Bayesian Joint Modeling of Longitudinal Trajectories and Health Outcome: A Broad Evaluation of Mean and Variation Features in Risk Profiles and Model Assessments

by

Bei Jiang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2014

Doctoral Committee:

Professor Michael R. Elliott, Co-Chair
Professor Naisyin Wang, Co-Chair
Professor Siobán D. Harlow
Professor Timothy D. Johnson
Assistant Professor Lu Wang

ACKNOWLEDGEMENTS

I am deeply indebted to my co-chairs, Drs. Naisyin Wang and Michael Elliott for their patience, guidance and support in countless ways. They inspire me to think critically, guide my academic development and provide me valuable feedback throughout the entire process of my dissertation research. Most importantly, their passionate commitment to research and overwhelming enthusiasm in statistics have greatly impacted me. It has been a great honor and privilege to work with them.

I would like to thank my dissertation committee members, Dr. Siobán Harlow, Dr. Timothy Johnson and Dr. Lu Wang for their insightful comments that have greatly improved this dissertation. Particularly, I am very grateful to Dr. Harlow for sharing her expertise in women reproductive aging, Dr. Johnson for teaching me Bayesian statistics and offering me the opportunity to be exposed to brain imaging research, and Dr. Wang for helpful suggestions on my professional development and friendship.

I also wish to express my thanks to Dr. Mary Sammel at the University of Pennsylvania for providing us the interesting data used in this dissertation, always promptly and patiently answering our questions about the data, discussing the directions of the modeling approaches and pointing out crucial features in the data.

Finally, thanks to my family. I am forever indebted to them for their endless love and consistent support throughout my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Bayesian Joint Modeling of Longitudinal Trajectories and Health Outcome: A Broad Evaluation of Mean and Variation Features in Risk Profiles and Model Assessments

by

Bei Jiang

Co-Chair: Professors Michael R. Elliott and Naisyin Wang

This dissertation consists of methodology developments and applications for joint modeling of repeated measurements of health risk factors (i.e., longitudinal trajectories) and health outcome data. Joint modeling is a natural choice to link longitudinally observed covariates and disease outcomes: a mixed-effect model postulates the longitudinal trajectories and features extracted therein serve as predictors in a primary disease-outcome model. In the first chapter, we consider joint models that incorporate information from both long-term mean trends and short-term variability in the longitudinal submodel. We then utilize both multiple shared random-effects (MSRE) and latent class (LC) approaches to predict a binary disease outcome in the primary model. We develop simulation studies to compare and contrast these two modeling strategies; in particular, we study in detail the effects of the primary model misspecification. In the second chapter, we develop a joint modeling method that uses the individual-level longitudinal measurements of follicle stimulating hormone (FSH) to predict the occurrence of severe hot flashes in a manner that distinguishes long-term trends of the mean trajectory, cumulative change captured by the deriva-

tive of mean trajectory, and short-term residual variability. Our method allows the potential effects of longitudinal trajectories on the health risks to vary and accumulate over time. We further utilize the proposed methods to narrow down the critical time windows of increased health risks. The third chapter is a detailed study of model assessment. We evaluate six Bayesian model assessment criteria in the context of a model that simultaneously considers a set of longitudinal predictors and a primary outcome, connected through either LC or MSRE predictors. We focus on two evaluation aspects: goodness-of-fit adjusted for the complexity of the models, and prediction evaluation based on both training and test samples as well as their contrasts. An interesting result is that when the data are generated under an MSRE mechanism but fit assuming an LC mechanism, a very highly predictive "artifact" can be generated under certain scenarios. The consequence of this phenomenon is that an over-optimistic classification estimate can be built on such an artifact.

The methods developed in all three papers are applied to data from the Penn Ovarian Aging Study, a 13-year longitudinal study comprised of a population-based sample of 436 women aged 35-47 years selected via random-digit dialing in Philadelphia, PA during 1996-97 and followed through 2010. The study goal is to explore the associations between reproductive hormone levels (follicle stimulating hormone or FSH) and symptoms in the transition to menopause (severe hot flashes).

CHAPTER I

Introduction

Epidemiology studies or clinical trials often measure both the risk factors and disease outcome repeated over time in order to seek useful features in these longitudinal risk profiles to predict disease progression. In previous developments of statistical tools to analyze such data, the focus has often been on a single longitudinal risk process with the goal of using the true underlying longitudinal process (i.e., mean profile) to relate to the disease process. It is therefore natural to model both the observed longitudinal and disease processes jointly to gain more efficient inference, where a longitudinal submodel is outlined for the observed longitudinal risk profiles $\mathbf{Y}$ and a primary disease outcome model is defined to link the true underlying longitudinal process and disease process $\mathbf{O}$ together. The common approach is to assume a latent process $\mathbf{u}$ underlying both $\mathbf{Y}$ and $\mathbf{O}$ such that $\mathbf{u}$ account for all correlations between $\mathbf{Y}$ and $\mathbf{O}$, i.e., $[\mathbf{Y},\mathbf{O}] = \int [\mathbf{Y},\mathbf{O} \mid \mathbf{u}] d\mathbf{u} = \int [\mathbf{Y} \mid \mathbf{u}][\mathbf{O} \mid \mathbf{u}] d\mathbf{u}$. Within this general joint modeling framework, the longitudinal submodel is usually a mixed effect model and the latent variable $\mathbf{u}$ takes one of the two forms: the random effects, defined to capture the individual variations from the population mean, and the latent class, defined to capture the heterogeneity in the random effects in the sense of Verbeke and Lesaffre (1996) such that subjects in a particular latent class have similar shapes in the longitudinal risk profiles and share the same disease risk, possibly conditional on

other covariates. Each approach has its own focuses and strengths as we will explore in later chapters.

Many joint models based on these two approaches have been developed in the context of cancer research and HIV/AIDS studies, mainly focusing on novel developments of longitudinal and outcome submodels. Such innovations include the use of semi- and non-parametric longitudinal models to accommodate more flexible risk profiles and developing outcome submodels to accommodate more complex disease outcomes, such as zero-inflated outcomes, competing events, multiple failure times and cure fractions. For example, Rizopoulos and Ghosh (2011) linked multiple longitudinal processes of different types to a time-to-event outcome. Thorough reviews of the topic are given by Ibrahim et al. (2001), Tsiatis and Davidian (1996), Ibrahim et al. (2010) and Rizopoulos (2012); and a recent special issue of *Lifetime Data Analysis* edited by Chen and Gustafson (2011) was dedicated to various aspects of joint modeling of longitudinal and time-to-event data. However, these previous joint modeling strategies have focused on using the true underlying longitudinal process (i.e., mean profile) to predict a disease outcome while treating the residual variability (i.e., variance) in the longitudinal profile as a nuisance parameter such that it does not influence the disease progression. Recently, Elliott et al. (2012) proposed a new method focusing on multiple shared random effects (MSRE) with the individual level variance in the longitudinal profiles as another latent variable to link both sources of data. This new method laid out a new framework of joint modeling of longitudinal data and disease outcome data by linking the features in not only the long term mean trend but also short term variability in the longitudinal trajectories to predict a disease outcome.

This dissertation contributes to this fast growing area by developing novel joint models to combine information from both mean trajectory and residual variability in longitudinal predictors of key health outcomes. Motivating our work is the Penn

Ovarian Aging Study (Freeman et al. 2001; Manson et al. 2001), a longitudinal study consisting of a population-based sample of 436 women aged 35-47 years selected via random digit dialing in Philadelphia County, PA during 1996-97 and followed through 2010. One of the major goals of this study is to explore associations between reproductive hormone levels and symptoms in the transition to menopause. Follicle stimulating hormone (FSH) is of particular interest because it is known to stimulate folliculogenesis, an important factor in ovarian aging. Our aim is to develop flexible modeling tools that allow us to uncover new dynamic features in the various aspects of the FSH evolution history that predict the menopausal symptoms of interest (in particular, cross-sectional onset of severe hot flashes and severity of hot flashes).

We base our developments within the framework of joint models of longitudinal and disease outcome data. We outline one longitudinal submodel for the longitudinal history of FSH levels, and one primary outcome model for symptoms of interest along with assumptions of how the FSH evolution process relates to subsequent menopausal symptoms. In particular, the formulation of a longitudinal submodel for the FSH data takes into account the following data features: 1) the FSH data are highly unbalanced with available measurements across women taken at different ages and ranging from 1 to 26 out of the maximum 26 possible measurements; 2) overall, the FSH evolution pattern is relatively flat between age 35 and 40 and then has an increasing period later on where there exist both acute and gradual increase periods; 3) the fluctuations of the FSH levels are highly variably among these women, reflecting low or high short term variabilities around the mean trends. Motivated by specific scientific questions, we consider association structures that go beyond the common SRE and LC approaches to relate the important features in the FSH evolution process to menopausal symptoms.

In Chapter II, we start our analysis with the popular shared random effect and

latent class approach but built within the new framework proposed by Elliott et al. (2012) to investigate the interesting features of FSH trajectories that are manifest in both the mean trends and the short term fluctuations. However, an important limitation of the shared random effects approach is the difficulty of interpreting the effects when the random effects are associated with complex functions of time (for example, spline representations). To avoid this difficulty, we seek features in FSH trajectories from the perspective of linear deviations from a population trend estimated by the LOWESS method; the individual level mean trends are therefore assumed to deviate from the LOWESS estimated population trend in the direction of intercept and slope while also accounting for different rates of change in FSH levels among women. Finite mixture lognormal distributions are used to model the individual level residual variance because evidence suggests that the highly variable short term fluctuations in the FSH evolution process cluster into low or high categories. Although shared random effect and latent class approaches have their own focuses, a lack of existing knowledge about the strengths and weaknesses of each approach motivates us further to compare and contrast these two approaches with the goal of providing useful insights into interpreting our data analysis results. This investigation reveals interesting phenomena that are new to the literature.

While there are advantages of using summary measures of the longitudinal trajectories to relate to the disease outcome, as in Chapter II, it is also of clinical interest to investigate the contribution of the history of the FSH levels over the entire late reproductive period to the menopausal symptoms by quantifying the cumulative time varying effects and identifying the critical age range when the elevated FSH levels lead to increased risks of more severe symptoms. To meet this aim, in Chapter III, we consider alternative association structures to link the FSH evolution process and menopausal symptoms. In particular, we borrow existing tools from functional data analysis literature while building our models within the framework proposed

by Elliott et al. (2012), with the goal to distinguish between long-term mean trends and short-term variability. This proposed association structure relates the primary outcome of interest to an integral of long-term mean trend $\mu(t)$ over certain time windows, i.e., $\int_T \mu(t)\theta(t)dt$, where $\theta(t)$ is the functional coefficient function capturing the cumulative time varying effect due to the longitudinal history. We consider the Bayesian penalized B spline approach by Lang and Brezger (2004) to achieve smooth but flexible representation of $\theta(t)$. Further, we consider a robust semi-parametric model to smooth the underlying FSH trajectories by 1) associating random effects with B spline basis functions in a mixed effect model and using penalty term to shrink unnecessary fluctuation towards zero; and 2) assuming within-subject residual errors follow a t distribution rather than a typical normal distribution to avoid the potential influence of outlying observations, where a mixture of log normal distributions is assumed to allow for potential over-dispersion of the within-subject scale parameters. In contrast to the two-step procedure used in Chapter II to simplify the random effects structure, this formulation of the longitudinal submodel is more flexible and makes the evaluation of an alternative association structure $\int_T \mu'(t)\theta_1(t)dt$ possible, where $\mu'(t) = \partial\mu(t)/\partial t$ and conceptually captures the cumulative changes in the long-term mean trends. The implementation of the new proposed joint modeling approach in Bayesian paradigm is straightforward. It allows for simultaneous evaluation of the uncertainty in estimating $\theta(t)$ or $\theta_1(t)$ by providing point-wise credible intervals and eventually leads to easy identification of the critical age range of increased risks.

In Chapter IV, we focus on the model assessment for the joint models considered in Chapter II. In practice, such joint models can be used as prognostic tools to predict prospective event risks for a particular patient, given the patient's history of longitudinal measurements. However, investigations studying the pros and cons of joint LC and MSRE models in Chapter II reveal that different numbers of mixture compo-

nents as well as the assumed association structure to link the longitudinal submodel and primary outcome model (either in the form of MSRE or LC for these random effects) can greatly affect the target individualized predictions. In these settings, it is relevant to consider how to choose among these candidate models and how to assess which model best predicts the risk of the health outcome of interest. Motivated by these considerations, in this chapter, we evaluate several Bayesian model assessment criteria, including a recently proposed one, WAIC by Watanabe (2010). Some criteria are modified following the Bayesian principle when necessary to accommodate the joint modeling framework that analyzes longitudinal predictors and binary health outcome data. We base our evaluation on empirical numerical studies and focus on two evaluation aspects: goodness-of-fit adjusted for the complexity of the models, reflected by the numbers of latent features in the longitudinal trajectories that are part of the hierarchical structures in the joint models, and prediction evaluation based on both training and test samples as well as their contrasts.

Finally, we conclude this dissertation in Chapter V, summarizing our major findings, exploring in some depth issues that were set aside in the main chapters, and discussing limitations and suggestions for future research.

CHAPTER II

Joint Modeling of Cross-Sectional Health Outcomes and Longitudinal Predictors via Mixtures of Means and Variances

2.1 Introduction

Joint models are a natural choice to link longitudinally observed covariates and disease outcomes. Many joint models have been developed in the context of cancer research and HIV/AIDS clinical trials, where a mixed-effect model is outlined for the longitudinal trajectories and a primary outcome model is defined for the disease outcome. The two models are usually linked together in one of two ways: (1) shared random effects (SRE) models, where a functional form of the random effects in the mixed effect model is a covariate in the outcome model, and (2) latent class (LC) models, where there exists heterogeneity (precisely, latent classes) in the mean profiles of the longitudinal trajectories, and the subjects in a particular latent class share the same risk of event, possibly conditional on other covariates. Each approach has its strengths.

For SRE models, the random effects are assumed to capture the main features in the longitudinal trajectories that predict the outcomes. The concept of "shared parameters" was used in Wu and Carroll (1988) to model non-ignorable missing data.

Articles by Tsiatis and Davidian (2004) and Ibrahim et al. (2010) and the book by Ibrahim et al (2001) offer excellent general reviews of these models.

In the LC model literature, growth mixture models (Verbeke and Lesaffre (1996), Muthén and Shedden (1999), Muthén et al. (2002)) are often used to classify longitudinal trajectories. Proust-Lima et al (2012) studied the joint LC modeling in detail and contrasted its use with that of joint SRE models. To the best of our knowledge, their work was the first to investigate and compare the goodness of fit and prediction accuracy of joint LC versus joint SRE modeling. They illustrate and compare model performance using data from a prostate cancer study.

In this paper, we use both approaches to investigate the association between longitudinal hormone levels and menopausal symptoms for a group of middle-aged women. The Penn Ovarian Aging Study (Freeman et al. 2001; Manson et al. 2001) is a longitudinal study consisting of a population-based sample of 436 women aged 35-47 years selected via random digit dialing in Philadelphia County, PA during 1996-97. At each annual assessment period, study personnel visited each subject two times approximately one month apart for an in-person interview and a blood sample for hormone measurements. One of the major goals of this study is to explore associations between reproductive hormone levels and symptoms in the transition to menopause. Changes in hormone levels alter menstrual bleeding patterns prior to the cessation of menstruation (menopause) marking the end of a woman's reproductive years. This course of events, which can last 5 or more years, coincides for a majority of women with the development of hot flashes, sleep disorders, and bone loss, among other symptoms. While researchers have focused on the associations between these symptoms and hormone levels, the impact of within woman rate of change and variability in hormones, such as Follicle Stimulating Hormone (FSH), is not well understood.

To address such questions, we investigate methods that can model different as-

pects of the dynamic process of hormone change and variability and simultaneously link them with the symptoms of interest. A novel feature of our modeling effort is to link both longitudinal profiles and residual variability of the FSH measures to risk of experiencing severe hot flashes. While most joint models have treated within-subject variability as a nuisance parameter, recently a small literature has developed to evaluate the associations between within-subject variability in longitudinal data and the primary outcomes (Sammel 2001; Elliott 2007, 2012). Because of biological considerations that subject-level hormone fluctuation may accentuate menopausal symptoms (Gracia et al. 2004; Freeman et al. 2006), we develop a joint modeling approach that simultaneously examines the association between subject-level mean trajectories and variability of FSH, and the binary indicator of severe hot flashes.

Because of evidence that these trajectories or variabilities are heterogenous and cluster into possibly clinically relevant groupings, we develop a first-stage growth mixture model for FSH hormone that also includes latent classes for the subject-level variability. We consider two second-stage outcome models: one based on a "multiple shared random effects" (MSRE) model that uses the individual level trajectories and variabilities to predict severe hot flash experience, and the other based on an LC model that uses the latent classes themselves to predict the experience of severe hot flashes. Our key focus is not on one second-stage model or the other, but their contrast. In addition to model fit and model checking efforts, we develop simulation studies to examine the robustness and predictive accuracy of each approach when the model may or may not be correctly specified. While Proust-lima et al.'s work (2012) mainly focuses on comparing different modeling strategies on a real data example, our simulation investigation sheds light on the potential impact of model misspecification in terms of model estimation and predictive power.

2.2 Joint models and corresponding approaches

The joint modelling approach consists of a model for the longitudinal trajectories and a primary model for the outcomes.

- The longitudinal submodel is a generalized growth mixture model (Muthén and Shedden 1999) with subject-specific mean trajectories and residual variances:

$$\begin{aligned} y_{ij}|\mathbf{b}_i, \sigma_i^2 &\sim \mathrm{N}\{f(\mathbf{b}_i; t_{ij}), \sigma_i^2\} \\ D_i &\sim \text{Multinomial}(\pi_1^D, ..., \pi_{K_D}^D) \\ \mathbf{b}_i|D_i = d &\sim \mathrm{N}(\boldsymbol{\beta}_d, \Sigma_d), d = 1, ..., K_D \\ C_i &\sim \text{Multinomial}(\pi_1^C, ..., \pi_{K_C}^C) \\ \sigma_i^2|C_i = c &\sim \text{log-N}(\mu_c, \tau^2), c = 1, ..., K_C \end{aligned} \tag{2.1}$$

  where, y_{ij} denotes the longitudinal covariate for the i^{th} subject at time t_{ij}, $j = 1, ..., n_i$, $i = 1, ..., n$, $\mathbf{b}_i$ are random effects that reflect the subject-level trajectory patterns, and σ_i^2 is the residual variance. D_i and C_i define the latent classes for the longitudinal means and individual variance memberships, respectively.

- The primary outcome model is a probit regression model:

$$\Phi^{-1}(\mathrm{P}(o_i = 1)) = \mathbf{Z}_i'\boldsymbol{\eta}, \tag{2.2}$$

  where o_i denotes the binary indicator of outcome, $\mathbf{Z}_i$ denotes a vector of the covariates in the probit model for subject i, $i = 1, \cdots, n$ and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. For the LC probit model, $\mathbf{Z}_i$ contains the latent class membership indicators D_i and C_i; while for the MSRE probit model, $\mathbf{Z}_i$ contains the specific shared random effects and residual variances. Other fully-observed baseline predictors may be included in $\mathbf{Z}_i$ as well.

2.2.1 Likelihood specification

Let $\boldsymbol{\phi} = (\pi_d^D,\ \boldsymbol{\beta}_d,\ \Sigma_d,\ d = 1,...,K_D;\ \pi_c^C,\ \mu_c,\ c = 1,...,K_C,\ \tau^2,\ \boldsymbol{\eta})'$, where we assume each parameter in $\boldsymbol{\phi}$ has independent prior distribution with the joint prior distribution denoted by $\pi(\boldsymbol{\phi})$, and $\mathbf{z}$ include all unobserved latent variables, i.e., $\mathbf{z} = (\mathbf{b}, \boldsymbol{\sigma}, \mathbf{C}, \mathbf{D})'$. The observed data $\mathbf{x}$ consists of the longitudinal profiles $\mathbf{y}_1, ..., \mathbf{y}_n$ and the observed outomes $o_1, ..., o_n$. Then the complete data likelihood of $\boldsymbol{\phi}$ based on the complete data $(\mathbf{x}, \mathbf{z})$ is given by

$$
\begin{aligned}
f(\mathbf{x},\mathbf{z}|\boldsymbol{\phi}) \propto \Bigg\{ \prod_{i=1}^{n} \Bigg[& \prod_{d} \left[\pi_d (2\pi)^{-\frac{r}{2}} |\Sigma_d|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\mathbf{b}_i - \boldsymbol{\beta}_d\right)' \Sigma_d \left(\mathbf{b}_i - \boldsymbol{\beta}_d\right) \right\} \right]^{\mathrm{I}(D_i = d)} \\
& \times \prod_{c} \left[\pi_c (2\pi\tau^2)^{-\frac{1}{2}} \sigma_i^{-2} \exp\left\{ -\frac{1}{2\tau^2} \left(\log\sigma_i^2 - \mu_c\right)^2 \right\} \right]^{\mathrm{I}(C_i = c)} \\
& \times \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2} \left\{ y_{ij} - f(\mathbf{b}_i; t_{ij}) \right\}^2 \right] \\
& \times \Phi(\mathbf{Z}_i'\boldsymbol{\eta})^{o_i} \left\{ 1 - \Phi(\mathbf{Z}_i'\boldsymbol{\eta}) \right\}^{1-o_i} \Bigg] \Bigg\} \pi(\boldsymbol{\phi})
\end{aligned}
\tag{2.3}
$$

2.2.2 Prior specification and posterior computation

We propose a fully Bayesian approach to estimate model parameters. Weakly informative conjugate priors were used whenever possible. For the mixture normal distribution of the random effects, we let $\boldsymbol{\beta}_d \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$, $\mathbf{V} = n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the linear regression estimator of $\mathbf{y}$ on the design matrix of $\mathbf{t}$ defined by $f(\cdot; t_{ij})$. This corresponds to a prior "single observation" data-driven inflated covariance structure centered at a null model, and avoids improper posteriors resulting from the possibility that some of the latent classes are not represented in the data (Elliott et al. 2005). For the the variance-covariance matrix for the random effects Σ_d, we use the prior proposed by Kass and Natarajan (2006): $\Sigma_d \sim$ Inverse-Wishart$(\mathrm{df} = r, \Lambda)$, where $\Lambda = r\left(\sum_{i=1}^{n} \widehat{\mathrm{Cov}}(\tilde{\mathbf{b}}_i)^{-1}/n\right)^{-1}$, $\tilde{\mathbf{b}}_i$ is given by OLS estimator of $\mathbf{b}_i$ for subject i and r is the dimension of $\mathbf{b}_i$. This prior restrains the eigenvalues of the variance-covariance ma-

trices away from 0, avoiding the improper posterior that can result from unbounded likelihoods when the variance-covariance matrix is unrestricted in normal mixture models (Day 1969).

For the mixture log normal distribution for the residual variances, we used diffuse priors: $\mu_c \sim \mathrm{N}(0, v)$, $\tau^{-2} \sim \mathrm{Gamma}(a, b)$ with $v = 1000$ and $a = b = .001$. For the class membership probabilities, we assume conjugate Dirichilet$(4, ..., 4)$ on both $\pi^C = (\pi_1^C, ..., \pi_{K_C}^C)$ and $\pi^D = (\pi_1^D, ..., \pi_{K_D}^D)$ (Frühwirth-Schnatter 2006); this is equivalent to assuming *a priori* 4 observations in each class, avoiding the existence of empty classes. Lastly, we let $\theta \sim \mathrm{N}(\mathbf{0}, (9/4)\mathbf{I})$ in the probit regression, where $(9/4)\mathbf{I}$ is chosen to bound the estimated outcome probabilities to be away from 0 and 1 (Garrett and Zeger 2000; Elliott et al. 2007, and Neelon et al., 2011).

Gibbs sampling is used to obtain draws from the corresponding posterior distributions. For $(\boldsymbol{\theta} \mid \mathbf{C}, \mathbf{D}, \mathbf{O})$ we use the Albert and Chib (1993) data augmentation method for probit regression models. The draws of $(\sigma_i^2 | C_i, \mu_c, \tau^2, \mathbf{b}_i, o_i, \{y_{ij}\}_j)$ for $i = 1, ..., n$ are obtained the inverse cumulative distribution method. The exact specifications of all priors and MCMC sampling procedures are provided in the Appendix A.

