Research*, 29(3):374–393, 2001.

[Jones, 2011] R. H. Jones. Bayesian Information Criterion for Longitudinal and Clustered Data. *Statistics in Medicine*, 30:3050–3056, 2011.

[Jung and Wickrama, 2008] T. Jung and K A S Wickrama. An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling. *Social and Personality Psychology Compass*, 2(1):302–317, 2008.

[Keribin, 2000] C. Keribin. Consistent Estimation of the Order of Mixture Models. *Sankhya, Series A, The Indian Journal of Statistics*, 62:49–66, 2000.

[Koop, 2003] G. Koop. Qualitative and Limited Dependent Variables. In *Bayesian Econometrics*, chapter 9, pages 209–234. John Wiley & Sons, Hoboken, NJ, 2003.

[Liang *et al.*, 2009] Y. Liang, W. Lu, and Z. Ying. Joint Modeling and Analysis of Longitudinal Data with Informative Observation Times. *Biometrics*, 65:377–384, 2009.

[Lin *et al.*, 2004] H. Lin, C. E. McCulloch, and R. A. Rosenheck. Latent Pattern Mixture Models for Informative Intermittent Missing Data in Longitudinal Studies. *Biometrics*, 60(2):295–305, 2004.

[Little, 2003] R. J. A. Little. The Bayesian Approach to Sample Survey Inference. In R. L. Chambers and C. J. Skinner, editors, *Analysis of Survey Data*, pages 49–52. John Wiley and Sons Ltd, West Sussex, 2003.

[Little, 2004] R. J. Little. To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, 99(466):546–556, 2004.

[Lowe *et al.*, 2015] S. R. Lowe, S. Joshi, R. H. Pietrzak, S. Galea, and M. Cerdá. Mental Health and General Wellness in the Aftermath of Hurricane Ike. *Social Science & Medicine*, 124:162–170, 2015.

[McCulloch and Rossi, 1994] R. McCulloch and P. E. Rossi. An Exact Likelihood Analysis of the Multinomial Probit Model. *Journal of Econometrics*, 64:207–240, 1994.

[McCulloch *et al.*, 2016] C. E. McCulloch, J. M. Neuhaus, and R. L. Olin. Biased and Unbiased Estimation in Longitudinal Studies with Informative Visit Processes. *Biometrics*, 72(4):1315–1324, 2016.

[Muthen and Muthen, 2017] L. K. Muthen and B. O. Muthen. *Mplus User's Guide*. Muthen & Muthen, Los Angeles, 8 edition, 2017.

[Muthen *et al.*, 2002] B. Muthen, C. H. Brown, K. Masyn, B. Jo, S. Khoo, C. Yang, C. Wang, S. G. Kellam, J. B. Carlin, and J. Liao. General Growth Mixture Modeling for Randomized Preventive Interventions. *Biostatistics*, 3(4):459–475, 2002.

[National Institute of Mental Health, 2016] National Institute of Mental Health. Post-Traumatic Stress Disorder. *Mental Health Information, Health Topics*, 2016.

[Neelon *et al.*, 2011] B. Neelon, G. K. Swamy, L. F. Burgette, and M. L. Miranda. A Bayesian Growth Mixture Model to Examine Maternal Hypertension and Birth Outcomes. *Statistics in Medicine*, 30(22):2721–2735, 2011.

[Norris *et al.*, 2002] F. H. Norris, M. J. Friedman, P. J. Watson, C. M. Byrne, and K. Kaniasty. 60,000 Disaster Victims Speak: Part I. An Empirical Review of the Empirical Literature, 1981-2001. *Psychiatry*, 65(3):207–239, 2002.

[Norris *et al.*, 2009] F. H. Norris, M. Tracy, and S. Galea. Looking for Resilience: Understanding the Longitudinal Trajectories of Responses to Stress. *Social Science & Medicine*, 68:2190–2198, 2009.

[Patterson *et al.*, 2002] B. H. Patterson, C. M. Dayton, and B. I. Graubard. Latent Class Analysis of Complex Sample Survey Data: Application to Dietary Data. *Journal of the American Statistical Association*, 97(459):721–741, 2002.

[Rice, 2016] H. Rice. Hurricane Ike Worst Storm in Decades. *Houston Chronical*, oct 2016.

[Roy, 2003] J. Roy. Modeling Longitudinal Data with Nonignorable Dropouts. *Biometrics*, 59:829–836, 2003.

[Rubin, 1976] D. B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976.

[Sanderson *et al.*, 1988] M. Sanderson, C. Scott, and J. F. Gonzalez. 1988 National Maternal and Infant Health Survey: Methods and Response Characteristics. *Vital and Health Statistics*, 2(125):1–48, 1988.

[Schwarz, 1978] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[Spiegelhalter *et al.*, 2002] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Linde. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, B Methodology*, 64(4):583–639, 2002.

[Sun *et al.*, 2007] Jianguo Sun, Liuquan Sun, and Dandan Liu. Regression Analysis of Longitudinal Data in the Presence of Informative Observation and Censoring Times. *Journal of the American Statistical Association*, 102(480):1397–1406, 2007.

[Tanner and Wong, 1987] M. A. Tanner and W. H. Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[Valliant *et al.*, 2009] R. Valliant, T. Adams, and J. Wagner. Sample Design Documentation Galveston Bay Recovery Survey 2008-2009. Technical report, Survey Research Operations, Production Sampling Group, University of Michigan Survey Research Center, Ann Arbor, 2009.

[Verbeke and Lesaffre, 1996] G. Verbeke and E. Lesaffre. A Linear Mixed-Effects Model with Heterogeneity in the Random-Effects Population. *Journal of the American Statistical Association*, 91(433):217–221, 1996.

[Wedel *et al.*, 1998] M. Wedel, F. Hofstede, and J. E. M. Steenkamp. Mixture Model Analysis of Complex Samples. *Journal of Classification*, 15(2):225–44, 1998.

[Weiskopf and Weng, 2013] N. G. Weiskopf and C. Weng. Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research. *Journal of the American Medical Informatics Association*, 20:144–151, 2013.

[Wu and Carroll, 1988] M. C. Wu and R. J. Carroll. Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics*, 44(1):175–188, 1988.