Three chains from diverse starting points were run and Gelman-Rubin statistics (Gelman et al. 2003) were used to assess the convergence of the MCMC chains. For the well-documented issue of "label switching" in finite mixture modeling (Redner and Walker 1984), we applied the post-processing relabeling algorithm (Stephens, 2000) in which all possible permutations of class assignments are considered at each iteration of the Gibbs sampler. The permutation of the class assignment was chosen to maximize the posterior probability so that the labeling of classes was consistent with the previous assignments. For our models with $K_D = 2$ or $K_C = 2$, there is little evidence of label switching and Stephens's relabeling algorithm converged very rapidly with the class labels from the initial MCMC output. But for models with more classes ($K_D = 3$ or $K_C = 3$), where label switching happens more frequently, we

initialized the class labels prior to running the algorithm by "reordering" the labels within chains based on visual inspection, since the convergence speed of Stephens's algorithm may depend on the quality of initial labels.

2.2.3 The choice of the number of classes

The choice of the number of latent classes is known to be a challenging problem; see discussions by McLachlan and Peel (2000) and Beunckens et al. (2008). We considered two criteria: the deviance information criterion (DIC), proposed by Spiegelhalter et al. (2002), and the logarithm of the pseudomarginal likelihood (LPML) (Geisser and Eddy 1979). DIC uses the discrepancy between the posterior mean of the deviance $\overline{D(\boldsymbol{\phi})} = E_{\boldsymbol{\phi}}\left\{-2\log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\}$ and the deviance evaluated at the posterior mean $D(\overline{\boldsymbol{\phi}}) = -2\log f\left\{\mathbf{x} \mid E(\boldsymbol{\phi} \mid \mathbf{x})\right\}$ to estimate the effective number of degrees of freedom in the model p_D. DIC is then given by the analog of the Akaike Information Criterion (AIC):

$$\mathrm{DIC}(\mathbf{x}) = \overline{D(\boldsymbol{\phi})} + p_D = 2\overline{D(\boldsymbol{\phi})} - D(\overline{\boldsymbol{\phi}}) = -4\mathrm{E}_{\boldsymbol{\phi}}\left\{\log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2\log f\left\{\mathbf{x} \mid \mathrm{E}(\boldsymbol{\phi} \mid \mathbf{x})\right\}.$$

In our setting, $f(\mathbf{x} \mid \boldsymbol{\phi})$, where $\mathbf{x} = (\mathbf{y}, \mathbf{o})'$ consisting of the fully-observed data, is not available in closed form; instead we use the approach outlined in Celeux et al. (2006) to obtain

$$\mathrm{DIC}(\mathbf{x}) = \mathrm{E}_{\mathbf{z}}\{\mathrm{DIC}(\mathbf{x}, \mathbf{z})\} = -4\mathrm{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\log f(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2\mathrm{E}_{\mathbf{z}}\left[\log f\left\{\mathbf{x}, \mathbf{z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z})\right\} \mid \mathbf{x}\right]$$

where integration over the latent variables $\mathbf{z} = (\mathbf{b}, \boldsymbol{\sigma}, \mathbf{C}, \mathbf{D})'$ is obtained via numercial methods; detailed are provided in Appendix A.

LPML corresponds to a Bayesian cross-validation measure, defined as LPML = $\sum_{i=1}^{n} \log(\mathrm{CPO}_i)$, where $\mathrm{CPO}_i = f(\mathbf{y}_i, o_i | \mathbf{y}_{(-i)}, \mathbf{o}_{(-i)})$ represents a cross-validated posterior predictive density for $(\mathbf{y}_i, o_i)$ given the data excluding $(\mathbf{y}_i, o_i)$ (denoted by $(\mathbf{y}_{(-i)}, \mathbf{o}_{(-i)})$).

The model with higher value of LPML provides better fit to the data (Ibrahim et al., 2001). Details of the LPML computation are also provided in Appendix A.

2.2.4 Goodness of fit evaluation

We assessed the model goodness of fit to the data in two ways. First, we examined the posterior predictive distributions (PPDs; Gelman et al. 2003), where a PPD p value close to 0.5 implies a good fit of the model to the data. For the longitudinal trajectories, we use repeated posterior draws to compute the PPD p values

$$\mathrm{P}\left\{T_i(\mathbf{y}_i;\mathbf{b}_i,\sigma_i^2) < T_i(\mathbf{y}_i^{\mathrm{rep}};\mathbf{b}_i,\sigma_i^2)\right\},$$

where we consider a χ^2-like statistic, $T_i(\mathbf{y}_i;\mathbf{b}_i,\sigma_i^2) = \sum_j \{y_{ij} - f(\mathbf{b}_i;t_{ij})\}^2/\sigma_i^2$ for subject i. For the outcome o_i, we compute $\mathrm{P}\left(T^{\mathrm{rep}} < T^{\mathrm{obs}}\right)$ using repeated posterior draws, where $T^{\mathrm{obs}} = n^{-1}\sum_i \mathrm{I}(o_i = 1)$, and $T^{\mathrm{rep}} = n^{-1}\sum_i o_i^{\mathrm{rep}}$ with o_i^{rep} following a Bernoulli distribution of the success probability $\Phi(\mathbf{Z}_i'\boldsymbol{\eta})$.

Second, we assessed the discriminatory ability of the model using receiver-operating characteristic (ROC) curves, in particular the area under the ROC curve (AUC). ROC curves plot true positive rate (TP) versus false positive rate (FP) for all possible cutoffs based on predicted $\mathrm{P}(o_i = 1) = \Phi(\mathbf{Z}_i'\boldsymbol{\eta})$ obtained from (4.3). The ROC curve and AUC were computed at each MCMC iteration using the ROCR package in R (Sing et al. 2005). The ROC is computed by ordering the observations $(i) = 1,\ldots,n$ so that $\hat{\mathrm{P}}(o_{(i)} = 1) \geq \hat{\mathrm{P}}(o_{(i+1)} = 1)$, computing changepoints $c = 2,\ldots,n_c$, $n_c \leq n$ where the observations change from positive to negative (i.e., $o_{(c-1)} = 1$, $o_{(c)} = 0$), and plotting $\sum_{(i)=1}^{c}(1 - o_{(i)})/\sum_{(i)=1}^{n}(1 - o_{(i)})$ on the x-axis versus $\sum_{(i)=1}^{c} o_{(i)}/\sum_{(i)=1}^{n} o_{(i)}$ on the y-axis. Area under the ROC is then computed using a trapezoidal approximation. The posterior mean AUC is calculated as the average AUC's across MCMC iterations. To obtain the posterior mean and the pointwise 95% credible interval of ROC curve,

we select 250 points equally spaced along the FP axis and take the vertical average or 95% quantiles of TP's at the 250 chosen points. This approach is referred to as vertical averaging of ROC curves at fixed FP rates by Fawcett (2006).

2.3 Simulations

In this section, we conduct simulation studies to evaluate the inferential properties of the LC and MSRE approaches when the model is correctly specified and when it is not (i.e., the data are generated under the LC model but analyzed using the MSRE model, and vice versa). We consider four scenarios for the longitudinal model with different levels of overlap between two mixtures in both mean profiles and variance patterns, crossed with two cross-sectional outcome models. For the longitudinal observations, we generate data from the following model with two mean profile and two variance classes,

$$\begin{aligned} \mathrm{y}_{ij}|\mathbf{b}_i, \sigma_i^2 &\sim \mathrm{N}(b_{0i} + b_{1i}t_{ij}, \sigma_i^2) \\ \mathbf{b}_i &\sim \pi_d \mathrm{N}(\beta_1, \Sigma_1) + (1-\pi_d)\mathrm{N}(\beta_2, \Sigma_2) \\ \log\sigma_i^2 &\sim \pi_c \mathrm{N}(\mu_1, \tau^2) + (1-\pi_c)\mathrm{N}(\mu_2, \tau^2) \end{aligned} \tag{2.4}$$

where $i = 1,...,200$ and $t_{ij} = 0,1,2,...,n_i$ with $n_i \equiv 20$. For $k = 1,2$, we let $\beta_k = (\beta_{k1}, \beta_{k2})'$ and

$$\Sigma_k = \begin{pmatrix} \omega_{k1}^2 & \rho_k \omega_{k1}\omega_{k2} \\ \rho_k \omega_{k1}\omega_{k2} & \omega_{k2}^2 \end{pmatrix}.$$

We let $\beta_1 = (0,0)'$ and $\beta_2 = (2\sqrt{2}, 2\sqrt{2})'$, $\rho_1 = 0$ and $\rho_2 = -0.6$, and $\mu_1 = -2$ and $\mu_2 = -.5$ in all scenarios. Thus the mean of the two bivariate normals differs by 4 throughout, while the mean log of the variances are separated by 1.5. Our four longitudinal model scenarios are defined by $(\omega^2, \tau^2)' = (2,.25),(1,.25),(2,.06)$, and $(1,.06)$, respectively, where $\omega = \omega_{11} = \omega_{12} = \omega_{21} = \omega_{22}$.

Figure 2.1 shows the 95% contours for the two components in the mixture mean profiles and the density plots of the mixture log-variance classes in each of the four scenarios: both mean and variance classes heavily overlapping (scenario # 1, the most challenging in terms of identification), only the variance classes heavily overlapping (scenario # 2), only the mean classes heavily overlapping (scenario # 3), neither the mean nor the variance classes heavily overlapping (scenario # 4, the least challenging in terms of identification). In all scenarios, we let $\pi_d = 0.35$ and $\pi_c = 0.65$.

For each of the simulation scenarios proposed for the longitudinal observations, the following two underlying probit models are considered for health outcome:

1. **latent class (LC) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \theta_0 + \theta_1 \mathrm{I}(D_i = 2) + \theta_2 \mathrm{I}(C_i = 2) + \theta_3 \mathrm{I}(D_i = 2, C_i = 2) \tag{2.5}$$

2. **multiple shared random effect (MSRE) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \gamma_0 + \gamma_1 b_{i0} + \gamma_2 b_{i1} + \gamma_3 \sigma_i^2 + \gamma_4 b_{i0}\sigma_i^2 + \gamma_5 b_{i1}\sigma_i^2 \tag{2.6}$$

where $D_i = 1$ corresponds to the mean profile class $\mathrm{N}((\mathbf{0},\mathbf{0})', \Sigma_1)$, and $C_i = 1$ corresponds to the variance class $\mathrm{N}(-2, \tau^2)$ in the longitudinal submodel (4.8). We replace $\boldsymbol{\eta}$ in (4.3) by $\boldsymbol{\theta}$ for the LC and by $\boldsymbol{\gamma}$ for the MSRE probit primary models to simplify the task of presentation; θ_0 and γ_0 are chosen for each scenario so that the outcome prevalence is approximately 50 percent.

To investigate the robustness to model mis-specification for each approach, we generated data from LC and MSRE primary models from equations (4.9) and (4.10) under each of the four longitudinal mixture scenarios, and then applied the approaches assuming the LC and MSRE structure to all generated data sets regardless of how the data were generated. (For scenarios in which the assumed model is

mis-specified, we generated observations from 10,000 subjects, obtained the corresponding maximum likelihood estimates constructed under the assumed model, and repeated the process 1000 times to obtain the averages of the estimated parameters. We then used these average estimates as the "true" parameters for that simulation scenario. This practice allows us to compare the robustness for the two different modeling considerations under the same simulation scenario.) For each scenario, we simulate 100 data sets and report the Monte Carlo bias, standard deviation (SD), mean squared error (MSE) and 95% credible interval coverage (95% COV) for the corresponding parameters under the correctly and incorrectly specified models. We also consider the correct classification of subjects to latent classes based on the posterior mean of $P(D_i = d \mid \mathbf{y}, \mathbf{o})$ and $P(C_i = c \mid \mathbf{y}, \mathbf{o})$.

Table 2.1 shows the bias, SD, MSE, and nominal coverage for the probit regression parameters in the primary outcome model; see Tables 2.5-2.8 for the complete results for all parameters in the model. Results were quite similar for scenarios #1 and #3: the mean parameters are estimated with bias and sometimes less than nominal coverage, even when the model is correctly specified, due to the difficulties in separating the two with overlapping distributions (the parameters associated with the first mean component tend to be better estimated since they compose a larger fraction of the sample). Incorrectly specifying the MSRE model has less impact on longitudinal mean estimation than incorrectly specifying the LC model. Incorrect specification damages the estimation of the mixture proportions for scenarios #1 and #3; even when correctly specified, the LC model still yields somewhat biased and undercovered estimates of the mixture proportions, again due to the difficulties in separating mixture components. In general, both methods performed well in estimation of longitudinal mean components for scenarios #2 and #4, although some bias and coverage issues remained for the slope of the second component in scenario #2. The variance components of the longitudinal model were generally well estimated

under all scenarios, regardless of model misspecification.

For estimation of the primary outcome model, the correctly specified LC model has some bias and modest undercoverage for the estimation of the longitudinal mean effect in scenarios #1 and #3, again resulting from the difficulty in separation of the clusters. The correctly specified MSRE model generally does well, with some very modest downward bias in the estimation of the individual residual variance effects. When incorrectly specified, the LC model is biased, overestimating the main effects of the longitudinal mean and variance class but underestimating the interaction effect between these two classes, which eventually leads to very poor coverage. The MSRE appears to be more robust, with little bias and good coverage of the implied target MSRE parameters under LC model. Under scenarios #2 and #4, estimation of the primary outcome model appears to be approximately correct under both the LC and MSRE models and largely robust to model misspecification, although using the LC model when the MSRE model is correct under scenario #2 yields some bias and reduction of coverage due to the imperfect separation of the variances mixtures. While our findings appear to suggest the use of MSRE approach over that of LC if the goal is the estimation of regression parameters in the primary outcome model, such a conclusion must be tempered by the fact that, if the LC model is actually correct, the target parameters under the MSRE model may be of less interest.

We next turn our attention to evaluating the correctness of outcome prediction and mis-classification of mixture classes. Again, generating data from the 4 longitudinal scenarios and LC vs MSRE primary models, we evaluate the true area under the ROC curves (AUC) as well as the corresponding values predicted by assuming the LC/MSRE models, respectively, in Table 2.2 (a), and compare those with the mean AUC obtained under the true values of the model. The misclassification rates for the mixture components are given in Table 2.2 (b). We denote $\tilde{C}_i = \text{argmax}_c \hat{\pi}^C_{ic}$ and $\tilde{D}_i = \text{argmax}_d \hat{\pi}^D_{id}$, where $\hat{\pi}^C_{ic}$ and $\hat{\pi}^D_{id}$ are the posterior mean assignment probabilities

of class-membership of the mean profile and variance class for subject i, respectively. We also use "$\tilde{D} = i|D = j$" indicate the case of the true $D = j$ while the predicted $\tilde{D} = i$ and equivalently define the cases for variable C.

When the model is correctly specified, the repeated sampling mean of the posterior mean AUC is fairly close to the repeated sampling mean of the AUC obtained using the true parameter value. When the LC model is correctly specified, the empirical 95% credible intervals of AUC when the mean mixtures overlap (scenarios #1 and #3) are wider than the truth, while the empirical 95% credible intervals of AUC when the mean mixtures are well-separated (scenarios #2 and #4) are of similar length to the truth, reflecting the difficulty in separating the mean mixture components in the LC model and therefore larger variabilities in the predictive power. The empirical 95% credible intervals of AUC under all 4 scenarios are all very close to the truth when the MSRE model is correctly specified, since the main outcome model is a function of the continuous subject-level mean and variance parameters and hence not affected by any degree of difficulty in separating the mean clusters. When the MSRE model is used and LC is the truth, it loses a little predictive power only when the mean mixtures overlap (scenarios #1 and #3); otherwise the results are similar to that of the correctly specified LC model. However, when the MSRE model is the truth and the LC model is used, it is interesting to note that, when it is more difficult to separate the components within the mean class and there exists considerable potential for misclassification as in scenario #1 and #3, the LC model can achieve either much higher or lower AUC than the AUC based on ML estimates of probit LC submodel given known class memberships. Figure 2.2 presents two typical data examples out of the total 100 data sets that have either very low or very high AUC estimated by the LC model. In both examples, the AUC's by the correctly specified MSRE model are very close to the truth. However, the much higher AUC by LC compared to the MLE estimate given known class membership suggests that the LC model has some ability

to create "artificial clusters" of very high predictive performance when MSRE model is the truth, while the MSRE AUC estimates suffered from imperfect prediction of the true probabilities of the outcome. On the other hand, when almost all subjects are being assigned to one mean class by the LC model, prediction of the outcome is solely dependent on the variance class and consequently the LC model had worse predictive performance than the ML estimates of probit LC submodel given known class membership. This phenomenon is an unique artifact of joint LC modeling, and may be due to the fact that the mixture classification is computed given both the longitudinal data $\mathbf{y}$ and the outcome $\mathbf{o}$. When the information from the longitudinal data is weak, the class membership can be "artificially" explained by the outcome $\mathbf{o}$. With the outcome $\mathbf{o}$ being binary, classes were artificially created to match the two outcome groups of $o = 0$ and $o = 1$. These two commonly observed cases presented in Figure 2.2 illustrate this phenomenon which leads to overly inflated variation for LC estimated AUC (i.e., considerably wide 95% credible intervals of AUC given in Table 2.2); what is observed in Figure 2.2 (a) and (b) also leads to the high misclassification rate of class membership. This phenomenon for joint LC modeling also happens when the data are generated with the LC model, but the effect is much less prominent. To our knowledge this has not been previously noted in the literature and could have strong implications for model interpretation.

Finally, in terms of misclassification rates of mean class membership, we see consistent patterns with the above discussion. The LC model tends to perform reasonably well when correctly specified; when LC is fit using MSRE data, the misclassification rates for the mean classes are high. In addition, the LC approach seems to perform a bit better than the MSRE when the model is correctly specified, where the latter deteriorates for Scenarios # 1 and # 3, regardless of whether the model is correct or not. This is consistent with what we have observed for the MSRE longitudinal mean parameters. When the mixture components are well separated, as in scenario

#4, both approaches perform well regardless of model specification. Variance classes are generally well estimated; there is some modest reduction in accuracy under Scenarios # 1 and # 2, where the separation is less complete, but results are robust to model misspecification.

2.4 Analysis of Penn Ovarian Aging Study data

One goal of the Penn Ovarian Aging Study is to determine to what extent the longitudinal Follicle Stimulating Hormone (FSH) observations are predictive of severe hot flashes during the follow-up period. After taking into account the population level non-linear trend of these annual measurements, we seek to evaluate whether subject-specific trends in FSH as well as underlying latent class membership determined by these trends are associated with severe hot flashes. Out of the 436 women in the study, we restrict our analysis to the 245 who a) reported having not experienced severe hot flash symptoms at baseline and b) had at least 3 follow-up measurements of FSH levels. Hormone values were treated as missing if a woman was pregnant, breast feeding or taking exogenous hormones during the 13 years of follow-up, since these circumstances will dramatically affect FSH levels. A total of 4,244 hormone values were observed during follow-up, with a minimum of 3 and a maximum of 26 per woman. Of the 245 women without severe hot flash symptoms at baseline, 118 (48.2%) had experienced severe hot flashes at least once during the study.

We let y_{ij} denote the mean detrended log(FSH) obtained from the lowess residuals (Figure 2.3) and o_i denote the severe hot flash indicator: $o_i = 1$ if had experienced severity score ≥ 2 at least once during study. Preliminary analysis using standard linear mixed effects (LME) models indicated that a random intercept and linear slope model is sufficient to capture the trends in the residual subject-level trajectories. In particular, orthogonality of design matrices can help improve chain mixing and con-

vergence. Thus we let $f(\mathbf{b}_i; t_{ij}) = b_{i0} + b_{i1} t_{ij}$, where t_{ij} is the first degree orthogonal polynomial of subject age, and b_{i0} and b_{1i} are the subject specific random intercepts and slopes, respectively. We then jointly model the FSH mean profile and variance to predict the onset of severe hot flash using our models in (4.2) and (4.3). We examine the strategies of using the primary probit LC and MSRE models under the joint modelling framework, as presented in section 2.2. We also adjust for additional baseline covariates log(BMI) and smoking status in both primary models.

For all the models considered in this section, we ran three MCMC chains of 50,000 iterations, with the first 10,000 iterations discarded as burn-in. We retained every 10^{th} draw to reduce autocorrelation. We assessed convergence of the chains using the Gelman-Rubin statistic $\hat{R}$, based on the ratio of the total variance across the chains to the within-chain variance. The maximum value among all population and subject-specific parameters was less than 1.1, indicating convergence.

Given the moderate sample size, we consider the models with a maximum of three components for both mean profile and variance classes: i.e., we let $K_D = 1,2,3$ and $K_C = 1,2,3$. As shown in Table 2.3, both the likelihood and LPML measures (note: log likelihood is computed as the posterior mean of the log likelihood) suggest that the models with multiple components in both mean profile and variance classes fit the data much better than models with single-component mean profile and variance classes under both joint MSRE and LC models. In constrast, DIC favors simpler models for both joint modeling strategies. The likelihood is essentially flat for the MSRE model for $K_D \geq 2, K_C \geq 2$; for the LC model, the likelihood continues to decrease as the number of mean and variance classes increase. Under the DIC, a model with one class for the mean trajectory and two for the variance fits best for both the MSRE and LC models, with a two-class model for both the mean and variance a somewhat close second for both the MSRE and LC models. Accounting for model predictive performance with LPML, the two-class mean and one-class variance provides the best fit

for the MSRE model, with the two-class mean and two-class variance a close second; similar results are obtained for the LC model, with the best (now two-class mean and variance) and second-best (two-class mean and one-class variance) switched from the MSRE model. Hence we focus on the $K_D = 1, K_C = 2$ class in the remainder of the analysis, although we report the results for the $K_D = 2, K_C = 2$ as well.

Figure 2.4 shows the mean trajectories and the variances for the $K_D = 1, K_C = 2$ model (MSRE on the left side, LC on the right side), indicating the bimodal nature of the posterior means of the individual variances. Table 2.4 reports the results for both the $K_D = 1, K_C = 2$ and $K_D = K_C = 2$ MSRE and LC models. There is little difference between the LC and MSRE model with respect to the estimation of the longitudinal submodel. The two class mean model separates the mean trajectories into two approximately equal-sized classes, with one a "null class" with slope and intercept near zero, the other a "high and rising" class with an slope of .21 under the MSRE model and .20 under the LC model, and an intercept of .16 under both models. All models suggest that a little more than one in five women (23% under MSRE, 21% under LC) belong to a low residual variance class (centered at .07(MSRE)/.06(LC)), while the remainder belong to a higher variance class centered at .32(MSRE)/.31(LC).

For the outcome, both the MSRE model and the LC model suggests positive and highly significant association between subject-level variance and risk of hot flash. For a non-smoking woman at mean BMI of 27.7 with detrended FSH slope and intercept at the population mean, the probability of experiencing a severe hot flash under the MSRE model with $K_D = 1, K_C = 2$ is 29.7% (19.0%, 41.7%) if her residual variance is at the Class 1 mean and 45.9% (37.8%, 54.0%) if her residual variance is at the Class 2 mean. The difference is greater under the LC model with $K_D = 1, K_C = 2$, where non-smoking women with mean BMI have a predicted hot flash probability of 17.2% (4.1%, 32.0%) if they are in the low variance class (Class 1) and 51.6% (43.0%, 60.2%) if they are in the high variance class (Class 2). The MSRE model also provides some

evidence that subjects with higher baseline FSH are at reduced risk of hot flash (95% credible interval barely excludes 0 under $K_D = 1, K_C = 2$ and barely includes 0 under $K_D = K_C = 2$); the LC model with $K_D = 1, K_C = 2$ by definition excludes mean trajectories from influencing the outcome, but even assuming $K_D = K_C = 2$ provides no evidence that the different mean classes are associated with the hot flash outcomes. Neither the MSRE or LC models provide any evidence of interaction between the subject-level residual trajectories and subject-level variances (interactions can only be estimated for the $K_D = K_C = 2$ LC model). All models provided marginal evidence to support smoking at baseline as contributing to higher risk of severe hot flash; this evidence was somewhat weaker under the LC model than under the MSRE model due to somewhat greater uncertainty in the smoking effect. Baseline BMI was not found to be important to the risk of severe hot flash under any model.

2.4.1 Model Fit and Model Checking

For final joint MSRE and LC models with $K_D = 1, K_C = 2$, the resulting histograms of the 245 PPD p values shown in Figure 2.3 indicate good fit of the LOWESS mean detrended log(FSH) values. The PPD p values range from 0.06 to 0.94 with median 0.53 under MSRE model and range from 0.09 to 0.92 with median 0.53 under the LC model. The top row in Figure 2.7 shows three randomly selected individual fits with PPD p values between 0.1 and 0.9, where both MSRE and LC models lead to almost identical and fairly good fits by visual inspection; the bottom row in Figure 2.7 shows the total three individual fits with PPD p values less than 0.1 or greater than 0.9 under either model, indicating that the small PPD p values appear to be driven by the individual outlying points and large p values are caused by the "almost perfect" fits. Web Figure 2.8 shows that only about 4% of the points are not covered by the subject-level posterior predictive intervals under both LC and MSRE models, again indicating good fit of the longitudinal trajectories except a few outlying obser-

vations. For the outcome submodel, the PPD p values are 0.519 and 0.533 under MSRE and LC models respectively, indicating a good fit to the onset of severe hot flash.

Further, we found that the MSRE model had somewhat greater predictive power as assessed by ROC curves (posterior mean of AUC=.685) than the LC model (posterior mean of AUC=.648) (Figure 2.5), although a comparison of AUCs suggests that the difference in the performance was not clearly delineated (ΔAUC is .037 (-0.040, 0.114)). Comparing the posterior mean of MSRE (.679) and LC models (.688) when $K_D = K_C = 2$ suggests that the mean trajectories do carry a modest amount of information to predict hot flash experience, since information about the trajectories can only be used in the LC outcome model when $K_D > 1$.

2.4.2 Summary

In our analysis of FSH and severe hot flash data from Penn ovarian aging study, there appeared to be weak evidence for two types of subject-level FSH residual trajectories (one “null" with slope and intercept near zero, the other with higher intercept and increasing values over age), and strong evidence for two clearly distinct underlying clusters of subject-level FSH residual variances (one low, one high). There was no strong evidence in either model that the mean profile was associated with such risks, although the MSRE model gave some evidence that subjects with higher predicted baseline (age 35) residual FSH measures were less likely to experience hot flashes. However, both the MSRE and LC modeling strategies revealed that higher within-subject variability (which can be interpreted as larger short-term fluctuations) contributes to significantly increased risk of severe hot flash while adjusting for baseline covariates of smoking and BMI. This important message from the data could have been overlooked if we had treated variability as a nuisance parameter, as is usually done even in joint longitudinal predictor/cross-sectional outcome settings.

In selecting between the LC and MSRE models for the Penn ovarian aging study data, measures of model fit give some preference to the MSRE model. Although, with the variances being the only key predictor outside of baseline smoking status, their clear separation into two components suggests that, per the simulation results, both the LC and MSRE modeling strategies will behave well regardless of the true outcome model generation mechanism.

2.5 Concluding Remarks

The joint LC and SRE models were originally proposed to link the important characteristics or features in the longitudinal mean trajectories to the primary health outcome of interest. While the joint LC model focuses on the clustering of those features (mean profile class), the joint SRE model focuses on the feature itself (random effect itself). In this paper, we consider an extended version of joint LC and SRE modeling by also considering the latent class in the variance structure. Therefore, our modeling approach allows discovery of predictive longitudinal features that distinguish between short-term variability (defined by variance class) and long-term trend (defined by mean profile class).

Both LC and MSRE models are built upon different assumptions that are generally difficult to verify without knowing the truth. However, relatively little attention has been paid to the potential impact of model misspecification in the joint modeling framework. This work provides guidance concerning the potential impact of choosing the wrong model to link both longitudinal and health outcome data under our extended LC and MSRE modeling strategy. In particular, we conducted simulation studies to investigate 1) the performance of LC and MSRE models, and 2) their robustness to model misspecification, when assuming two components in each latent class–mean profile class and variance class. For the longitudinal parameters, the MSRE model performed poorly when the components of the latent classes were not

well separated, and well when they were; these results were consistent whether the outcome model data were generated under the MSRE model or the LC model. In contrast, the LC model outperformed the MSRE model when the components of the latent classes were not well separated, but was more sensitive to model misspecification unless components were well-separated. In terms of the parameter estimation and outcome prediction in the primary outcome model, LC and MSRE modeling strategy performed differently. First, the MSRE approach was not as sensitive to small separations of latent classes as LC approach because correct class assignment is more critical to estimating the outcome model parameters under LC modeling strategy. Second, the MSRE approach was more robust to model misspecification: it enjoyed smaller elevated biases and maintained more sensible coverage than LC approach under model misspecification, while the misspecified MSRE AUC measure was almost identical to the truth while LC approach suffered considerable loss of predictive power when misspecified. When the LC model is misspecified, the loss of predictive power phenomenon can be explained by the potential loss of information when replacing a continuous variable by a discretized version in the regression analysis. On the other hand, even when the information is fully captured by the categorical variable (i.e., the MSRE model is misspecified relative to the LC model), the use of its continuous version can help fully recover the true information. These two advantages motivate us to recommend MSRE modeling strategy to achieve the goal of inference in the primary outcome model. However, the interpretation in MSRE model is not as easy as LC model, where one can relate the outcome risk to distinct features identified by the various latent classes.

This work can be extended in a variety of manners. For example, the assumption of a simple linear or low-order polynomial function for the longitudinal predictors could be relaxed to allow for a penalized spline or functional regression model. This may provide a more non-parametric parsing of "short term" and "long term" subject-

level variability, if sufficient data are available at the subject level to allow estimation of such terms. Also, developing methods to compensate for missing data in both the longitudinal predictors and outcome measures, particularly under non-missing-at-random mechanisms, will have practical application as well.

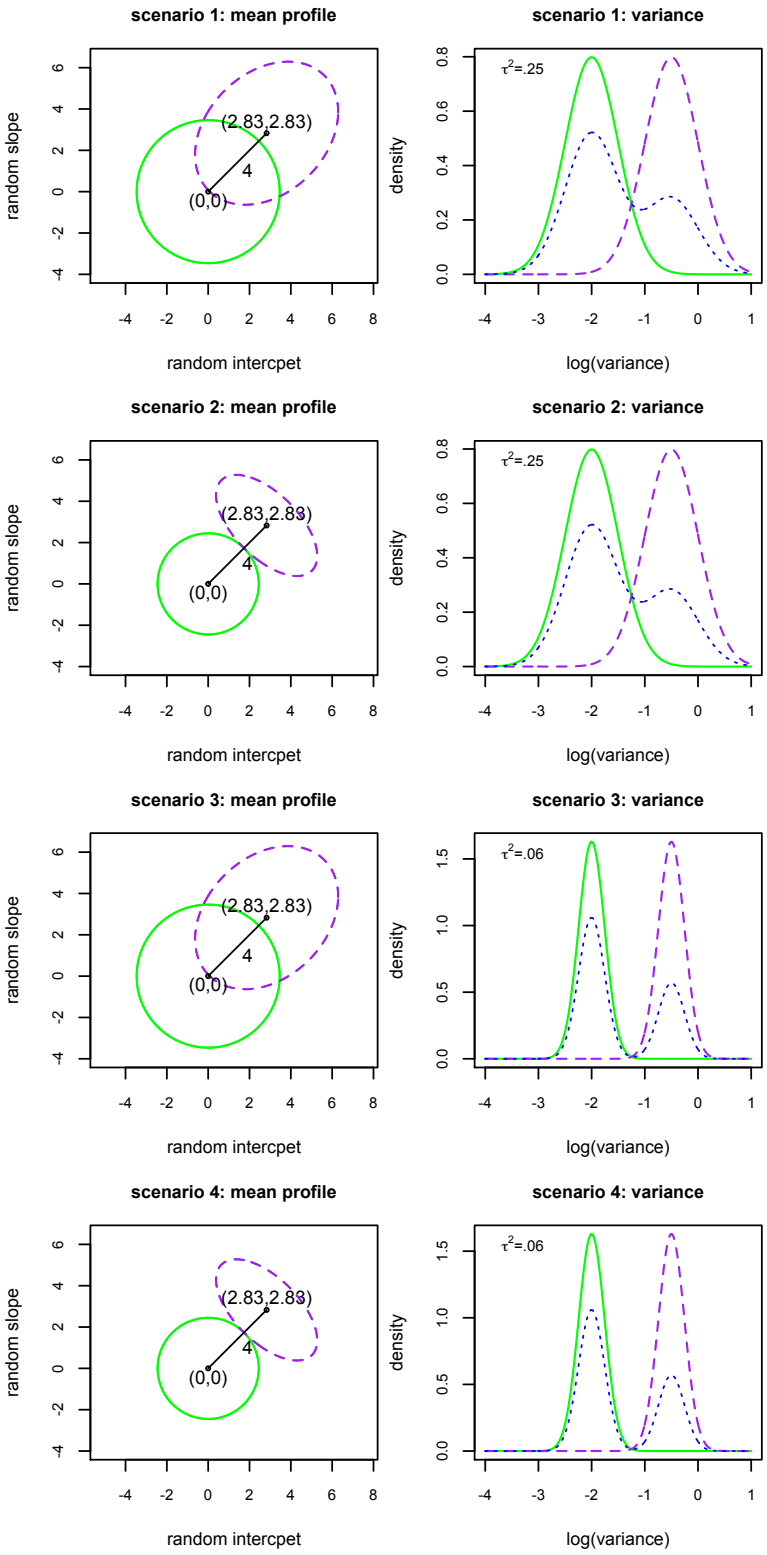

Figure 2.1: Simulation setup for the mean profiles and variance classes: left column: 95% contour plots of the two components for mean profile class; right column: density plots of the two components for variance class (dotted curves are the density curves for the variances).

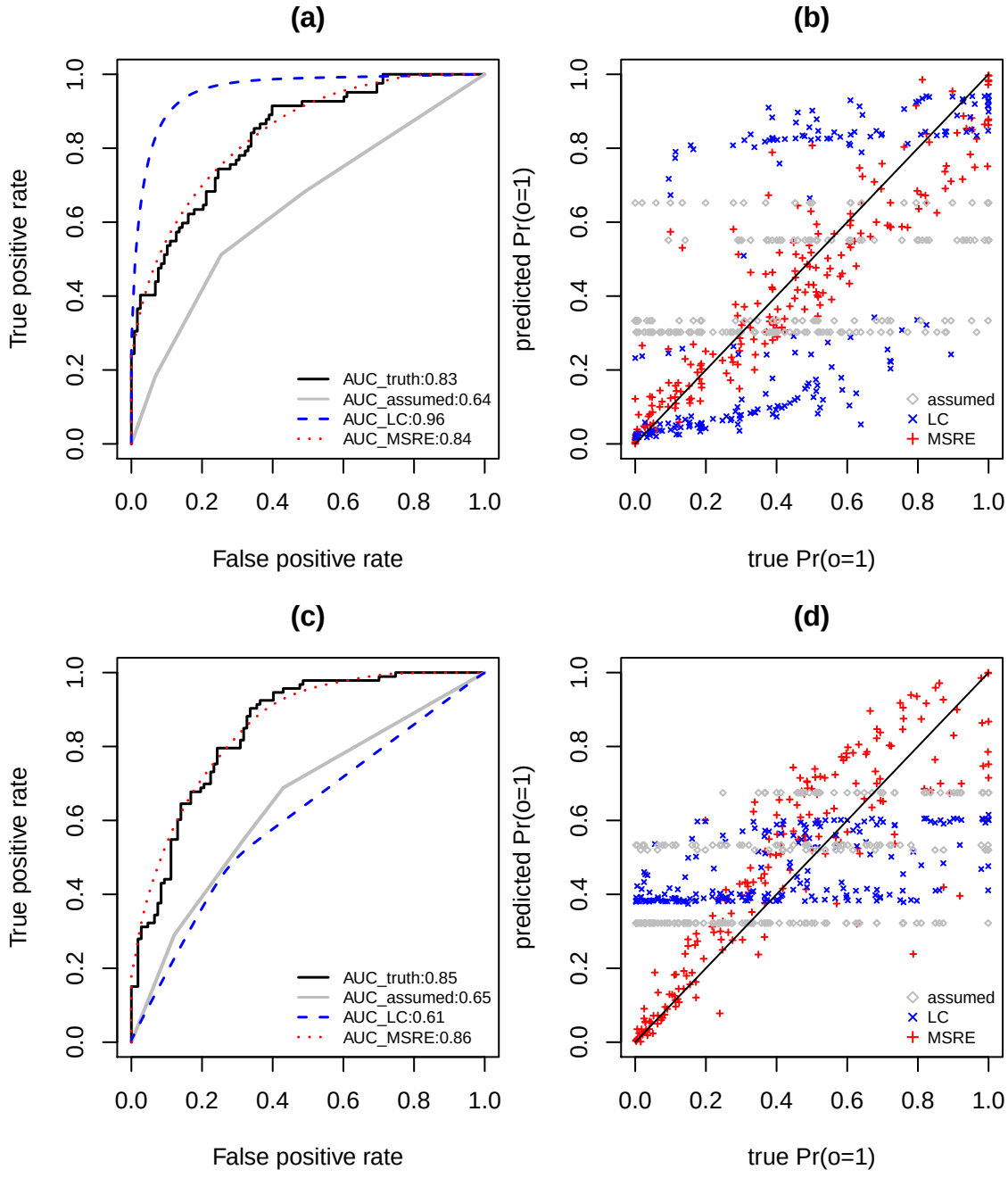

Figure 2.2: Two typical ROC's when the truth is joint MSRE model: (a) and (b) are from the data set where "artificial mean clusters" are created by joint LC model; (c) and (d) are from the data set when an almost empty mean cluster is created by joint LC model. Note: "assumed" refers to the ML estimates of LC probit submodel given known class memberships.

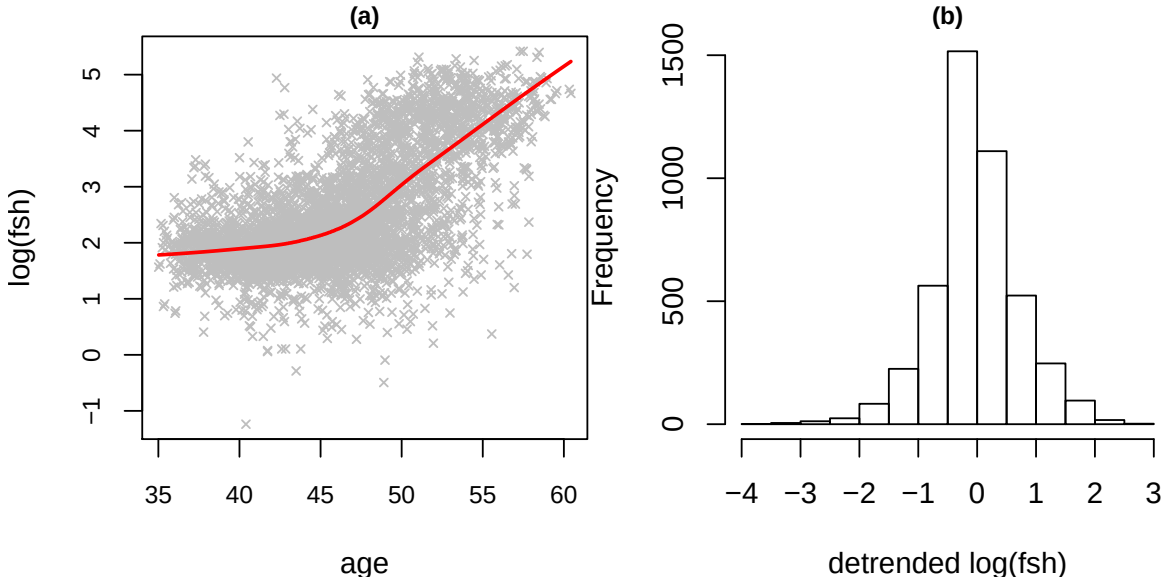

Figure 2.3: (a) Estimated population longitudinal trend by lowess; (b) histogram of detrended log(FSH).

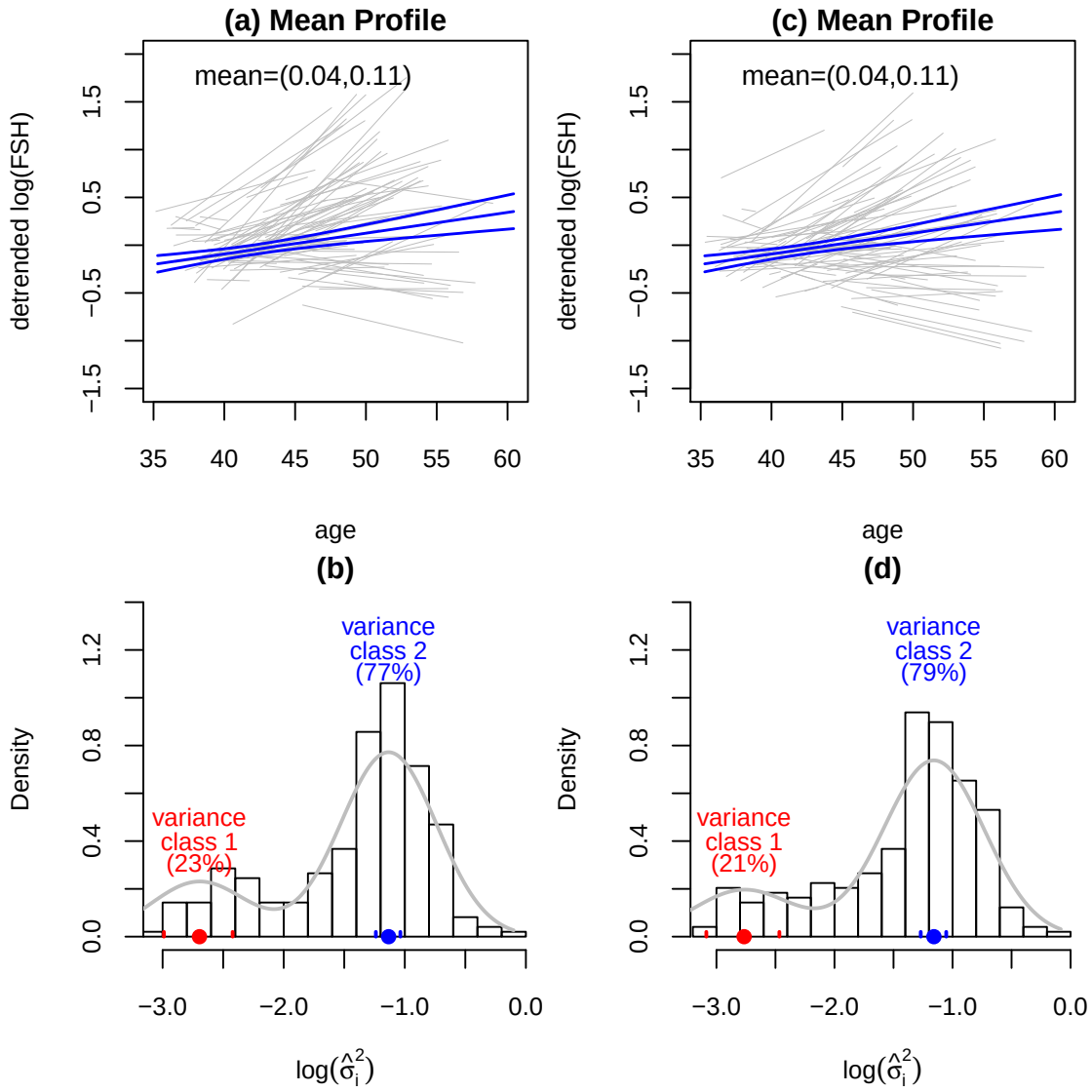

Figure 2.4: Posterior pointwise 95% credible intervals for the mean profile classes and the histograms of log-variances with $K_D = 1, K_C = 2$: (a) and (c): under joint MSRE model and (b) and (d): under joint LC model.

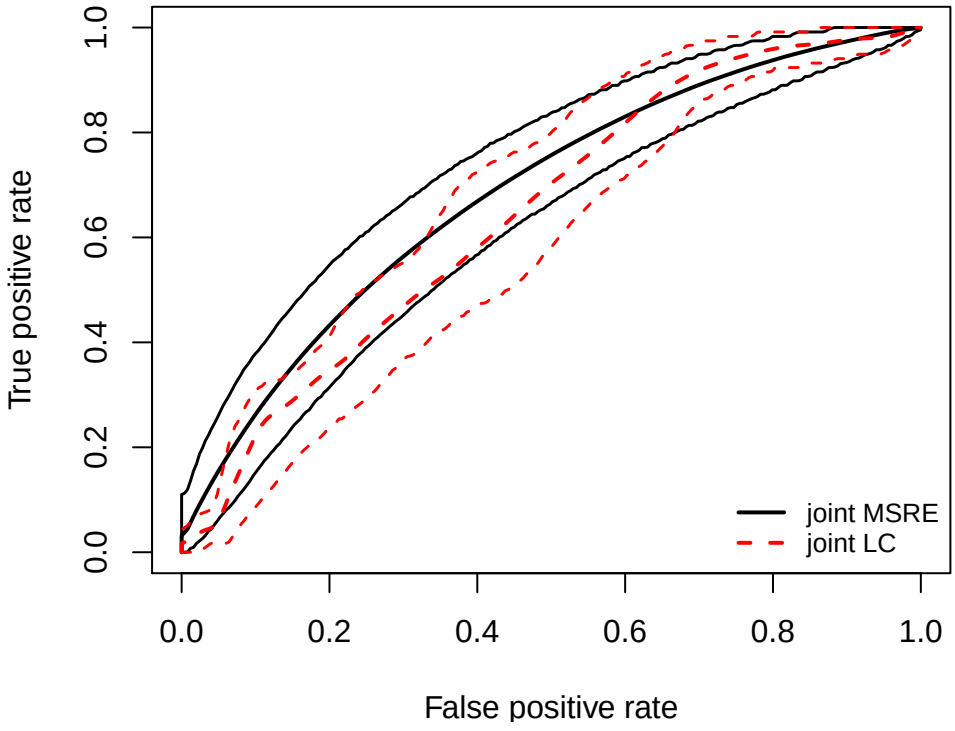

Figure 2.5: Posterior average of the receiver operating characteristic (ROC) curves under joint MSRE model (average AUC=0.685 with 95% CI (0.632, 0.736)) and joint LC model (average AUC=0.648 with 95% CI (0.590, 0.699)) both with $K_D 1, K_C = 2$ in the analysis of Penn ovarian aging data.

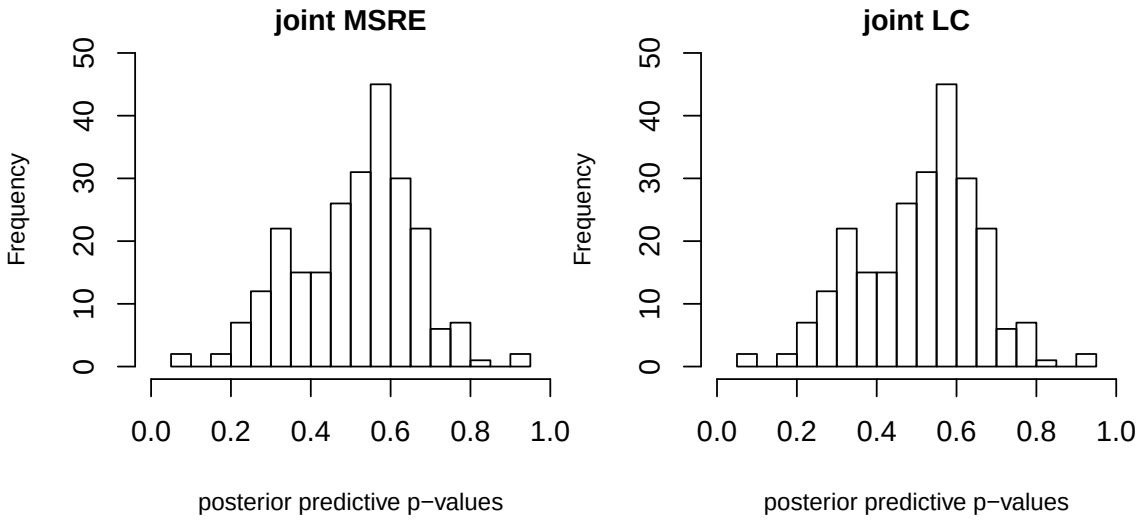

Figure 2.6: Posterior predictive p values under joint LC and MSRE models with $K_D = 1, K_C = 2$.

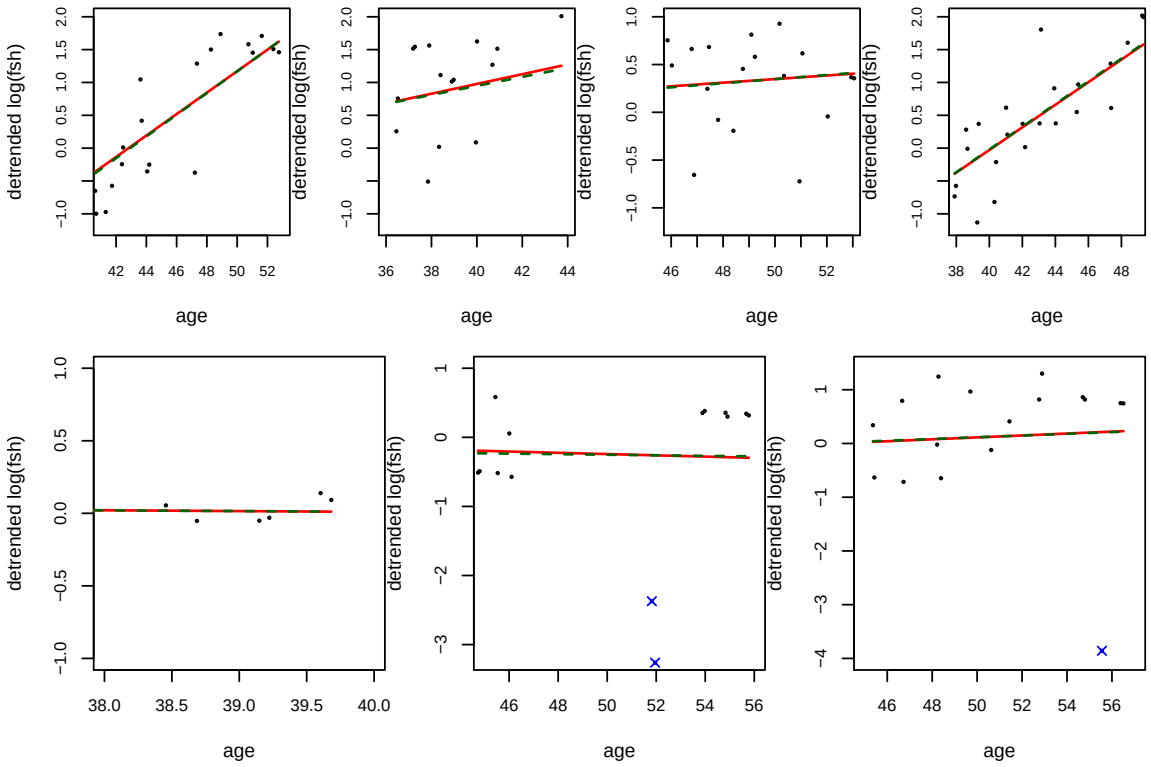

Figure 2.7: The individual fits in the analysis of Penn ovarian aging data with solid line: under joint MSRE model and dashed line: under joint LC model; top row: 3 randomly selected individual fits with PPD p values between 0.1 or great than 0.9 and bottom row: individuals with PPD p values less than 0.1 or great than 0.9.

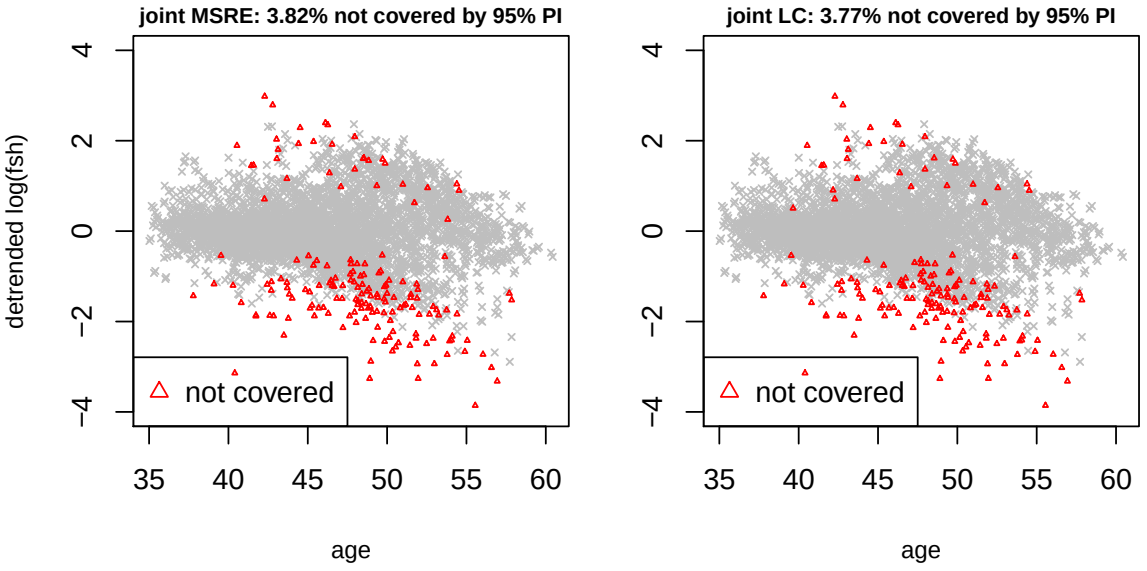

Figure 2.8: Scatter plot of detrended log(FSH) versus age with red points not being covered by subject-specific 95% posterior predictive intervals with $K_D = 1, K_C = 2$ in models for the analysis of Penn ovarian aging data: left: joint MSRE models and right: joint LC models.

Table 2.1: Simulation results from 100 datasets of size, $n = 200$.

(a) Generated from longitudinal scenario # 1

		True LC structure					**True MSRE structure**				
Assumed Structure		TRUE	BIAS	SD	RMSE	95% COV	TRUE	BIAS	SD	RMSE	95% COV
LC	θ_0	−0.80	−0.65	0.50	0.82	0.89	−0.40	−1.00	1.13	1.51	0.35
	θ_1	1.80	0.61	0.66	0.90	0.88	−0.11	2.40	2.34	3.35	0.35
	θ_2	−0.20	0.20	0.69	0.72	0.98	0.53	1.64	1.76	2.41	0.36
	θ_3	−0.30	−0.28	0.82	0.87	0.97	0.16	−3.69	3.73	5.24	0.35
MSRE	γ_0	−0.32	0.00	0.21	0.21	0.95	−1.00	0.19	0.24	0.30	0.92
	γ_1	0.20	0.01	0.11	0.11	0.96	1.00	−0.09	0.16	0.18	0.95
	γ_2	0.18	−0.01	0.11	0.11	0.96	−1.00	0.04	0.17	0.18	0.96
	γ_3	−0.22	−0.15	0.60	0.62	0.92	2.00	−0.52	0.58	0.78	0.87
	γ_4	−0.04	0.01	0.32	0.32	0.93	−2.00	0.29	0.36	0.46	0.90
	γ_5	−0.04	0.06	0.30	0.30	0.94	2.00	−0.14	0.38	0.41	0.95

(b) Generated from longitudinal scenario # 2

		True LC structure					**True MSRE structure**				
Assumed Structure		TRUE	BIAS	SD	RMSE	95% COV	TRUE	BIAS	SD	RMSE	95% COV
LC	θ_0	−0.80	−0.06	0.25	0.25	0.98	−0.48	−0.19	0.30	0.36	0.89
	θ_1	1.80	0.13	0.37	0.39	0.98	0.06	0.03	0.45	0.45	0.92
	θ_2	−0.20	−0.07	0.52	0.52	0.96	0.65	0.52	0.66	0.84	0.84
	θ_3	−0.30	−0.05	0.69	0.69	0.96	−0.08	−0.09	0.91	0.92	0.87
MSRE	γ_0	−0.66	0.01	0.24	0.24	0.98	−1.00	0.07	0.24	0.25	0.97
	γ_1	0.28	−0.02	0.12	0.12	0.96	1.00	−0.05	0.17	0.17	0.94
	γ_2	0.28	0.02	0.11	0.11	0.94	−1.00	0.02	0.17	0.17	0.93
	γ_3	−0.22	−0.28	0.56	0.62	0.97	2.00	−0.33	0.59	0.68	0.96
	γ_4	−0.05	0.12	0.33	0.35	0.89	−2.00	0.17	0.37	0.40	0.95
	γ_5	−0.05	−0.02	0.29	0.30	0.94	2.00	−0.06	0.39	0.39	0.97

(c) Generated from longitudinal scenario # 3

		True LC structure					**True MSRE structure**				
Assumed Structure		TRUE	BIAS	SD	RMSE	95% COV	TRUE	BIAS	SD	RMSE	95% COV
LC	θ_0	−0.80	−0.60	0.42	0.74	0.87	−0.41	−1.02	1.15	1.54	0.34
	θ_1	1.80	0.53	0.55	0.77	0.91	−0.12	2.52	2.28	3.40	0.31
	θ_2	−0.20	0.10	0.62	0.63	1.00	0.57	1.30	1.50	1.98	0.34
	θ_3	−0.30	−0.11	0.77	0.77	1.00	0.15	−3.42	3.15	4.65	0.36
MSRE	γ_0	−0.28	−0.01	0.20	0.20	0.96	−1.00	0.16	0.22	0.28	0.96
	γ_1	0.19	0.02	0.13	0.13	0.89	1.00	−0.09	0.15	0.18	0.92
	γ_2	0.20	0.00	0.14	0.14	0.93	−1.00	0.04	0.16	0.17	0.92
	γ_3	−0.37	−0.12	0.47	0.49	0.98	2.00	−0.42	0.53	0.68	0.94
	γ_4	−0.07	−0.10	−0.03	0.34	0.92	−2.00	0.25	0.35	0.43	0.94
	γ_5	−0.07	−0.04	0.04	0.36	0.93	2.00	−0.10	0.35	0.37	0.96

(d) Generated from longitudinal scenario # 4

		True LC structure					**True MSRE structure**				
Assumed Structure		TRUE	BIAS	SD	RMSE	95% COV	TRUE	BIAS	SD	RMSE	95% COV
LC	θ_0	−0.80	−0.01	0.19	0.19	0.97	−0.50	−0.05	0.24	0.25	0.91
	θ_1	1.80	0.00	0.25	0.25	0.99	0.06	0.05	0.30	0.30	0.90
	θ_2	−0.20	−0.08	0.51	0.52	0.95	0.69	0.10	0.38	0.40	0.94
	θ_3	−0.30	0.09	0.57	0.58	0.95	−0.08	−0.09	0.49	0.49	0.95
MSRE	γ_0	−0.62	−0.01	0.23	0.23	0.98	−1.00	0.14	0.25	0.29	0.93
	γ_1	0.29	−0.01	0.12	0.12	0.97	1.00	−0.08	0.15	0.17	0.90
	γ_2	0.29	0.03	0.13	0.13	0.94	−1.00	0.05	0.17	0.18	0.93
	γ_3	−0.36	−0.16	0.69	0.71	0.95	2.00	−0.44	0.58	0.73	0.93
	γ_4	−0.09	0.07	0.35	0.35	0.95	−2.00	0.22	0.36	0.43	0.90
	γ_5	−0.08	−0.02	0.30	0.30	0.97	2.00	−0.10	0.42	0.43	0.94

Table 2.2: (a) Mean Area under the ROC curves for the prediction of outcome and (b) misclassification rates for mixture membership. Left columns: data generated from the LC model; right columns: data generated from the MSRE model. LC-Assumed refers to AUC results obtained fitting a probit model using the known latent classes as predictors when the data are generated under the MSRE model, and similarly MSRE-Assumed refers to AUC results obtained fitting a probit model using the known random effects and variances as predictors when the data are generated under the LC model. "Percentile" refers to the 2.5 and 97.5 percentils of the AUC computed under the true parameters across the simulations; "95% CI" refers to mean of the lower and upper 95% credible intervals across simulations.

(a) Area under the ROC curves

	TRUE: joint LC model				**TRUE: joint MSRE model**			
	Scenario				**Scenario**			
	# 1	# 2	# 3	# 4	# 1	# 2	# 3	# 4
Truth								
mean	0.80	0.81	0.81	0.81	0.84	0.85	0.83	0.84
Percentile	(0.75, 0.86)	(0.75, 0.86)	(0.75, 0.87)	(0.75, 0.86)	(0.79, 0.89)	(0.80, 0.90)	(0.77, 0.88)	(0.78, 0.89)
LC								
mean	0.80	0.82	0.80	0.81	0.85	0.69	0.83	0.64
95% CI	(0.58, 0.91)	(0.75, 0.88)	(0.63, 0.92)	(0.75, 0.86)	(0.63, 0.97)	(0.58, 0.82)	(0.60, 0.96)	(0.56, 0.72)
LC Assumed								
mean	—	—	—	—	0.64	0.63	0.65	0.64
95% CI					(0.58, 0.70)	(0.56, 0.70)	(0.58, 0.70)	(0.58, 0.72)
MSRE								
mean	0.76	0.80	0.77	0.81	0.84	0.85	0.83	0.83
95% CI	(0.69, 0.83)	(0.73, 0.85)	(0.71, 0.85)	(0.74, 0.86)	(0.79, 0.89)	(0.79, 0.90)	(0.76, 0.88)	(0.77, 0.89)
MSRE Assumed								
mean	0.77	0.80	0.78	0.81	—	—	—	—
95% CI	(0.69, 0.83)	(0.74, 0.85)	(0.72, 0.85)	(0.75, 0.87)				

(b) Misclassification rates (%) for the mixture component membership

	TRUE: joint LC model				**TRUE: joint MSRE model**			
	$\tilde{D}$=2|D=1	$\tilde{D}$=1|D=2	$\tilde{C}$=2|C=1	$\tilde{C}$=1|C=2	$\tilde{D}$=2|D=1	$\tilde{D}$=1|D=2	$\tilde{C}$=2|C=1	$\tilde{C}$=1|C=2
				Scenario #1				
LC	10	2	5	6	19	31	4	8
MSRE	33	0	5	5	32	1	5	6
				Scenario #2				
LC	0	0	5	6	0	1	5	8
MSRE	0	0	5	6	0	0	5	6
				Scenario #3				
LC	9	2	1	2	18	32	1	2
MSRE	32	2	1	2	32	2	1	2
				Scenario #4				
LC	0	0	1	2	1	0	1	2
MSRE	1	0	1	2	1	0	1	2

Table 2.3: Model comparison statistics from different joint models for the analysis of Penn ovarian aging data.

Number of Mean Classes	**Number of Variance Classes**					
	LPML			**DIC**		
	1	2	3	1	2	3
	(a) Joint MSRE Model					
1	-3825.1	-3838.3	-3845.4	7293.3	**7240.1**	7433.1
2	**-3808.9**	-3817.8	-3828.6	7294.9	7249.4	7444.7
3	-3820.7	-3835.3	-3834.6	7377.4	7321.3	7516.2
	(b) Joint LC Model					
1	-3921.4	-3843.4	-3846.6	7289.8	**7245.5**	7472.2
2	-3830.2	**-3828.3**	-3832.4	7291.7	7260.0	7495.9
3	-3855.8	-3846.8	-3842.9	7375.7	7347.6	7567.2

Table 2.4: Posterior estimates of the model parameters under joint MSRE and LC models with $K_D = 1, 2$ and $K_C = 1, 2$ for the analysis of Penn ovarian aging data.

	MSRE Model						**LC Model**					
	$K_D = 1, K_C = 2$			$K_D = 2, K_C = 2$			$K_D = 1, K_C = 2$			$K_D = 2, K_C = 2$		
	mean	se	95% CI	mean	se	95% CI	mean	se	95% CI	mean	se	95% CI
β_{11}	0.041	0.031	(-0.02, 0.102)	-0.078	0.045	(-0.167, 0.01)	0.039	0.031	(-0.022, 0.101)	-0.075	0.051	(-0.165, 0.019)
β_{12}	0.111	0.025	(0.061, 0.16)	0.006	0.037	(-0.068, 0.079)	0.109	0.025	(0.06, 0.158)	0.009	0.04	(-0.062, 0.084)
β_{21}				0.158	0.071	(0.025, 0.299)				0.156	0.075	(0.006, 0.309)
β_{22}				0.212	0.058	(0.102, 0.329)				0.201	0.061	(0.083, 0.324)
ω_{11}^2	0.2	0.023	(0.158, 0.249)	0.078	0.022	(0.037, 0.125)	0.201	0.023	(0.159, 0.249)	0.08	0.041	(0.039, 0.132)
ω_{12}^2	0.102	0.013	(0.079, 0.13)	0.045	0.018	(0.013, 0.085)	0.103	0.013	(0.08, 0.132)	0.039	0.021	(0.011, 0.081)
ω_{21}^2				0.334	0.067	(0.226, 0.492)				0.335	0.074	(0.215, 0.497)
ω_{22}^2				0.155	0.036	(0.094, 0.237)				0.171	0.04	(0.102, 0.259)
ρ_1	0.667	0.056	(0.549, 0.765)	0.935	0.041	(0.831, 0.983)	0.668	0.057	(0.549, 0.769)	0.913	0.08	(0.759, 0.979)
ρ_2				0.46	0.125	(0.185, 0.665)				0.473	0.135	(0.189, 0.693)
π_1^D				0.524	0.083	(0.352, 0.677)				0.525	0.086	(0.346, 0.687)
μ_1	-2.696	0.145	(-2.989, -2.422)	-2.696	0.149	(-2.999, -2.411)	-2.766	0.158	(-3.086, -2.467)	-2.772	0.165	(-3.1, -2.461)
μ_2	-1.134	0.051	(-1.239, -1.037)	-1.141	0.053	(-1.249, -1.038)	-1.158	0.055	(-1.269, -1.053)	-1.168	0.06	(-1.271, -1.064)
τ^2	0.16	0.037	(0.098, 0.24)	0.161	0.038	(0.098, 0.245)	0.185	0.042	(0.114, 0.278)	0.186	0.041	(0.117, 0.274)
π_1^C	0.227	0.039	(0.153, 0.307)	0.226	0.04	(0.152, 0.306)	0.212	0.039	(0.138, 0.291)	0.209	0.04	(0.138, 0.29)
γ_0 (intercept)	-0.471	0.945	(-2.331, 1.372)	-0.546	0.937	(-2.368, 1.284)						
γ_1 (log(BMI))	-0.064	0.282	(-0.611, 0.488)	-0.041	0.279	(-0.59, 0.506)						
γ_2 (smoking)	0.374	0.189	(0.004, 0.745)	0.377	0.186	(0.011, 0.744)						
$\gamma_3(b_{i0})$	-0.949	0.492	(-1.946, -0.012)	-0.809	0.47	(-1.75, 0.072)						
$\gamma_4(b_{i1})$	0.589	0.581	(-0.546, 1.71)	0.523	0.562	(-0.565, 1.656)						
$\gamma_5(\sigma_i^2)$	1.644	0.596	(0.504, 2.857)	1.708	0.605	(0.548, 2.922)						
$\gamma_6(b_{i0} \times \sigma_i^2)$	0.138	1.17	(-2.104, 2.442)	0.063	1.146	(-2.107, 2.338)						
$\gamma_7(b_{i1} \times \sigma_i^2)$	0.644	1.329	(-1.925, 3.209)	0.361	1.324	(-2.21, 2.898)						
θ_0 (intercept)							-0.911	0.829	(-2.518, 0.679)	-1.429	0.999	(-3.548, 0.396)
θ_1 (log(BMI))							-0.026	0.239	(-0.487, 0.441)	-0.088	0.204	(-0.485, 0.306)
θ_2 (smoking)							0.334	0.184	(-0.022, 0.694)	0.383	0.204	(-0.009, 0.793)
θ_3 (D=2)										1.273	1.369	(-1.442, 4.093)
θ_4 (C=2)							1.038	0.362	(0.463, 1.805)	1.928	0.941	(0.527, 4.063)
θ_5 (D=2,C=2)										-1.618	1.484	(-4.607, 1.341)

Table 2.5: Simulation results from 100 datasets of size, $n = 200$, generated from longitudinal scenario # 1 and the primary probit (a) LC, (b) MSRE models. Left columns: fitted assuming the LC model; right column: fitted assuming the MSRE model.

(a) TRUE: joint LC model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	−0.19	−0.19	0.34	0.39	0.91	−0.53	−0.53	0.95	1.08	0.87
β_{12}	0.00	−0.04	−0.04	0.22	0.23	0.96	0.06	0.06	0.36	0.37	0.92
β_{21}	2.83	2.39	−0.44	0.23	0.49	0.40	1.79	−1.04	0.18	1.06	0.00
β_{22}	2.83	2.31	−0.52	0.21	0.56	0.18	1.73	−1.10	0.15	1.11	0.00
ω^2_{11}	2.00	1.83	−0.17	0.62	0.64	0.90	0.60	−1.40	0.63	1.54	0.57
ω^2_{12}	2.00	1.90	−0.10	0.67	0.68	0.89	0.44	−1.56	1.03	1.87	0.19
ω^2_{21}	2.00	2.56	0.56	0.48	0.74	0.74	3.66	1.66	0.38	1.71	0.03
ω^2_{22}	2.00	2.65	0.65	0.52	0.83	0.65	3.75	1.75	0.40	1.79	0.04
ρ_1	0.00	−0.12	−0.12	0.22	0.26	0.94	−0.72	−0.72	0.20	0.75	0.80
ρ_2	0.60	0.66	0.06	0.06	0.09	0.77	0.69	0.09	0.04	0.10	0.44
π_d	0.35	0.27	−0.08	0.07	0.11	0.78	0.05	−0.30	0.05	0.30	0.03
μ_1	−2.00	−1.95	0.05	0.11	0.12	0.98	−1.98	0.02	0.09	0.09	0.99
μ_2	−0.50	−0.61	−0.11	0.22	0.25	0.87	−0.58	−0.08	0.19	0.20	0.90
τ^2	0.25	0.33	0.08	0.12	0.14	0.90	0.30	0.05	0.10	0.11	0.95
π_c	0.65	0.63	−0.02	0.05	0.06	0.97	0.63	−0.02	0.04	0.05	0.98
θ_0	−0.80	−1.45	−0.65	0.50	0.82	0.89					
θ_1	1.80	2.41	0.61	0.66	0.90	0.88					
θ_2	−0.20	0.00	0.20	0.69	0.72	0.98					
θ_3	−0.30	−0.58	−0.28	0.82	0.87	0.97					
γ_0	−0.32						−0.32	0.00	0.21	0.21	0.95
γ_1	0.19						0.20	0.01	0.11	0.11	0.96
γ_2	0.18						0.17	−0.01	0.11	0.11	0.96
γ_3	−0.22						−0.38	−0.15	0.60	0.62	0.92
γ_4	−0.04						−0.02	0.01	0.32	0.32	0.93
γ_5	−0.04						0.02	0.06	0.30	0.30	0.94

(b) TRUE: joint MSRE model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	0.85	0.85	0.91	1.25	0.33	−0.53	−0.53	1.03	1.16	0.85
β_{12}	0.00	1.27	1.27	0.95	1.59	0.36	0.13	0.13	0.45	0.47	0.87
β_{21}	2.83	2.02	−0.81	0.29	0.86	0.14	1.81	−1.02	0.25	1.05	0.02
β_{22}	2.83	1.05	−1.78	0.57	1.87	0.00	1.69	−1.14	0.33	1.18	0.00
ω^2_{11}	2.00	2.78	0.78	1.69	1.87	0.24	0.70	−1.30	0.87	1.56	0.57
ω^2_{12}	2.00	2.28	0.28	1.43	1.46	0.16	0.49	−1.51	1.08	1.86	0.18
ω^2_{21}	2.00	3.26	1.26	0.58	1.39	0.38	3.60	1.60	0.54	1.69	0.06
ω^2_{22}	2.00	4.07	2.07	0.63	2.16	0.06	3.70	1.70	0.55	1.79	0.06
ρ_1	0.00	0.25	0.25	0.71	0.75	0.32	−0.71	−0.71	0.27	0.76	0.74
ρ_2	0.60	0.77	0.17	0.08	0.19	0.29	0.68	0.08	0.13	0.15	0.46
π_d	0.35	0.46	0.11	0.31	0.33	0.03	0.07	−0.28	0.13	0.31	0.04
μ_1	−2.00	−1.93	0.07	0.08	0.10	0.86	−1.97	0.03	0.08	0.09	1.00
μ_2	−0.50	−0.47	0.03	0.15	0.16	0.90	−0.56	−0.06	0.18	0.19	0.91
τ^2	0.25	0.30	0.05	0.08	0.10	0.93	0.30	0.05	0.09	0.10	0.93
π_c	0.65	0.67	0.02	0.04	0.05	0.92	0.63	−0.02	0.04	0.05	0.98
θ_0	−0.40	−1.40	−1.00	1.13	1.51	0.35					
θ_1	−0.11	2.29	2.40	2.34	3.35	0.35					
θ_2	0.53	2.17	1.64	1.76	2.41	0.36					
θ_3	0.16	−3.53	−3.69	3.73	5.24	0.35					
γ_0	−1.00						−0.81	0.19	0.24	0.30	0.92
γ_1	1.00						0.91	−0.09	0.16	0.18	0.95
γ_2	−1.00						−0.96	0.04	0.17	0.18	0.96
γ_3	2.00						1.48	−0.52	0.58	0.78	0.87
γ_4	−2.00						−1.71	0.29	0.36	0.46	0.90
γ_5	2.00						1.86	−0.14	0.38	0.41	0.95

Table 2.6: Simulation results from 100 datasets of size, $n = 200$, generated from longitudinal scenario # 2 and the primary probit (a) LC, (b) MSRE models. Left columns: fitted assuming the LC model; right column: fitted assuming the MSRE model.

(a) TRUE: joint LC model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	0.01	0.01	0.13	0.13	0.93	0.01	0.01	0.13	0.13	0.94
β_{12}	0.00	−0.02	−0.02	0.11	0.11	0.98	−0.02	−0.02	0.11	0.11	0.99
β_{21}	2.83	2.91	0.08	0.09	0.12	0.89	2.91	0.08	0.09	0.12	0.89
β_{22}	2.83	2.69	−0.14	0.09	0.17	0.64	2.69	−0.14	0.09	0.17	0.65
ω^2_{11}	1.00	0.99	−0.01	0.16	0.16	0.98	0.99	−0.01	0.16	0.16	0.97
ω^2_{12}	1.00	1.03	0.03	0.19	0.19	0.93	1.03	0.03	0.20	0.20	0.93
ω^2_{21}	1.00	0.99	−0.01	0.14	0.14	0.92	0.99	−0.01	0.15	0.15	0.89
ω^2_{22}	1.00	1.01	0.01	0.12	0.12	0.97	1.01	0.01	0.12	0.12	0.96
ρ_1	0.00	0.00	0.00	0.12	0.12	0.98	0.00	0.00	0.12	0.12	0.98
ρ_2	−0.60	−0.59	0.01	0.07	0.07	0.93	−0.59	0.01	0.07	0.07	0.93
π_d	0.35	0.35	0.00	0.03	0.03	0.94	0.35	0.00	0.03	0.03	0.93
μ_1	−2.00	−1.95	0.05	0.11	0.12	0.97	−1.99	0.01	0.08	0.08	0.98
μ_2	−0.50	−0.62	−0.12	0.21	0.24	0.93	−0.56	−0.06	0.15	0.17	0.93
τ^2	0.25	0.33	0.08	0.13	0.15	0.91	0.29	0.04	0.09	0.10	0.92
π_c	0.65	0.63	−0.02	0.05	0.05	0.97	0.63	−0.02	0.05	0.05	0.97
θ_0	−0.80	−0.86	−0.06	0.25	0.25	0.98					
θ_1	1.80	1.93	0.13	0.37	0.39	0.98					
θ_2	−0.20	−0.27	−0.07	0.52	0.52	0.96					
θ_3	−0.30	−0.35	−0.05	0.69	0.69	0.96					
γ_0	−0.66						−0.65	0.01	0.24	0.24	0.98
γ_1	0.28						0.26	−0.02	0.12	0.12	0.96
γ_2	0.28						0.30	0.02	0.11	0.11	0.94
γ_3	−0.22						−0.49	−0.28	0.56	0.62	0.97
γ_4	−0.05						0.07	0.12	0.33	0.35	0.89
γ_5	−0.05						−0.07	−0.02	0.29	0.30	0.94

(b) TRUE: joint MSRE model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	0.02	0.02	0.17	0.17	0.93	0.01	0.01	0.13	0.13	0.94
β_{12}	0.00	−0.01	−0.01	0.18	0.18	0.98	−0.02	−0.02	0.11	0.11	0.99
β_{21}	2.83	2.91	0.08	0.12	0.14	0.88	2.91	0.08	0.09	0.12	0.89
β_{22}	2.83	2.68	−0.14	0.12	0.19	0.65	2.69	−0.14	0.09	0.17	0.65
ω^2_{11}	1.00	1.00	0.00	0.18	0.18	0.96	0.99	−0.01	0.16	0.16	0.97
ω^2_{12}	1.00	1.06	0.06	0.34	0.35	0.91	1.03	0.03	0.20	0.20	0.92
ω^2_{21}	1.00	0.98	−0.02	0.16	0.16	0.89	0.99	−0.01	0.15	0.15	0.91
ω^2_{22}	1.00	1.00	0.00	0.14	0.14	0.94	1.01	0.01	0.12	0.12	0.95
ρ_1	0.00	0.01	0.01	0.13	0.13	0.97	0.01	0.01	0.12	0.12	0.97
ρ_2	−0.60	−0.59	0.01	0.07	0.07	0.93	−0.59	0.01	0.07	0.07	0.95
π_d	0.35	0.36	0.01	0.05	0.05	0.92	0.35	0.00	0.03	0.03	0.93
μ_1	−2.00	−1.94	0.06	0.10	0.11	0.95	−1.99	0.01	0.08	0.08	0.97
μ_2	−0.50	−0.55	−0.05	0.17	0.18	0.92	−0.54	−0.04	0.15	0.16	0.93
τ^2	0.25	0.33	0.08	0.11	0.14	0.87	0.29	0.04	0.09	0.10	0.92
π_c	0.65	0.65	0.00	0.05	0.05	0.96	0.63	−0.02	0.05	0.05	0.98
θ_0	−0.48	−0.68	−0.19	0.30	0.36	0.89					
θ_1	0.06	0.09	0.03	0.45	0.45	0.92					
θ_2	0.65	1.17	0.52	0.66	0.84	0.84					
θ_3	−0.08	−0.18	−0.09	0.91	0.92	0.87					
γ_0	−1.00						−0.93	0.07	0.24	0.25	0.97
γ_1	1.00						0.95	−0.05	0.17	0.17	0.94
γ_2	−1.00						−0.98	0.02	0.17	0.17	0.93
γ_3	2.00						1.67	−0.33	0.59	0.68	0.96
γ_4	−2.00						−1.83	0.17	0.37	0.40	0.95
γ_5	2.00						1.94	−0.06	0.39	0.39	0.97

Table 2.7: Simulation results from 100 datasets of size, $n = 200$, generated from longitudinal scenario # 3 and the primary probit (a) LC, (b) MSRE models. Left columns: fitted assuming the LC model; right column: fitted assuming the MSRE model.

(a) TRUE: joint LC model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	−0.14	−0.14	0.27	0.31	0.91	−0.49	−0.49	1.11	1.21	0.80
β_{12}	0.00	−0.04	−0.04	0.20	0.20	0.97	0.12	0.12	0.54	0.55	0.81
β_{21}	2.83	2.44	−0.39	0.20	0.44	0.47	1.81	−1.02	0.28	1.06	0.06
β_{22}	2.83	2.34	−0.49	0.19	0.53	0.22	1.66	−1.17	0.43	1.25	0.01
ω^2_{11}	2.00	1.88	−0.12	0.54	0.56	0.96	0.68	−1.32	0.91	1.61	0.48
ω^2_{12}	2.00	1.93	−0.07	0.55	0.55	0.95	0.38	−1.62	0.94	1.87	0.16
ω^2_{21}	2.00	2.49	0.49	0.48	0.68	0.80	3.62	1.62	0.61	1.73	0.08
ω^2_{22}	2.00	2.61	0.61	0.51	0.79	0.71	3.69	1.69	0.68	1.82	0.02
ρ_1	0.00	−0.10	−0.10	0.23	0.25	0.93	−0.71	−0.71	0.30	0.77	0.68
ρ_2	0.60	0.65	0.05	0.06	0.08	0.86	0.66	0.06	0.20	0.21	0.43
π_d	0.35	0.28	−0.07	0.07	0.09	0.76	0.07	−0.28	0.15	0.32	0.02
μ_1	−2.00	−2.00	0.00	0.04	0.04	0.95	−2.00	0.00	0.04	0.04	0.96
μ_2	−0.50	−0.50	0.00	0.06	0.06	0.95	−0.51	−0.01	0.06	0.06	0.91
τ^2	0.06	0.06	0.00	0.02	0.02	0.89	0.06	0.00	0.02	0.02	0.94
π_c	0.65	0.65	0.00	0.03	0.03	0.96	0.65	0.00	0.03	0.03	0.97
θ_0	−0.80	−1.40	−0.60	0.42	0.74	0.87					
θ_1	1.80	2.33	0.53	0.55	0.77	0.91					
θ_2	−0.20	−0.10	0.10	0.62	0.63	1.00					
θ_3	−0.30	−0.41	−0.11	0.77	0.77	1.00					
γ_0	−0.28						−0.30	−0.01	0.20	0.20	0.96
γ_1	0.19						0.22	0.02	0.13	0.13	0.89
γ_2	0.20						0.20	0.00	0.14	0.14	0.93
γ_3	−0.37						−0.49	−0.12	0.47	0.49	0.98
γ_4	−0.07						−0.10	−0.03	0.34	0.34	0.92
γ_5	−0.07						−0.04	0.04	0.36	0.36	0.93

(b) TRUE: joint MSRE model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	0.77	0.77	1.15	1.39	0.26	−0.50	−0.50	1.01	1.12	0.79
β_{12}	0.00	1.26	1.26	0.93	1.57	0.30	0.14	0.14	0.52	0.54	0.80
β_{21}	2.83	2.02	−0.80	0.27	0.85	0.13	1.81	−1.02	0.27	1.06	0.04
β_{22}	2.83	1.03	−1.80	0.62	1.90	0.01	1.65	−1.18	0.47	1.27	0.01
ω^2_{11}	2.00	2.85	0.85	1.67	1.88	0.24	0.68	−1.32	0.81	1.55	0.46
ω^2_{12}	2.00	2.39	0.39	1.44	1.49	0.13	0.42	−1.58	1.04	1.89	0.15
ω^2_{21}	2.00	3.14	1.14	0.82	1.40	0.40	3.61	1.61	0.71	1.76	0.08
ω^2_{22}	2.00	3.98	1.98	0.91	2.18	0.12	3.66	1.66	0.77	1.83	0.03
ρ_1	0.00	0.31	0.31	0.68	0.75	0.29	−0.70	−0.70	0.30	0.76	0.67
ρ_2	0.60	0.73	0.13	0.24	0.27	0.26	0.65	0.05	0.24	0.25	0.41
π_d	0.35	0.48	0.13	0.31	0.34	0.05	0.08	−0.27	0.16	0.32	0.02
μ_1	−2.00	−1.99	0.01	0.04	0.04	0.96	−2.00	0.00	0.04	0.04	0.95
μ_2	−0.50	−0.49	0.01	0.06	0.06	0.95	−0.49	0.01	0.06	0.06	0.93
τ^2	0.06	0.06	0.00	0.03	0.03	0.91	0.06	0.00	0.02	0.02	0.90
π_c	0.65	0.66	0.01	0.03	0.03	0.96	0.65	0.00	0.03	0.03	0.94
θ_0	−0.41	−1.43	−1.02	1.15	1.54	0.34					
θ_1	−0.12	2.40	2.52	2.28	3.40	0.31					
θ_2	0.57	1.87	1.30	1.50	1.98	0.34					
θ_3	0.15	−3.27	−3.42	3.15	4.65	0.36					
γ_0	−1.00						−0.84	0.16	0.22	0.28	0.96
γ_1	1.00						0.91	−0.09	0.15	0.18	0.92
γ_2	−1.00						−0.96	0.04	0.16	0.17	0.92
γ_3	2.00						1.58	−0.42	0.53	0.68	0.94
γ_4	−2.00						−1.75	0.25	0.35	0.43	0.94
γ_5	2.00						1.90	−0.10	0.35	0.37	0.96

Table 2.8: Simulation results from 100 datasets of size, $n = 200$, generated from longitudinal scenario # 4 and the primary probit (a) LC, (b) MSRE models. Left columns: fitted assuming the LC model; right column: fitted assuming the MSRE model.

(a) TRUE: joint LC model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	0.00	0.00	0.12	0.12	0.97	−0.01	−0.01	0.12	0.12	0.98
β_{12}	0.00	0.00	0.00	0.11	0.11	0.98	0.00	0.00	0.11	0.11	0.97
β_{21}	2.83	2.90	0.08	0.09	0.12	0.87	2.90	0.07	0.10	0.12	0.85
β_{22}	2.83	2.67	−0.16	0.09	0.18	0.61	2.67	−0.16	0.09	0.18	0.61
ω^2_{11}	1.00	0.98	−0.02	0.19	0.19	0.94	0.98	−0.02	0.19	0.20	0.92
ω^2_{12}	1.00	0.99	−0.01	0.18	0.18	0.92	0.99	−0.01	0.19	0.19	0.92
ω^2_{21}	1.00	1.00	0.00	0.12	0.12	0.95	1.00	0.00	0.12	0.12	0.94
ω^2_{22}	1.00	1.01	0.01	0.13	0.13	0.95	1.01	0.01	0.13	0.13	0.97
ρ_1	0.00	0.00	0.00	0.13	0.13	0.97	0.00	0.00	0.13	0.13	0.96
ρ_2	−0.60	−0.59	0.01	0.07	0.07	0.92	−0.59	0.01	0.07	0.07	0.94
π_d	0.35	0.36	0.01	0.03	0.04	0.96	0.36	0.01	0.03	0.04	0.96
μ_1	−2.00	−2.00	0.00	0.04	0.04	0.96	−2.00	0.00	0.04	0.04	0.96
μ_2	−0.50	−0.50	0.00	0.06	0.06	0.93	−0.50	0.00	0.07	0.07	0.91
τ^2	0.06	0.06	0.00	0.02	0.02	0.93	0.06	0.00	0.02	0.02	0.91
π_c	0.65	0.64	−0.01	0.04	0.04	0.94	0.64	−0.01	0.04	0.04	0.92
θ_0	−0.80	−0.81	−0.01	0.19	0.19	0.97					
θ_1	1.80	1.80	0.00	0.25	0.25	0.99					
θ_2	−0.20	−0.28	−0.08	0.51	0.52	0.95					
θ_3	−0.30	−0.21	0.09	0.57	0.58	0.95					
γ_0	−0.62						−0.64	−0.01	0.23	0.23	0.98
γ_1	0.29						0.28	−0.01	0.12	0.12	0.97
γ_2	0.29						0.31	0.03	0.13	0.13	0.94
γ_3	−0.36						−0.52	−0.16	0.69	0.71	0.95
γ_4	−0.09						−0.02	0.07	0.35	0.35	0.95
γ_5	−0.08						−0.10	−0.02	0.30	0.30	0.97

(b) TRUE: joint MSRE model

		Assumed LC structure					**Assumed MSRE structure**				
	TRUE	MEAN	BIAS	SD	RMSE	95% COV	MEAN	BIAS	SD	RMSE	95% COV
β_{11}	0.00	0.00	0.00	0.12	0.12	0.96	−0.01	−0.01	0.12	0.12	0.96
β_{12}	0.00	0.00	0.00	0.11	0.11	0.97	0.00	0.00	0.11	0.11	0.97
β_{21}	2.83	2.90	0.07	0.10	0.12	0.84	2.90	0.07	0.10	0.12	0.85
β_{22}	2.83	2.67	−0.16	0.09	0.18	0.59	2.67	−0.16	0.09	0.18	0.60
ω^2_{11}	1.00	0.98	−0.02	0.20	0.20	0.93	0.98	−0.02	0.19	0.20	0.94
ω^2_{12}	1.00	0.99	−0.01	0.19	0.19	0.92	0.99	−0.01	0.19	0.19	0.92
ω^2_{21}	1.00	1.00	0.00	0.13	0.13	0.95	1.00	0.00	0.12	0.12	0.95
ω^2_{22}	1.00	1.01	0.01	0.13	0.14	0.95	1.01	0.01	0.13	0.13	0.97
ρ_1	0.00	0.00	0.00	0.13	0.13	0.98	0.00	0.00	0.13	0.13	0.96
ρ_2	−0.60	−0.59	0.01	0.07	0.07	0.94	−0.59	0.01	0.07	0.07	0.94
π_d	0.35	0.36	0.01	0.03	0.04	0.96	0.36	0.01	0.03	0.04	0.96
μ_1	−2.00	−1.99	0.01	0.04	0.04	0.96	−1.99	0.01	0.04	0.04	0.96
μ_2	−0.50	−0.49	0.01	0.06	0.06	0.94	−0.49	0.01	0.06	0.06	0.94
τ^2	0.06	0.06	0.00	0.02	0.02	0.94	0.07	0.01	0.02	0.02	0.91
π_c	0.65	0.65	0.00	0.04	0.04	0.92	0.65	0.00	0.04	0.04	0.94
θ_0	−0.50	−0.55	−0.05	0.24	0.25	0.91					
θ_1	0.06	0.11	0.05	0.30	0.30	0.90					
θ_2	0.69	0.79	0.10	0.38	0.40	0.94					
θ_3	−0.08	−0.16	−0.09	0.49	0.49	0.95					
γ_0	−1.00						−0.86	0.14	0.25	0.29	0.93
γ_1	1.00						0.92	−0.08	0.15	0.17	0.90
γ_2	−1.00						−0.95	0.05	0.17	0.18	0.93
γ_3	2.00						1.56	−0.44	0.58	0.73	0.93
γ_4	−2.00						−1.78	0.22	0.36	0.43	0.90
γ_5	2.00						1.90	−0.10	0.42	0.43	0.94

CHAPTER III

Modeling Short- and Long-Term Characteristics of Follicle Stimulating Hormone as Predictors of Severe Hot Flashes in Penn Ovarian Aging Study

3.1 Introduction

The Penn Ovarian Aging Study (Manson et al., 2001) is a longitudinal study consisting of a population based sample of 436 women aged 35-47 years selected via random digit dialing in Philadelphia County, PA during 1996-97. The study goal is to explore the associations between reproductive hormone levels and symptoms in the transition to menopause. Follicle stimulating hormone (FSH) is of particular interest because it is known to stimulate folliculogenesis, an important factor in ovarian aging. Thus there has been interest in using longitudinal FSH information to define menopause transition stages as discussed by Sowers et al. (2008). For example, it is well accepted that increasing FSH is an indicator of ovarian aging. However, the FSH levels are not consistently elevated throughout the whole period of late reproductive years. Sowers et al. (2008) found both acceleration and deceleration periods in the gradually increasing FSH levels for a group of women from the late reproductive years. Exploratory analysis of the FSH data in the Penn Ovarian Aging Study shows both acute and gradual increase periods of FSH levels of the population level.

It is therefore of clinical interest to investigate whether elevated FSH levels signal risks of severe menopausal symptoms. Moreover, clinicians also want to quantify the cumulative impact of the FSH histories on the severity of the menopausal symptoms (e.g., hot flashes). To meet these aims, we develop a joint modeling method that accommodates individual contributions of means, time-varying change rates in the long-term trends, and short-term residual variabilities as predictors of health outcomes of interest.

Joint models of longitudinal and health outcome data (mostly time-to-event) have been extensively developed in the literature. The early developments of such joint models were mainly motivated by HIV/AIDS clinical trials and cancer research (Tsiatis et al.,1995; Muthén and Shedden, 1999; Wang and Taylor, 2001; Law et al., 2002; Song et al. 2002; Brown and Ibrahim, 2003a, 2003b; Ibrahim et al., 2004; Yu et al., 2008, among many others). For example, interest has focused on using the "true" underlying longitudinal process (i.e., mean profile) of CD4 or viral load trajectories to relate to the time to progression to AIDS or death. It is therefore natural to model both the observed longitudinal trajectories and disease outcome jointly to gain more efficient inference results, where a longitudinal submodel in the form of a mixed effect model is outlined for the observed longitudinal trajectories and a primary disease outcome model is defined to link the "true" underlying longitudinal process and disease outcome together. Most of these developments have focused on using 1) a summary of important features in the longitudinal profiles, such as the random effects (REs) and the latent classes (LCs); or 2) the last available "true" value (i.e., a function form of REs) as a time dependent covariate, with the earlier values being considered irrelevant to the outcome of interest. These early joint modeling strategies were built assuming that the mean profile was the only feature in the "true" underlying longitudinal process that could predict a disease outcome. The residual variation in the longitudinal profile was treated as a nuisance parameter and as-

sumed not to influence the disease progression. Thorough reviews of the topic are given by Ibrahim et al. (2010) and Rizopoulos (2012). For a comparison of using REs versus LCs, please see Proust-Lima et al. (2012) and Chapter II of this dissertation. Recently, Elliott et al. (2012) and we in Chapter II proposed new methods to link the features in not only the mean profiles but also residual variation in the longitudinal trajectories to predict a disease outcome. However, these existing methods did not consider the potential time varying effects of the dynamic process in longitudinal trajectories.

In this work, we extend these existing approaches by borrowing the idea of relating scalar response and functional predictors in functional data analysis paradigm. Here, we treat the mean longitudinal trajectories as functional predictors linked to the health outcome through a standard functional regression model in the sense of Ramsay and Dalzell (1991) and James (2002) among many others. This modeling strategy implicitly allows the effects of FSH histories (i.e., FSH values up to a particular time point) that are represented by a functional coefficient curve to be time varying and accumulative over time. To the best of our knowledge, such a modeling strategy has not been considered in the joint modeling literature. To estimate the functional coefficient curve, we propose to use a Bayesian penalized spline approach. The advantage is that it also allows for simultaneous evaluation of the uncertainty of the estimated functional coefficient curves by providing pointwise Bayesian credible intervals, which leads to easy identification of critical time windows of increased risk of health outcome of interest, while in standard functional regression, such intervals are typically obtained by bootstrap methods.

For the longitudinal submodel development, the proposed work also extends that of Jiang et al. (2013), who focused on using the latent growth curve features derived from each individual's linear deviation from a population trend as predictors of a health outcome. A key contribution from that paper is to make contrast of pros

and cons of the use of latent classes and multiple shared random effects in joint mixture modeling under correctly or incorrectly specified models. Consequently, the focus differs from this work in terms of exploring model flexibility. Although the posterior predictive checking from Jiang et al. (2013) indicated adequate fit for the majority of the individual FSH trajectories, in the current work we explore more flexible ways of modeling FSH trajectories that can accommodate potentially time varying rates of change, as well as t- rather than just normal distributions to avoid the potential influence of outlying observations. In particular, we consider a robust semiparametric model using Bayesian penalized B splines, which were developed by Lang and Brezger (2004) as a Bayesian version of the penalized splines proposed by Eilers and Marx (1996). In contrast to fully parametric splines, penalized splines are not as sensitive to the exact number and location of the knots as long as enough knots are being used, since "unnecessary" knots will be smoothed away by shrinking random effects toward 0. Recently, the penalized spline approach has gained popularity in smoothing individual curves by associating random effects with a spline basis in mixed effect models under a frequentist framework. For example, Durban et al. (2005) modeled the individual heights of children suffering from acute lymphoblastic leukemia from a clinical trial conducted at Dana Farber Cancer Institute; Chen and Wang (2011) considered modeling longitudinal systolic blood pressure data from Framingham Heart Study. In our work, we explore the use of penalized splines to smooth longitudinal trajectories in the joint modeling framework. Finally, we increase modeling flexibility in two ways: allowing t-distributed errors and using mixtures in mean profiles. These extensions allow us to avoid the potential impact of outlying observations.

In summary, our work brings together advanced statistical ideas including functional data analysis, robust inference and joint longitudinal and outcome modeling in novel ways:

1. It extends conventional functional data analysis to the framework of joint modeling of both the longitudinal (functional predictor) and outcome data, which allows us to select different aspects of the features in the dynamics of longitudinal process as functional predictors. In particular, we focus on the the values or the gradients of the mean trajectories at certain time windows as potential long term functional predictors while also adjusting for the effect of short term variability captured by the variance of the residuals.

2. It uses flexible mixed effects models with B spline basis and heterogeneity (precisely, latent classes) in the first stage submodel, which allows longitudinal trajectories of uneven spacing and unequal length to be used as functional predictors.

3. It allows the effects of FSH histories (values or gradients) to be time varying and to accumulate over time. Statistical tests of the functional coefficient function in the primary outcome submodel can then be used to identify critical time windows where the true association exists. Using a Bayesian approach allows easy calculation of pointwise credible intervals for the functional coefficient functions in comparison to frequentist approaches.

4. It uses a robust model to accommodate outlying observations in the longitudinal data.

3.2 The proposed model

In this section, we present our flexible semiparametric model to use functional regression models for the longitudinal FSH levels to predict the ordinal outcome, severity of hot flashes.

- Longitudinal submodel:

$$
\begin{aligned}
Y_{ij}|\mathbf{b}_i &= \mu_i(t_{ij}) + \epsilon_{ij} \\
\epsilon_{ij} &\sim \mathrm{t}_v(0, \sigma_i^2) \\
D_i &\sim \text{Multinomial}(\pi_{i1}^D, ..., \pi_{iK_D}^D) \\
C_i &\sim \text{Multinomial}(\pi_{i1}^C, ..., \pi_{iK_C}^C) \\
\mu_i(t_{ij}) &= \sum_{l=1}^{L} b_{il}\phi_l(t_{ij}) \\
\mathbf{b}_i = (b_{i1}, ..., b_{iL})'|D_i = d &\sim \mathrm{N}(\boldsymbol{\beta}_d, \Sigma_d) \\
\log\sigma_i^2|C_i = c &\sim \mathrm{N}(\mu_c, \tau^2)
\end{aligned}
\tag{3.1}
$$

where Y_{ij} denotes the observed longitudinal FSH values for subject i, $i = 1, ..., n$ at time t_{ij}, $j = 1, ..., n_i$, $\mu_{ij} \equiv \mu_i(t_{ij})$ denotes the mean of Y_{ij} at time t_{ij} and the vector $\mu_i = (\mu_{i1}, ..., \mu_{in_i})^T$ defines the mean profile or trajectory for subject i. Note that μ_i can be interpreted as the mean long term trend for subject i, so that the vector of the derivatives $\mu_i' = (\mu_{i1}', ..., \mu_{in_i}')^T$ with respect to time t measures the degree of change of the trajectory in a long-term fashion.

To flexibly model the mean profile, we use truncated power splines consists of piecewise polynomials of certain order connected at pre-specified knot locations (Ruppert et al., 2003). Given the same order and knot locations, truncated power spline and B spline are equivalent in the sense that there exist unique one-to-one linear transformations between these two sets of basis functions (Ruppert et al., 2003), leading to the same fitted values from these two splines in the regression setup. However, the B spline is more numerically stable than the truncated power spline because the B spline basis functions are almost orthogonal while the truncated power spline basis functions are not. Therefore, we use B spline basis functions $\phi_l(t_{ij}) \equiv \phi_{l,d}(t_{ij})$, $l = 1, ..., L$ of de-

gree $d = 3$, where $\phi_{l,3}(t_{ij})$ is obtained by the recursion relation:

$$\phi_{l,d}(t_{ij}) = \frac{t_{ij} - \kappa_l}{\kappa_{l+d} - \kappa_l}\phi_{l,d-1}(t_{ij}) + \frac{\kappa_{l+1+d} - t_{ij}}{\kappa_{l+d+1} - \kappa_{l+1}}\phi_{l+1,d-1}(t_{ij})$$

for knots at points $\kappa_1, ..., \kappa_{L-d-1}$, where $\phi_{l,0}(t_{ij}) = I(\kappa_l \leq t_{ij} \leq \kappa_{l+1})$. The number of interior knots is denoted by $J_{\mu(t)}$, such that $\sum_{l=1}^{L} \phi_l(t) = 1$ with $L = J_{\mu(t)} + d + 1$. We defer discussion of selection of knot points to Section 3.2.5. Following Lang and Brezger (2004), we use a Gaussian random walk prior on the fixed effect coefficients $\beta_{d1}, ..., \beta_{dL}$ to penalize large differences among coefficients of the adjacent spline basis and therefore control the smoothness of the mean profile curve to avoid potential overfitting. The specific prior distributions are given in Section 3.2.3. Thus, the vector of fixed effect coefficients $\boldsymbol{\beta}_d = (\beta_{d1}, \cdots, \beta_{dL})^T$ determines the shape and also the smoothness of the mean profile for the d^{th} latent class, defined as $f_d(t_{ij}) = \sum_{l=1}^{L} \beta_{dl}\phi_l(t_{ij})$, $d = 1, \cdots, K_D$. The random coefficients b_{il}, $l, \cdots, L$ capture the individual deviations from the class specific mean profile.

The residual ϵ_{ij} denotes the deviation of Y_{ij} from the subject specific mean at t_{ij} and is assumed to follow a Student's t-distribution with v degrees of freedom, assuming mean 0 and scale σ_i^2. The value of v is assumed to be known. Thus the variance of Y_{ij} is equal to $\frac{v}{v-2}\sigma_i^2$, which can be interpreted as a measurement of short term variability around the mean profile $\mu_i(t_{ij})$. In the case of $v = \infty$, ϵ_{ij} is normally distributed with mean 0 and variance σ_i^2. To allow for over-dispersion of the within-subject scale parameter σ_i^2, we assume a mixture of log normal distributions.

- Outcome submodel is specified conditional on individual longitudinal mean tra-

jectories and variances and has the form:

$$
\begin{aligned}
&W_i \sim \mathrm{N}(\eta_i^W, 1), i = 1, ..., n. \\
&\eta_i^W = \alpha_0 + \mathbf{x}^T \boldsymbol{\lambda}_0 + \int_T \mu_i(t)\theta_0(t)dt, \\
&O_i = 0 \Leftrightarrow W_i < \omega_0, \\
&O_i = s \Leftrightarrow \omega_{s-1} < W_i \leq \gamma_s, s = 1, ..., S-1 \\
&O_i = S \Leftrightarrow W_i > \omega_{S-1}
\end{aligned}
\tag{3.2}
$$

where $\mathbf{x}$ is a vector of baseline covariates with associated (constant) parameter $\boldsymbol{\lambda}_0$, and the functional coefficient function $\theta_0(t)$ represents the effect of subject specific mean trend $\mu_i(t)$ at time t while adjusting for the mean trends at other time points within the time window T. The purpose of considering the integral over the chosen time domain T, i.e., $\int_T \mu_i(t)\theta_0(t)dt$ is to identify critical time windows of elevated outcome risks, which have several advantages over simply summing up over the observed time points t_{ij}, $j = 1, ..., n$. First, longitudinal observations often have missing values and can be measured at different time points (known as unbalanced data) and hence summation over the observed time points becomes problematic. Second, an integral over a chosen time domain implicitly uses the information at infinite time points within time window T while summation only uses the information at finitely observed time points. As in the mean profile trajectories, we let $\theta_0(t) = \sum_{k=1}^{K_0} \tilde{\theta}_{0k}\psi_k^0(t)$ for cubic B-spline basis functions $\psi_k^0(t)$, with $\tilde{\theta}_{0k}$ following a random walk prior, given in Section 3.2.3 to avoid overfitting. Given that we express $\mu_i(t)$ by $\mathbf{b}_i^T\phi(t)$ and $\theta_0(t)$ by $\psi^0(t)^T\tilde{\boldsymbol{\theta}}_0$, thus $\int_T \mu_i(t)\theta_0(t)dt = \int_T \mathbf{b}_i^T\phi(t)\psi^0(t)^T\tilde{\theta}_0 dt = \mathbf{b}_i^T\mathbf{G}_T^0\tilde{\boldsymbol{\theta}}_0$, where $\phi(t)$ is a vector of L basis functions chosen to express $\mu_i(t)$ in the longitudinal submodel and $\phi^0(t)$ is a vector of K_0 basis functions; $\mathbf{G}_T^0 = \int_T \phi(t)\psi^0(t)^T dt$. We can calculate or numerically evaluate $\mathbf{G}_T^0$ for any given spline basis functions

and the estimation of unknown parameters in the outcome primary model becomes fully parametric. For the multinomial probit model, we use the common constraint for the cutoffs $0 = \gamma_1 \leq \cdots \leq \gamma_S = \infty$.

Alternatively, one may postulate that the cumulative changes of the individual trajectories are potentially predictive of the outcome of interest. To accommodate such a possibility, we can consider the first derivative of $\mu_i(t)$, i.e., $\mu_i'(t) = \partial\mu_i(t)/\partial t$ as a functional predictor by taking advantage of the nice properties of B spline of continuity and write our alternative outcome submodel as follows:

$$\begin{aligned}
&W_i \sim \mathrm{N}(\eta_i^W, 1), i = 1, ..., n. \\
&\eta_i^W = \alpha_1 + \mathbf{x}^T \boldsymbol{\lambda}_1 + \int_T \mu_i'(t)\theta_1(t)dt, \\
&O_i = 0 \Leftrightarrow W_i < \gamma_0, \\
&O_i = s \Leftrightarrow \gamma_{s-1} < W_i \leq \gamma_s, s = 1, ..., S-1 \\
&O_i = S \Leftrightarrow W_i > \gamma_{S-1}
\end{aligned} \tag{3.3}$$

where similarly as for $\theta_0(t)$, functional coefficient function $\theta_1(t)$ has the interpretation of the effect of derivative of mean trend $\mu_i'(t)$ at time t while adjusting for the derivatives of mean trends at other time points within the time window T. To emphasize the fact that we can use different spline basis functions to express $\theta_1(t)$, we express $\theta_1(t) = \sum_{k=1}^{K_1} \tilde{\theta}_{1k}\psi_k^1(t)$ using a different set of B spline basis $\psi^1(t) = (\psi_1^1(t), ..., \psi_{K_1}^1(t))'$ and the associated coefficient vector $\tilde{\theta}_1 = (\tilde{\theta}_{11}, ..., \tilde{\theta}_{1K_0})'$. A penalized approach was used by requiring a random walk prior on $\tilde{\boldsymbol{\theta}}_1$, i.e., $\tilde{\theta}_{1l} \sim \mathrm{N}(\tilde{\theta}_{1l-1}, \tau_{\theta_1}^2)$, $k = 2, ..., K_1$. Similarly, we have $\int_T \mu_i'(t)\theta_1(t)dt = \int_T \mathbf{b}_i^T \phi'(t)\psi^1(t)^T \tilde{\boldsymbol{\theta}}_1 dt = \mathbf{b}_i^T \mathbf{G}_T^1 \tilde{\boldsymbol{\theta}}_1$, where $\phi'(t) = \partial\phi(t)/\partial t$ given $\phi(t)$ is a vector of L basis functions chosen to express $\mu_i(t)$ in the longtitudinal submodel and $\phi'(t)$ is a vector of K_1 basis functions; $\mathbf{G}_T^1 = \int_T \phi'(t)\psi^1(t)^T dt$.

3.2.1 Likelihood specification

Let $\boldsymbol{\phi} = (\pi_d^D, \boldsymbol{\beta}_d,\ \Sigma_d,\ d = 1,...,K_D;\ \pi_c^C,\ \mu_c,\ c = 1,...,K_C,\ \tau^2,\ \alpha_0,\ \boldsymbol{\lambda}_0,\ \tilde{\boldsymbol{\theta}}_0, \boldsymbol{\gamma})$, where we assume each parameter in $\boldsymbol{\phi}$ has independent prior distribution with the joint prior distribution denoted by $\pi(\boldsymbol{\phi})$, and $\mathbf{z}$ include all unobserved latent variables, i.e., $\mathbf{z} = (\mathbf{b}, \boldsymbol{\sigma}, \mathbf{C}, \mathbf{D})'$. The observed data $\mathbf{x}$ consists of the longitudinal profiles $\mathbf{y}_1, ..., \mathbf{y}_n$ and the observed outomes $o_1, ..., o_n$. Then the complete data likelihood of $\boldsymbol{\phi}$ based on the complete data $(\mathbf{x}, \mathbf{z})$ is given by

$$
\begin{aligned}
f(\mathbf{x}, \mathbf{z}|\boldsymbol{\phi}) \propto \prod_{i=1}^{n} \Bigg\{ & \prod_d \left[\pi_d (2\pi)^{-\frac{p}{2}} |\Sigma_d|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{b}_i - \boldsymbol{\beta}_d)' \Sigma_d (\mathbf{b}_i - \boldsymbol{\beta}_d) \right\} \right]^{\mathrm{I}(D_i = d)} \\
& \prod_c \left[\pi_c (2\pi\tau^2)^{-\frac{1}{2}} \sigma_i^{-2} \exp\left\{ -\frac{(\log\sigma_i^2 - \mu_c)^2}{2\tau^2} \right\} \right]^{\mathrm{I}(C_i = c)} \prod_{j=1}^{n_i} p\left\{ y_{ij}; \upsilon, \mathbf{b}_i, \sigma_i^2 \right\} \\
& \prod_{s=0}^{S} \left\{ \Phi(\gamma_s - \eta_i^W) - \Phi(\gamma_{s-1} - \eta_i^W) \right\}^{\mathrm{I}(o_i = s)} \Bigg\} \pi(\boldsymbol{\phi})
\end{aligned}
\tag{3.4}
$$

where

$$
p\left\{ y_{ij}; \upsilon, \mathbf{b}_i, \sigma_i^2 \right\} = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{\left(y_{ij} - \sum_{l=1}^{L} b_{il} \phi_l(t_{ij}) \right)^2}{2\sigma_i^2} \right\} & \text{if } \upsilon = \infty; \\ \frac{\Gamma(\frac{\upsilon+1}{2})}{\Gamma(\frac{\upsilon}{2}) \sqrt{\pi\upsilon\sigma_i^2}} \left\{ 1 + \frac{1}{\upsilon} \frac{\left(y_{ij} - \sum_{l=1}^{L} b_{il} \phi_l(t_{ij}) \right)^2}{2\sigma_i^2} \right\}^{-\frac{\upsilon+1}{2}} & \text{if } \upsilon < \infty. \end{cases}
$$

3.2.2 Data augmentation step to impute missing data

Given the minimum number of available repeatedly measured FSH levels in our final sample (ranging between 6 and 26 per woman), we are limited to the number of knots when choosing cubic B spline basis functions to express $\mu_i(t_{ij})$. To maximize the number of knots we can consider, we fill in those with fewer than 26 observations based on data augmentation within each iteration of Gibbs sampling (Chapter 10 in Little and Rubin, 2002). When assuming missing completely at random (MCAR) missing data mechanism, this data augmentation procedure proceeds as follows,

- draw $\mathbf{Y}_{\text{mis}}^{(t+1)}$ from $p(\mathbf{Y}_{\text{mis}} \mid \Phi)$
- draw $\Phi^{(t+1)}$ from $p(\Phi \mid \mathbf{X}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$

where Φ denotes all unknown variables, $\mathbf{Y}_{\text{mis}}$ denotes the missing longitudinal observations of FSH levels, and $\mathbf{X}_{\text{obs}}$ denotes all observed data including observed longitudinal observations and primary outcome of interest. The above simulation leads to draws from the joint distribution of $(\Phi, \mathbf{Y}_{\text{mis}})$ given observed data $\mathbf{X}_{\text{obs}}$. Therefore, this procedure leads to the same inference about Φ as when we only focus on the marginal distribution of Φ given observed data $\mathbf{X}_{\text{obs}}$. However, this trick allows us to put in more knots to fully take advantage of the penalized spline approach that is free from knot location selection given enough number of knots.

3.2.3 Prior specification

We propose a fully Bayesian approach to estimate model parameters. For the mixture normal distribution of the random effects, we assume a first-order Gaussian random walk prior as proposed by as Lang and Brezger (2004): $\beta_{dl} \sim \text{N}(\beta_{d,l-1}, \tau^2_{\beta d})$, $l = 2, ..., L$ with diffuse prior $\beta_{d1} \sim \text{N}(0, 100)$ for the initial coefficient, and $\tau^2_{\beta d} \sim \text{IG}(1, 0.005)$ to control the smoothness of the fitted curves. We do not impose restrictions on the structure of the variance-covariance matrix for the random effects Σ_d. To avoid problems with unbounded likelihoods in normal mixture models with unstructured variance-covariance matrices (Day 1969), we use the prior proposed by Kass and Natarajan (2006): $\Sigma_d \sim$ Inverse-Wishart$(\text{df} = r, \Lambda)$, where $\Lambda = r\left(\sum_{i=1}^{n} \widehat{\text{Cov}}(\tilde{\mathbf{b}}_i)^{-1}/n\right)^{-1}$, $\tilde{\mathbf{b}}_i$ is given by OLS estimator of b_i for subject i and r is the dimension of b_i.

For the mixture log normal distribution for the residual variances, we used diffuse priors: $\mu_c \sim \text{N}(0, v)$, $\tau^2 \sim \text{IG}(a, b)$ with $v = 1000$ and $a = b = .001$. For the class membership probabilities, we assume conjugate Dirichlet$(4, ..., 4)$ on both $\pi^C = (\pi^C_1, ..., \pi^C_{K_C})$ and $\pi^D = (\pi^D_1, ..., \pi^D_{K_D})$ (Frühwirth-Schnatter 2006); this is equivalent to assuming *a priori* 4 observations in each class, avoiding the existence of empty classes.

Lastly, in the probit submodel we assign independent priors N(0,9/4) for the α_0 and every element of $\boldsymbol{\lambda}_0$; for the coefficients associated with functional coefficient function $\theta_0(t)$, $\tilde{\boldsymbol{\theta}}_0 = (\tilde{\theta}_{01},...,\tilde{\theta}_{0K_0})'$, similarly we use a first-order Gaussian random walk prior, i.e., $\tilde{\theta}_{0k} \sim \text{N}(\tilde{\theta}_{0k-1}, \tau^2_{\theta_0})$, $k = 2,...,K_0$ with $\tilde{\theta}_{01} \sim \text{N}(0,9/4)$ and $\tau^2_{\theta_0} \sim \text{IG}(1,0.005)$, where the prior variance 9/4 is chosen to bound the probabilities of $o_i = 1$ to be away from 0 and 1 (Garrett and Zeger, 2000; Elliott et al., 2007 and Neelon et al., 2011). We put flat uniform priors on γ_s for $s \notin \{0,S\}$, that is, $\gamma_s \sim \text{Uniform}(-\infty,\infty)$.

3.2.4 Posterior computation

Gibbs sampling is used to obtain draws from the corresponding posterior distributions. For $(\alpha_0, \lambda_0, \tilde{\boldsymbol{\theta}} \mid \mathbf{b}, \boldsymbol{\sigma}, \mathbf{o})$ we use the Albert and Chib (1993) data augmentation method for probit regression models. The draws of $(\sigma_i^2 | C_i, \mu_c, \gamma^2, \mathbf{b}_i, o_i, W_i, \{y_{ij}\}_j)$ for $i = 1,...,n$ are obtained by the inverse cumulative distribution method. The exact specification of all priors and MCMC sampling procedures are provided in Appendix B.

Three chains from diverse starting points were run and Gelman-Rubin statistic $\sqrt{\hat{R}}$ (Gelman et al., 2003) (square root of total variance to within-chain variance ratio) were used to assess the convergence of the MCMC chains. For the population level parameters, the maximum $\sqrt{\hat{R}} = 1.030$ for models assuming less than 3 classes; and when assuming 3 classes for either mean profile or the variance class, the maximum $\sqrt{\hat{R}} = 1.184$. For the well-documented issue of "label switching" in finite mixture modeling (Redner and Walker 1984), we applied the post-processing relabeling algorithm (Stephens, 2000) in which all possible permutations of class assignments were considered at each iteration of the Gibbs sampler. The permutation of the class assignment was chosen to maximize the posterior probability so that the labeling of classes was consistent with the previous assignments. We post-process our chains using Stephen's algorithm to untangle the draws for model parameters.

All the calculations were performed by calling stand alone C++ codes in R, developed using an open source C++ library for statistical computation, the Scythe statistical library (Pemstein et al., 2007), which is available for free download at http://scythe.wustl.edu.

3.2.5 The choice of the number of classes and number of knots in penalized splines

We consider the deviance information criterion (DIC), proposed by Spiegelhalter et al. (2002) to select both the number of components for the latent classes and to choose the number of knots in the penalized splines. DIC uses the discrepancy between the posterior mean of the deviance $\overline{D(\boldsymbol{\phi})} = E_{\boldsymbol{\phi}}\left\{-2\log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\}$ and the deviance evaluated at the posterior mean $D(\overline{\boldsymbol{\phi}}) = -2\log f\left\{\mathbf{x} \mid E(\boldsymbol{\phi} \mid \mathbf{x})\right\}$ to estimate the effective number of degrees of freedom in the model p_D. DIC is then given by the analog of the Akaike Information Criterion (AIC):

$$\mathrm{DIC}(\mathbf{x}) = \overline{D(\boldsymbol{\phi})} + p_D = 2\overline{D(\boldsymbol{\phi})} - D(\overline{\boldsymbol{\phi}}) = -4\mathrm{E}_{\boldsymbol{\phi}}\left\{\log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2\log f\left\{\mathbf{x} \mid \mathrm{E}(\boldsymbol{\phi} \mid \mathbf{x})\right\}.$$

In our setting, $f(\mathbf{x} \mid \boldsymbol{\phi})$ where $\mathbf{x} = (\mathbf{y}_{\mathrm{obs}}, \mathbf{o})'$ consisting of the fully-observed data is not available in closed form; instead we use the approach outlined in Celeux et al. (2006) to obtain

$$\mathrm{DIC}(\mathbf{x}) = \mathrm{E}_{\mathbf{z}}\{\mathrm{DIC}(\mathbf{x}, \mathbf{z})\} = -4\mathrm{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\log f(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2\mathrm{E}_{\mathbf{z}}\left[\log f\left\{\mathbf{x}, \mathbf{z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z})\right\} \mid \mathbf{x}\right]$$

where integration over the latent variables $\mathbf{z} = (\mathbf{b}, \boldsymbol{\sigma}, \mathbf{C}, \mathbf{D}, \mathbf{y}_{\mathrm{mis}})'$ is obtained via numerical methods.

3.2.6 Goodness of fit evaluation

We assessed the model goodness of fit to the data in two ways: pivotal discrepancy measures (PDMs) (Yuan and Johnson 2012), which yield an overall goodness-of-fit measure for both the longitudinal predictor component and the ordinal outcome component of the joint model, and area under the receiver-operator characteristic (ROC) curve (AUC), a goodness-of-fit measure focusing on prediction of the ordinal outcome of interest.

PDMs generalize the pivotal quantities proposed by Johnson (2007); in contrast to more general posterior predictive distribution measure of fit (Gelman et al., 1996), they are defined to depend only on the data and the model parameters with a known distribution. If the model is correctly specified, the PDMs evaluated at the true parameter value and the draws from the posterior distribution should have the same sampling distribution. Therefore, model adequacy can be tested by treating the PDMs as a test statistic to obtain a uniformly distributed p value. However, the posterior samples of PDMs are not independent as they are all derived from the observed data (Johnson, 2004), thus p-value calculation is difficult. Instead, Yuan and Johnson (2012) focused on the upper bound of p values and hence the upper bound of a p value being less than 0.05 provided strong evidence of model inadequacy.

For the longitudinal trajectories, we define subject level PDMs to examine the fit of longitudinal trajectories, which are constructed in the following steps:

1. For each subject i, define the standardized residuals $z_{ij} = [y_{ij} - \mu_i(t_{ij})]/\sigma_i$, $j = 1,...,n_i$.

2. Partition z_{ij} into K groups according to the values of $\mu_i(t_{ij})$. We choose $K = 2$ given that the repeated measurements per subject ranges from 6 to 26.

3. If the assumed longitudinal submodel is correct, then $r_{ij} = P^{-1}\left(y_{ij}; v, \mu_i(t_{ij}), \sigma_i^2\right)$ is $[0,1]$ uniformly distributed, where $P(\cdot)$ is the corresponding cumulative dis-

tribution (CDF) function for y_{ij}. Let n_{ik} denote the total number of residuals in group k for subject i based on the partition in previous step, $k = 1,...,K$, and then paritition the interval $[0,1]$ into L sub-intervals of length p_l; we would then expect $n_{ik}p_l$ CDF transformed r_{ij}'s would fall into each of the L bins. We consider $L = 3$ intervals: $(0,\frac{1}{3}]$, $(\frac{1}{3},\frac{2}{3}]$ and $(\frac{2}{3},1]$ and hence $p_1 = p_2 = p_3 = \frac{1}{3}$. Let $\tilde{o}_{ikl}$ denote the observed r_{ij}, $j = 1,...,n_i$ in sub-interval l, $l = 1,...,L$ within group k, then $d_{ik}(\mathbf{y_i},\phi) = \sum_{l=1}^{L}\left(\frac{\tilde{o}_{ikl} - n_{ik}p_l}{\sqrt{n_{ik}p_l}}\right)^2$ is χ^2 distributed with degrees of freedom $L-1$ given that the assumed model is correct.

4. Sum the defined χ^2 statistic for each group k, we define our PDM for each subject i as $\sum_{k=1}^{K} d_{ik}(\mathbf{y},\phi)$, which is χ^2 distributed with degrees of freedom $K(L-1)$ when the assumed model is correct for subject i.

For the ordinal outcome of interest, our PDM is defined based on realized latent continuous residuals as considered by Gelman et al. (2000). That is, we define the standardized latent residuals $e_i = W_i - \eta_i^W$ with $\eta_i^W = \alpha_0 - \mathbf{x}_i^T\lambda_0 - \int_T \mu_i(t)\theta_0(t)dt$ and then partition e_i into $\tilde{K}$ groups according to the values of η_i^W, where we let $\tilde{K} = 10$ given our sample size is 235. Next, for $\Phi^{-1}(e_i)$, $i = 1,...,n$ we calculated PDM for outcome data by following steps similar to 2 and 3 above for individual trajectories.

Finally, we use repeated posterior draws to obtain the sampling distribution of PDMs and compute the upper bounds of the p values based on ordered statistics of PDMs using the equation given in Yuan and Johnson (2012). If the upper bound of p value is less than 0.05, it provided strong evidence of inadequate fits.

Second, we assessed the prediction of the outcome using receiver-operator characteristic (ROC) curves, in particular the area under the ROC curve (AUC). ROC curves plot true positive rate (TP) versus false positive rate (FP) for all possible cut-offs based on predicted $\mathrm{P}(o_i = s) = \Phi(\mathbf{Z}_i'\eta)$ obtained from (3.2) for $s = 0,...,S$. The ROC curve and AUC were computed at each MCMC iteration using the ROCR package in R (Sing et al. 2005). The ROC is computed by ordering the observations $(i) = 1,...,n$ so

that $\hat{P}(o_{(i)} = 1) \geq \hat{P}(o_{(i+1)} = 1)$, computing changepoints $c = 2, \ldots, n_c$, $n_c \leq n$ where the observations change from positive to negative (i.e., $o_{(c-1)} = 1$, $o_{(c)} = 0$), and plotting $\sum_{(i)=1}^{c}(1 - o_{(i)})/\sum_{(i)=1}^{n}(1 - o_{(i)})$ on the x-axis versus $\sum_{(i)=1}^{c} o_{(i)}/\sum_{(i)=1}^{n} o_{(i)}$ on the y-axis. Area under the ROC is then computed using a trapezoidal approximation. The posterior mean AUC is calculated as the average AUC's across MCMC iterations. To obtain the posterior mean and the pointwise 95% credible interval of ROC curve, we choose 250 points equally spaced along the FP axis and take the vertical average or 95% quantiles of TP's at the 250 chosen points. This approach is referred to as vertical averaging of ROC curves at fixed FP rates by Fawcett (2006).

3.3 Predicting severity of hot flash from longitudinal follicle stimulating hormone data

In the Penn Ovarian Aging Study, participating women had their hormone measures taken annually during the early follicular phase of a menstrual cycle for 2 sequential menstrual cycles, with up to 13 years of follow-up available at the time of our analysis. Hormone values were censored if a woman was pregnant, breast feeding or taking exogenous hormones during the follow-up. This led to the number of FSH levels available for each woman ranging from 1 to 26 out of the maximum 26. We focus our analysis on the 234 women who 1) had not experienced hot flash symptoms at baseline, 2) had baseline measurements of BMI and smoking status (0 or 1) that are to be included as baseline covariates in the outcome submodel, and 3) had at least 6 measurements of FSH levels. We let y_{ij} denote the natural log transformed FSH levels i.e., log(FSH) and o_i denote the ordinal outcome of interest, severity of hot flashes (0, 1, and 2), defined as $o_i = 0$ if never had severe hot flashes (that is, severity score < 2 throughout the follow up period); $o_i = 1$ if had severe but not more severe hot flashes (that is, severity score at least once =2 or once =3 that occurred

before 40 yrs. old) and $o_i = 2$ if had more severe hot flash (that is, severity score at least once =3 after 40 yrs old). In our final sample, 117 (50%) never experienced any severe hot flashes during follow-up (severity score=0), 80 (34%) had a severity score of 1, and 37 (16%) had a severity score of 2. The average number of available FSH levels per woman is 18.7 (range: 6-26) in our final sample. Since it is commonly believed that most women start to experience menopausal related symptoms between the age of 45 and 50 and reach menopause by the age 55, we consider $T = [45, 55]$ as a potential risk time window in our analysis for the impact of changes in FSH levels on risk of severe hot flashes.

We use model 3.1 to describe longitudinal measured FSH and model 3.2 to relate long- and short-term FSH characteristics to the occurrence of severe hot flashes. Preliminary analysis suggested using cubic B spline basis functions with 1 to 3 inner knots to express $\mu_i(t_{ij})$ and cubic B spline basis functions with 1 to 5 inner knots to express the functional coefficient function $\theta_0(t)$. Thus we consider models with 1, 3 or 5 knots; putting these knots at the equally spaced quantiles of the distinctly observed ages of these women as recommended by Ruppert et al. (2003). This is equivalent to assuming piecewise cubic orthogonal polynomials connected at those chosen knot locations. Next, we consider the number of components for both mean profile and variance classes. Previous analysis of fitting mixture distributions for both the random effects and variances (Chapter II) successfully identified 1 mean profile class and 2 variance classes under normality assumption for ϵ_{ij}. However, our current approach assumes a t distribution for ϵ_{ij} that will potentially impact the effect of any outliers on estimation of the mean trajectories, which may alter the optimal numbers of components for the mean profile and variance classes. With all these considerations, we consider $K_D = 1, 2$ and 3 and $K_C = 1, 2$ and 3 in our analysis. To deal with outlying observations that cannot be fitted well by normal model, Jeffreys (1973) (page 65) suggested using t distribution with degrees of freedom within the

range of $[4, 15]$.

We attempted to estimate the degrees of freedom by treating it as a true parameter in our model, but found its estimation unstable without use of an informative prior. Hence we perform a sensitivity analysis, comparing results from a normal model (df=∞) with a submodel with df=4 and 7, respectively. We chose these three scenarios as representative settings to reflect the assumptions of presence of extreme outliers, mild outliers or absence of outliers relative to a normal distribution in the FSH data.

Table 3.1 presents the DIC statistics for all models considered: 1,2, or 3 latent classes for the mean profiles and variances; normal, t (df=7) and t (df=4) assumptions for the errors in longitudinal submodel; and 1,3 or 5 knots for the longitudinal trajectories or functional varying coefficient function respectively. In general, DIC suggests that joint models with t (df=4) assumption for the longitudinal submodel fits the data much better than t (df=7), and much better than the normal model. Under this t assumption, $K_D = K_C = 2$ is selected for both df=4 and df=7. Given these selected number of components for both the mean profile and variance classes for each model, DIC further suggests that 1 knot (i.e., $J_{\mu(t)} = 1$) at 46.6 years of age for the longitudinal trajectories and 3 knots (i.e., $J_{\theta_0(t)} = 3$) at 41.6, 46.6 and 51.5 years of age for the functional varying coefficient function offers the best balance between goodness of fit and smoothness under all these three longitudinal submodel assumptions. Next, we will only focus on these best fitting models summarized as:

- best fitting normal model: $K_D = 1, K_C = 2$ with $J_{\mu(t)} = 1$ at 46.6 years of age and $J_{\theta_0(t)} = 3$ at 41.6, 46.6 and 51.5 years of age
- best fitting t (df=7) model: $K_D = K_C = 2$ with $J_{\mu(t)} = 1$ at 46.6 years of age and $J_{\theta_0(t)} = 3$ at 41.6, 46.6 and 51.5 years of age
- best fitting t (df=4) model: $K_D = K_C = 2$ with $J_{\mu(t)} = 1$ at 46.6 years of age and

$J_{\theta_0(t)} = 3$ at 41.6, 46.6 and 51.5 years of age

For these best fitting models, PDMs also confirmed our previous finding based on model selection criterion DIC that the t (df=4) model fits the longitudinal FSH trajectories better than the t (df=7) and normal distribution. Figure 3.1 shows the upper bounds of the p values based on PDMs for longitudinal trajectories fitted by all three final models. If the upper bound of a p value is less than 0.05, there is strong evidence of inadequate fit. We see that the normal model fits the large majority of subjects well, with 40 individual trajectories being considered to have inadequate fit by PDMs. Out of these 40 individual trajectories, assuming t distribution with 7 degree of freedom improved the fits of 27 individual trajectories, leaving still 13 individual trajectories with inadequate fit; among the 13 individual trajectories, further assumption of even more heavy tailed t distribution with 4 degrees of freedom still resulted in 8 individual trajectories to have inadequate fit. Further examination of these 8 (3.4% of total 234 trajectories) individual level observed trajectories indicates the existence of varying degrees of outlying observations. For example, the first row in Figure 3.2 shows three trajectories that are considered to have inadequate fits by all three best fitting models based on PDMs (i.e., upper bound of p values are less than 0.05). The second row in Figure 3.2 shows 3 trajectories that have upper bounds of p values less than 0.05 by our best fitting normal model but upper bounds of p values greater than 0.05 by both our best fitting t (df=7) and t (df=4) models, while the third row shows 3 trajectories that have upper bounds of p values less than 0.05 by our best fitting normal and t (df=7) models but upper bounds of p values greater than 0.05 by both our best fitting t (df=4) models. Clearly, these plots suggest that t models with 4 and 7 degrees of freedom show considerably less influence by outlying observations than the normal model and they both have almost identical fits visually. Finally, the last row of Figure 3.2 shows three trajectories that have upper bounds of p values greater than 0.05 by all three of our best fitting models:

the normal and t models with df=4 and df=7 show very similar fits. Therefore, the inadequate fit of longitudinal FSH trajectories identified by PDMs is likely due to these varying degrees of extreme outliers. Although we could consider even smaller degrees of freedom of t distribution or more heavily tailed distribution for the longitudinal submodel to accommodate these extreme outlying observations, the t model with either 4 or 7 degrees of freedom already shows almost identical robustness to them and seems to provide reasonably good fit to the FSH data.

Next, we contrast the estimation results from these models to demonstrate the influence of not appropriately accommodating outlying observations. Figure 3.3 presents the mean profile components and two variance components identified by the three best fitting models. Consistent with the finding reported in Chapter II, under the normal model assumption, a single-component mean profile is favored by DIC. In contrast, under both the t (df=7) and t (df=4) model assumptions, a two-component mean profile is favored by DIC: the major mean class (86% of women) whose FSH levels begin increasing in their late 40s and the minor mean class (14% of women) with increasing FSH levels starting around age 40 capturing a proportion of women who might transition into menopause at an earlier age. The variance class has different meanings under t and normal assumptions but in both scenarios measure the short term variations in FSH levels: according to their magnitudes, both t and normal models would classify them to either "low" and "high" classes. Based on the posterior estimates of these component-specific parameters given in Table 3.2, we can further see more subtle differences in these estimated mixture components under varying assumptions.

In general, under all these three model assumptions, we have reached the same broad conclusions: the high short term variability in the FSH levels is strongly associated with increased risks of more severe hot flash; smoking is marginally associated and there was no association with BMI or the individual mean trajectories between

age 45 and 55. The most dramatic difference between the different df models occurs for the estimated functional coefficient $\theta_0(t)$. Figure 3.4 (a), (b) and (c) show the estimated functional coefficient $\theta_0(t)$ by our best fitting normal, t (df=7) and t (df=4) models, respectively. The estimated $\theta_0(t)$ under our best fitting normal model tends to have larger effect size (larger magnitude in $\theta_0(t)$) before age 53 and an overall wider pointwise 95% credible interval than the estimated $\theta_0(t)$'s under our best fitting t (df=4) and t (df=7) models. All three coefficient curves suggest that, when adjusting for the whole history of mean FSH levels over the age range of age 45 to age 55, higher mean FSH levels before age 53 reduce risk of severe hot flashes, while higher mean FSH levels between age 53 and age 55 increase this risk, but there is no conclusive evidence of a true association between the FSH trajectory histories and the risk of more severe hot flash.

Finally, to consider the effect of the derivative of the mean profile $\mu_i'(t)$, we focus on the best-fitting t (df=4) model, for which the lower bound of p value based on PDM, is 1.0, indicating a good fit to the risks of different severities of hot flash. Figure 3.5(a) considers the effect of cumulative changes in the mean profiles across the age range $\tilde{T} = [45, 55]$, while Figure 3.5(b) considers the equivalent effect across the age range $\tilde{T} = [50, 55]$, potentially a more clinically relevant age range since the median age of menopause is 51 and therefore the hormone dynamics in this time window are more likely to play a role in the menopausal related symptoms. When fit over the wider age range, higher values of $\mu_i'(t)$ decrease risk slightly before age 50 and increase it over age 50, although the 95% credible intervals include 0 by a wide margin. In contrast, a more narrowly-focused age range of $\tilde{T} = [50, 55]$ suggested significantly increased risk of severe hot flash associated with higher values of $\mu_i'(t)$ in the age range of 52.5-55, with $\hat{\theta}_1(52.5) = 0.408$ (95% CI=0.019, 0.843) and $\hat{\theta}_1(55) = 0.514$ (95% CI=0.003, 1.290).

Figure 3.6 shows the receiver-operator characteristic (ROC) curves for the best-

fitting t (df=4) model, comparing the use of the $\mu_i(t)$ and $\mu_i'(t)$ between age 45 and 55 to discriminate each of the hot flash severities (0, 1 and 2), along with the other predictors (residual variance, BMI, and smoking status). These ROCs and their associated area under the curve (AUCs) suggest that using either functional predictors led to moderately accurate classifications of different hot flash severities. Visually, there is not much differences in these ROC curves; a further comparison of AUCs also suggests that the predictive performances by using $\mu_i(t)$ and $\mu_i'(t)$ have negligible differences (δ AUCs for severity 0, 1, and 2 are -0.012 (-0.097, 0.070), -0.002 (-0.073, 0.071) and -0.020 (-0.131, 0.091) respectively).

3.4 Conclusions and Discussion

In this paper we develop a novel joint modeling approach to answer the scientifically important research question of how long-term history of FSH values or its change rates affects the severity of hot flash, a symptom almost every woman experiences during the menopausal transition. While many joint models have been developed in the context of cancer research and HIV/AIDS clinical trials in the past decade, most methods focus on the features in the "true" underlying longitudinal process (i.e., mean profile) that take the forms of random effects or latent classes; or alternatively the last available "true" underlying value as a time-dependent covariate. Following Elliott et al. (2012) and Chapter II, we seek the useful longitudinal features in both the mean profiles and the short-term variability. Further we allow the mean of the longitudinal process and the corresponding derivatives to be time varying, and their effects toward the responses to be accumulative over time. To summarize, we propose a broadly applicable joint modeling approach that

1. makes efficient use of the available information in the longitudinal data, by including the short-term and long-term dynamic feature in the mean histories

as functional predictors in the second stage primary outcome model while also adjusting for short-term variability,

2. allows selection of longitudinal features within certain clinically relevant time windows to predict the health outcome of interest in the second stage primary outcome submodel, where the effects outside this particular time window are assumed to be negligible,

3. can handle balanced or unbalanced longitudinal data that are densely or sparsely measured.

To realize these modeling goals, we use a penalized spline approach to allow the flexible modeling of longitudinal features and the functional coefficient curve representing the time varying effect of longitudinal feature, although in principle they can also be fit by a more restrictive parametric approach. Since the ultimate goal is to simultaneously model both the mean trajectories and the residual variability but distinguish between their effects in the outcome submodel, we choose a t distribution to properly model residual variability to avoid the impact of outlying FSH values. In particular, we demonstrate the importance of assuming this robust distribution assumption instead of the typical normal assumption used in most of the joint modeling literature. However, due to the limited number of longitudinal observations for some women (i.e., ranging from 6 to 26), there is insufficient information in the data to assume individually varying degrees of freedom in t distribution, we are limited to assume a global degrees of freedom common to all trajectories. In addition, our attempts to use the data to estimate even the global degrees of freedom parameter using the informative exponential distribution proposed by Geweke (1993), the truncated uniform prior on the inverse of the degree of freedom suggested in Lange et al. (1989) and Gelman and Hill (2007) and the Jefferys prior derived by Fonseca et al. (2008) all failed: the estimated global degrees of freedom were always close to an

prior cutoff value, implying the existence of extreme outliers in the FSH data that tend to drive the degrees of freedom in t distribution to low values. Given that the fitted values are only modestly affected by different values of degree of freedom in t distribution (Lange et al. (1989)), we chose to fix the degrees of freedom parameter at a small number of fixed values and conduct a sensitivity analysis using DIC to choose among the models.

The proposed model also allows latent heterogeneities in both the individual level mean trajectories and the residual variability as in Chapter II. Under our best fitting t (df=4) model, as shown in Figure 3.3 (e), the mean FSH trajectories can be separated into two classes, one minor class with 14% of trajectories and the other major one with 86% of trajectories. Both classes are reflective of three typical FSH change patterns for women in the transition to menopause (Burger et al., 1999) in that FSH is relatively flat prior to the menopause transition, has an increasing period during the menopause transition, and will eventually plateau once women get about 2 years post menopause; but women in the minor class tend to have an earlier increasing time frame in the FSH trajectories along with higher FSH values than the women in the major class. As shown in Figure 3.7, the fitted mean FSH curves for the total 28 women assigned to the minor class and a random sample of 20 women assigned to the major class based on the posterior mode were plotted. This once again shows the heterogeneous nature in the mean FSH trajectories that is supported by our model selection criterion DIC and implies that the women in the minor class tend to reach menopause at a much earlier age. Also, as shown in Figure 3.3, even with the use of the t distribution to account for extreme outlying observations, it seemed that there still exists a true mixture in residual variability.

Generally, the functional coefficient curves $\theta_0(t)$ and $\theta_1(t)$ can be fit by any spline basis with or without penalty. In particular, if the shape of $\theta_0(t)$ or $\theta_1(t)$ is known, for example, $\theta_0(t)$ is a linear function, then we can let $\psi^0(t) = (1, t)$ and assume a regular

flat normal prior on the coefficients associated with basis function 1 and t. When the true shape of $\theta_0(t)$ or $\theta_1(t)$ is unknown, we recommend starting the analysis using a more flexible penalized approach to get some idea of the shape of $\theta_0(t)$ or $\theta_1(t)$, which may be further reduced to simple parametric form to stabilize estimation of model parameters and reduce the length of pointwise credible or confidence intervals for $\theta_0(t)$ or $\theta_1(t)$.

Finally, we feel that our proposed model lays out an example of efficiently using information in the longitudinal data to predict a primary outcome by borrowing various existing tools from the literature. Based on this framework, there are several directions for future work. Most generally, the longitudinal studies often measure several variables repeatedly, for example, in our case several other hormone profiles are also available. Developing methods to model these potentially correlated longitudinal trajectories simultaneously while also using this information effectively to predict or relate to the outcome of interest are areas for future research.

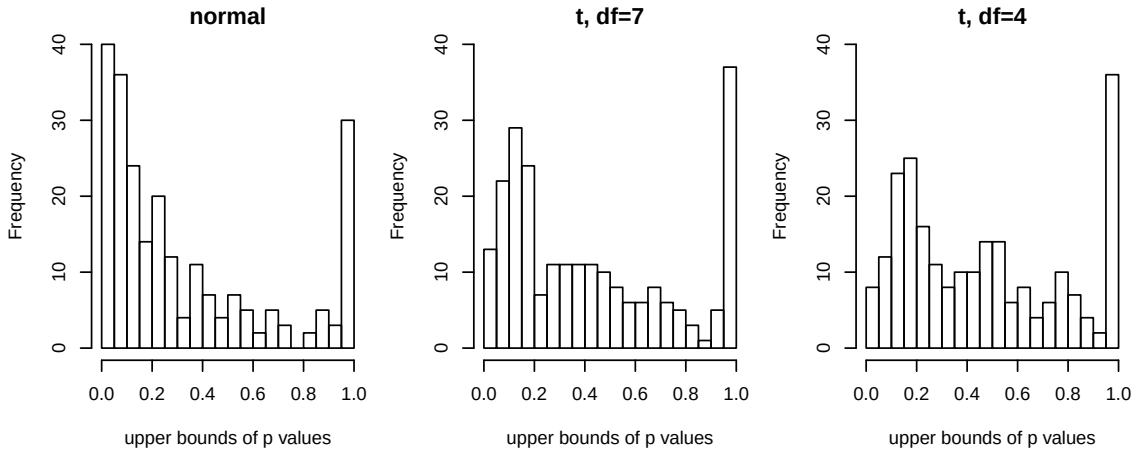

Figure 3.1: Upper bounds of p values based on PDMs for individual trajectories fit by our best fitting models with $\mu_i(t)$ $i = 1,...,n$ within the time window $T = [45,55]$ as a functional predictor in primary outcome submodel.

Table 3.1: DIC from different joint models for the analysis of Penn ovarian aging data, assuming normal, t (df=7) and t (df=4) distribution for the longitudinal submodel and using $\mu_i(t)$ $i = 1,...,n$ within the time window $T = [45,55]$ as a functional predictor in primary outcome submodel.

	$K_C=1$			$K_C=2$			$K_C=3$		
Model	$K_D=1$	$K_D=2$	$K_D=3$	$K_D=1$	$K_D=2$	$K_D=3$	$K_D=1$	$K_D=2$	$K_D=3$
Normal									
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=1$	11439.0	11477.2	11492.9	**11333.6**	11369.1	11399.3	11511.6	11545.1	11560.7
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=3$	11437.5	11487.9	11501.8	**11327.7**	11364.9	11386.7	11506.8	11542.9	11561.5
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=5$	11435.0	11480.5	11493.3	**11330.6**	11369.1	11385.7	11500.7	11552.2	11574.9
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=1$	11923.4	11912.4	11924.6	11809.6	11788.7	11798.9	12000.1	11977.5	11984.4
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=3$	11923.8	11901.3	11915.5	11807.0	11803.5	11799.8	11995.0	11971.6	11997.1
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=5$	11924.7	11892.4	11919.2	11799.7	11788.2	11801.4	11993.1	11965.6	11991.5
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=1$	12419.3	12400.5	12418.6	12319.9	12308.2	12316.5	12506.2	12489.0	12499.3
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=3$	12421.8	12398.8	12412.5	12317.6	12306.7	12320.6	12506.5	12486.7	12489.2
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=5$	12416.6	12399.3	12409.5	12317.0	12298.1	12307.5	12504.7	12472.7	12485.0
t, df=4									
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=1$	10335.0	10257.5	10271.0	10303.3	**10215.4**	10246.8	10425.0	10326.3	10347.2
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=3$	10333.2	10255.7	10272.5	10308.8	**10210.8**	10235.5	10419.9	10330.3	10374.1
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=5$	10331.2	10260.0	10273.9	10298.5	**10230.4**	10228.3	10432.3	10322.7	10371.9
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=1$	10831.8	10823.6	10826.4	10803.1	10774.6	10778.2	10947.6	10906.7	10889.1
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=3$	10830.0	10821.0	10833.2	10821.3	10776.0	10812.1	10929.6	10897.9	10934.2
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=5$	10828.0	10818.8	10822.3	10818.0	10780.1	10791.6	10936.8	10914.8	10922.0
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=1$	11280.6	11259.2	11256.8	11287.8	11255.8	11257.5	11406.5	11369.9	11397.4
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=3$	11275.4	11251.5	11256.8	11276.3	11251.4	11271.0	11393.9	11356.3	11382.0
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=5$	11278.3	11250.5	11265.0	11298.1	11253.6	11264.5	11409.9	11381.4	11384.1
t, df=7									
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=1$	10626.5	10585.0	10606.3	10566.9	**10518.2**	10533.3	10679.8	10603.3	10652.2
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=3$	10624.0	10584.2	10600.6	10567.8	**10511.5**	10532.0	10694.9	10633.9	10648.5
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=5$	10622.5	10579.8	10598.3	10558.1	**10512.0**	10536.6	10670.4	10615.5	10628.7
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=1$	11127.3	11114.8	11125.2	11065.8	11051.9	11067.9	11214.9	11205.2	11201.2
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=3$	11123.7	11116.2	11132.3	11074.7	11062.0	11061.8	11210.6	11195.2	11207.4
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=5$	11126.5	11115.4	11128.0	11069.1	11055.4	11056.6	11225.2	11185.2	11206.9
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=1$	11604.1	11582.4	11585.9	11570.0	11550.0	11544.7	11652.8	11651.3	11661.6
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=3$	11601.5	11577.1	11588.5	11572.0	11541.7	11547.4	11687.8	11644.0	11672.1
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=5$	11600.6	11586.8	11587.9	11569.2	11540.3	11548.9	11672.2	11671.7	11651.9

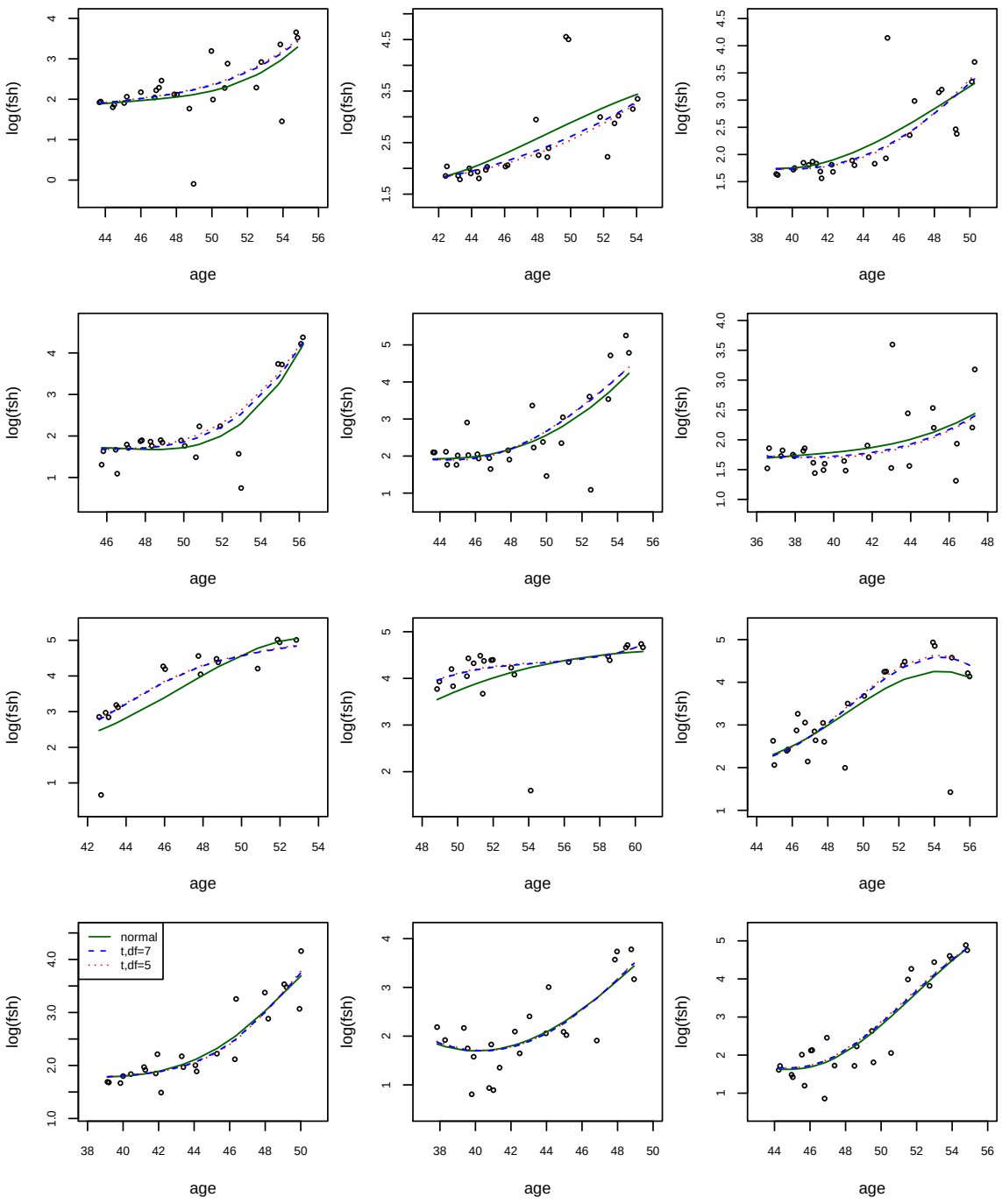

Figure 3.2: Selected individual trajectories fitted by our best-fitting joint models with $\mu_i(t)$ $i = 1,...,n$ within the time window $T = [45,55]$ as a functional predictor in primary outcome submodel: first row: fitted trajectories by all three models have upper bounds of p values less than 0.05; second row: fitted normal trajectories have upper bounds of p value less than 0.05 but fitted t (df=4) and t (df=7) trajectories have upper bounds of p value greater than 0.05; third row: fitted normal and t (df=7) trajectories have upper bounds of p value less than 0.05 but fitted t (df=4) trajectories have upper bounds of p value greater than 0.05; forth row: fitted trajectories by all three models have upper bounds of p values greater than 0.05.

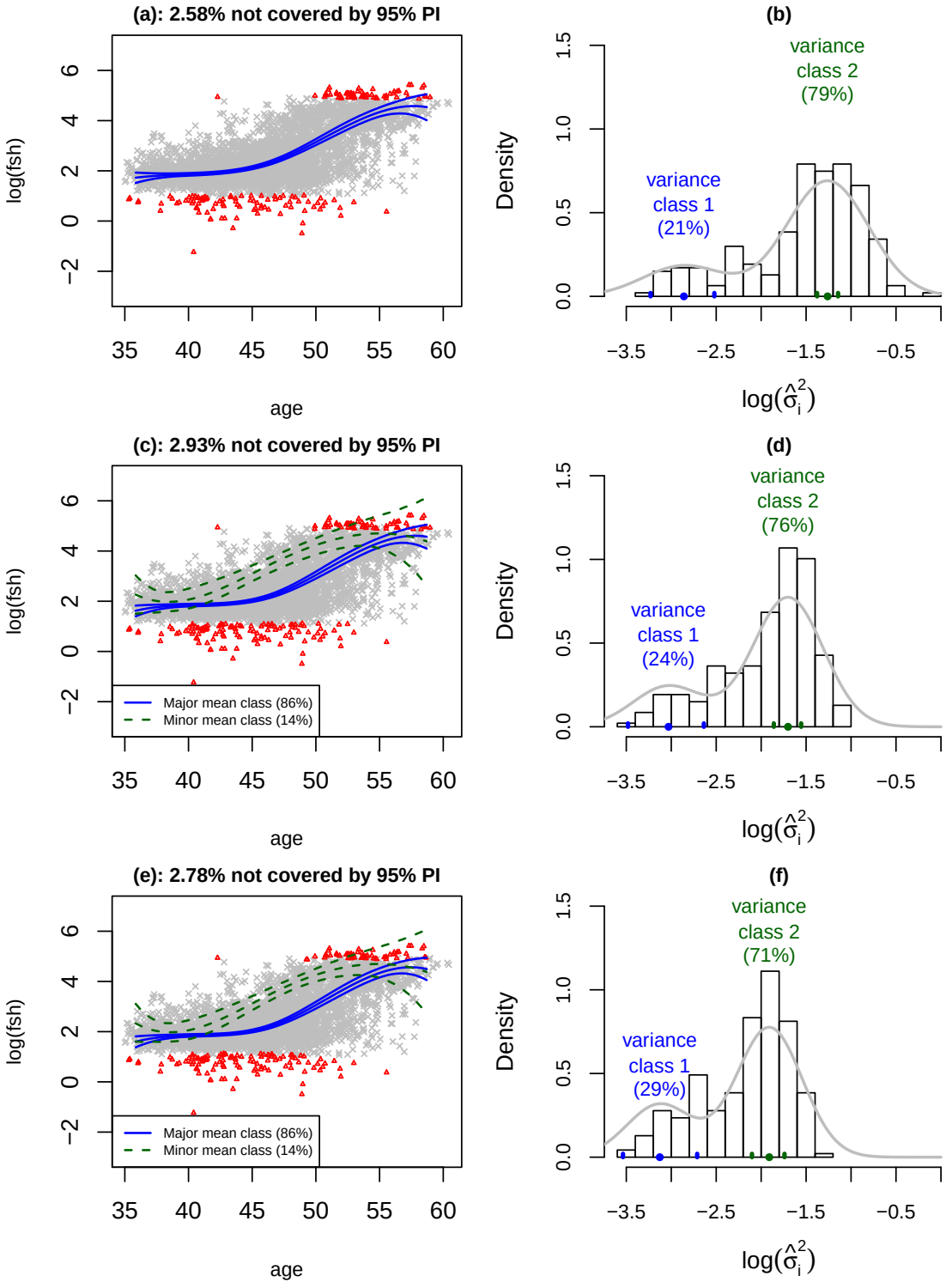

Figure 3.3: **longitudinal profiles** from our final models with $J_\mu = 1$ and $K_D = K_C = 2$ in longitudinal submodel; $\mu_i(t)$ as functional predictor with time window $T = [45, 55]$ and $J_{\theta_0} = 3$ in primary outcome submodel with different assumptions for longitudinal submodel: a) and b) under normal assumption; c) and d) under t assumption with df=7; e) and f) under t distribution with df=4.

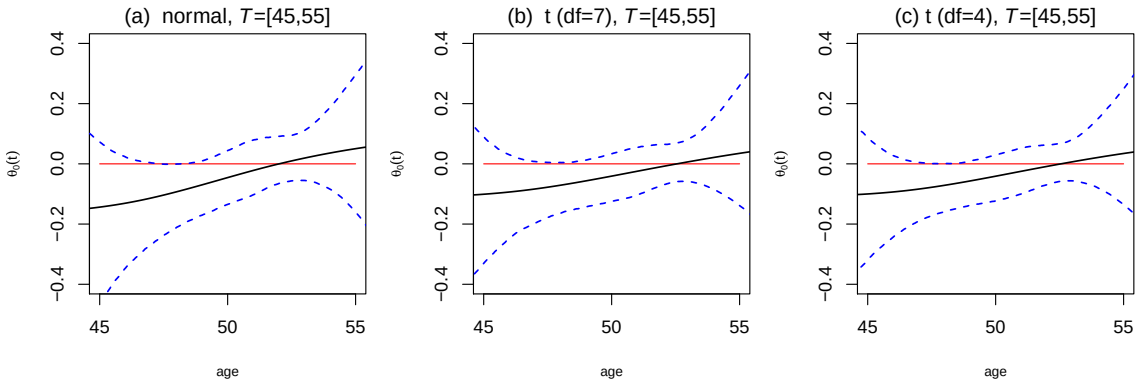

Figure 3.4: **Functional coefficient function** $\theta_0(t)$ from our best fitting t (df=4), t(df=7) and normal models with $\mu_i(t)$, $i = 1,...,n$ within the time window $T = [45,55]$ as a functional predictor in primary outcome submodel.

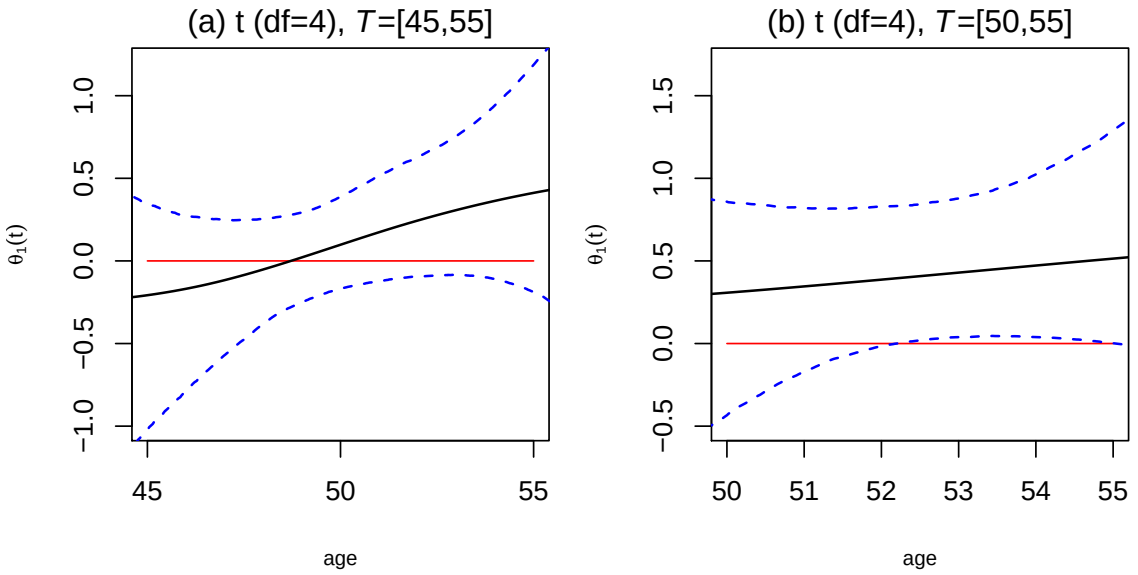

Figure 3.5: **Functional coefficient functions** $\theta_0(t)$ and $\theta_1(t)$ from our best-fitting model with $J_\mu = 1$ and $K_D = K_C = 2$ in longitudinal submodel with t (df=4) assumption; and $J_{\theta_0} = J_{\theta_1} = 3$ in primary outcome submodel: a) $\mu_i'(t)$ as functional predictor with $T = [45,55]$ and b) $\mu_i'(t)$ as functional predictor with $\tilde{T} = [50,55]$.

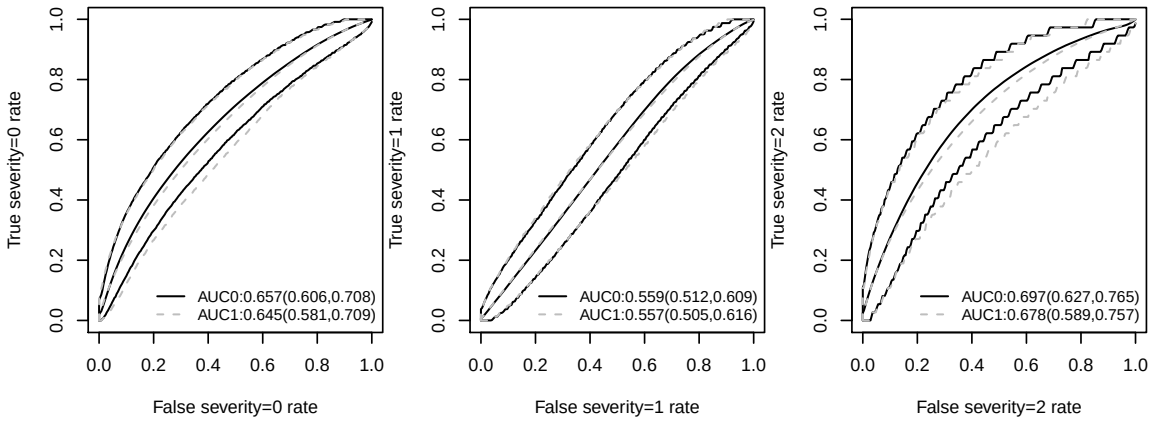

Figure 3.6: **ROC curves** from our final t model: AUC0 is obtained by using $\mu_i(t)$ with $J_{\theta_0(t)} = 3$ within the time window $T = [45,55]$ as a functional predictor in outcome submodel and AUC1 is obtained by using $\mu_i'(t)$ with $J_{\theta_1(t)} = 3$ within the time window $T = [45,55]$ as a functional predictor with in outcome submodel.

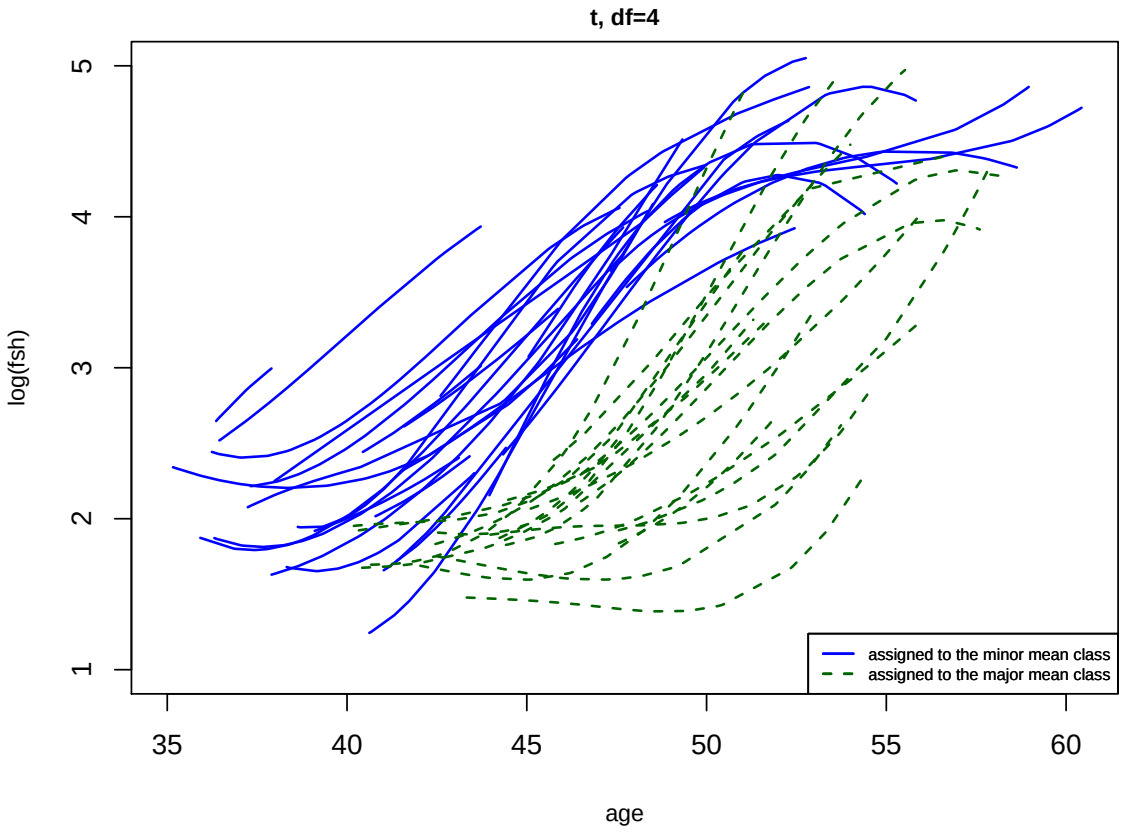

Figure 3.7: Individual trajectories that are assigned to the minor and major mean profile classes by our best fitting t (df=4) model with $\mu_i(t)$ $i = 1, ..., n$ within the time window $T = [45, 55]$ as a functional predictor in primary outcome submodel.

Table 3.2: Estimates of model parameters by our best fitting models with $\mu_i(t)$ $i = 1,...,n$ within the time window $T = [45,55]$ as a functional predictor in primary outcome submodel.

	normal			**t, df=7**			**t, df=4**		
p	mean	se	95% CI	mean	se	95% CI	mean	se	95% CI
β_{11}	1.666	0.145	(1.374, 1.95)	1.508	0.152	(1.195, 1.794)	1.475	0.152	(1.166, 1.766)
β_{12}	1.993	0.094	(1.808, 2.177)	2.118	0.094	(1.937, 2.308)	2.142	0.091	(1.963, 2.321)
β_{13}	1.286	0.153	(0.985, 1.585)	0.933	0.147	(0.643, 1.22)	0.889	0.144	(0.608, 1.175)
β_{14}	5.416	0.225	(4.981, 5.862)	5.506	0.24	(5.029, 5.969)	5.586	0.238	(5.115, 6.051)
β_{15}	4.253	0.445	(3.358, 5.114)	4.272	0.418	(3.44, 5.09)	4.122	0.396	(3.345, 4.897)
ω^2_{11}	0.94	0.265	(0.535, 1.556)	2.321	1.86	(0.786, 6.4)	2.357	1.597	(0.811, 6.39)
ω^2_{12}	0.606	0.137	(0.379, 0.915)	2.588	2.556	(0.698, 7.717)	2.838	2.357	(0.808, 8.107)
ω^2_{13}	2.022	0.396	(1.343, 2.887)	3.942	2.215	(1.327, 9.512)	3.644	1.911	(1.291, 8.433)
ω^2_{14}	5.68	0.927	(4.08, 7.692)	4.612	2.91	(1.376, 12.195)	4.232	2.51	(1.306, 10.495)
ω^2_{15}	9.43	3.235	(4.643, 17.328)	23.262	25.851	(3.639, 87.769)	17.166	16.618	(3.364, 61.141)
$\rho_{1,12}$	-0.763	0.073	(-0.877, -0.592)	-0.633	0.213	(-0.914, -0.096)	-0.653	0.202	(-0.917, -0.133)
$\rho_{1,13}$	0.661	0.098	(0.436, 0.816)	0.527	0.223	(0.002, 0.862)	0.54	0.214	(0.042, 0.857)
$\rho_{1,14}$	-0.199	0.146	(-0.468, 0.102)	-0.374	0.257	(-0.786, 0.203)	-0.381	0.249	(-0.783, 0.165)
$\rho_{1,15}$	-0.017	0.167	(-0.338, 0.31)	0.17	0.308	(-0.462, 0.707)	0.163	0.299	(-0.445, 0.688)
$\rho_{1,23}$	-0.818	0.047	(-0.894, -0.711)	-0.632	0.19	(-0.902, -0.169)	-0.644	0.182	(-0.903, -0.216)
$\rho_{1,24}$	0.459	0.103	(0.239, 0.64)	0.378	0.273	(-0.229, 0.808)	0.369	0.268	(-0.219, 0.802)
$\rho_{1,25}$	-0.172	0.15	(-0.452, 0.133)	-0.076	0.347	(-0.7, 0.589)	-0.061	0.333	(-0.671, 0.581)
$\rho_{1,34}$	-0.552	0.078	(-0.69, -0.385)	-0.861	0.116	(-0.974, -0.537)	-0.863	0.096	(-0.971, -0.606)
$\rho_{1,35}$	0.262	0.14	(-0.031, 0.517)	0.563	0.271	(-0.111, 0.909)	0.56	0.251	(-0.055, 0.897)
$\rho_{1,45}$	-0.744	0.08	(-0.865, -0.554)	-0.733	0.199	(-0.953, -0.198)	-0.722	0.197	(-0.948, -0.192)
β_{21}				2.5	0.527	(1.481, 3.574)	2.599	0.524	(1.612, 3.705)
β_{22}				1.233	0.439	(0.38, 2.118)	1.146	0.432	(0.275, 1.982)
β_{23}				4.129	0.493	(3.13, 5.103)	4.263	0.479	(3.336, 5.233)
β_{24}				5.267	0.544	(4.234, 6.387)	5.211	0.526	(4.196, 6.279)
β_{25}				3.996	1.371	(1.109, 6.594)	4.006	1.316	(1.313, 6.584)
ω^2_{21}				0.895	0.262	(0.507, 1.521)	0.909	0.265	(0.51, 1.535)
ω^2_{22}				0.55	0.125	(0.344, 0.835)	0.552	0.122	(0.347, 0.824)
ω^2_{23}				1.561	0.319	(1.027, 2.275)	1.597	0.323	(1.042, 2.303)
ω^2_{24}				5.367	0.891	(3.815, 7.292)	5.391	0.893	(3.835, 7.324)
ω^2_{25}				6.35	1.948	(3.326, 10.87)	6.017	1.744	(3.278, 10.062)
$\rho_{2,12}$				-0.769	0.07	(-0.881, -0.608)	-0.773	0.069	(-0.88, -0.613)
$\rho_{2,13}$				0.65	0.101	(0.421, 0.814)	0.657	0.1	(0.431, 0.818)
$\rho_{2,14}$				-0.253	0.157	(-0.545, 0.064)	-0.261	0.158	(-0.554, 0.065)
$\rho_{2,15}$				0.07	0.175	(-0.269, 0.41)	0.077	0.175	(-0.263, 0.417)
$\rho_{2,23}$				-0.858	0.041	(-0.923, -0.764)	-0.858	0.041	(-0.922, -0.763)
$\rho_{2,24}$				0.51	0.108	(0.279, 0.699)	0.511	0.106	(0.281, 0.7)
$\rho_{2,25}$				-0.282	0.151	(-0.557, 0.03)	-0.296	0.148	(-0.57, 0.008)
$\rho_{2,34}$				-0.686	0.073	(-0.809, -0.525)	-0.686	0.071	(-0.808, -0.529)
$\rho_{2,35}$				0.398	0.135	(0.112, 0.633)	0.4	0.132	(0.119, 0.637)
$\rho_{2,45}$				-0.745	0.085	(-0.871, -0.539)	-0.761	0.078	(-0.878, -0.575)
π^D_1				0.865	0.035	(0.787, 0.926)	0.861	0.033	(0.788, 0.918)
π^D_2				0.135	0.035	(0.074, 0.213)	0.139	0.033	(0.082, 0.212)
μ_1	-2.859	0.18	(-3.23, -2.521)	-3.03	0.223	(-3.481, -2.637)	-3.128	0.203	(-3.537, -2.712)
μ_2	-1.261	0.059	(-1.378, -1.144)	-1.701	0.078	(-1.859, -1.555)	-1.91	0.092	(-2.102, -1.74)
π^C_1	0.205	0.044	(0.126, 0.295)	0.242	0.07	(0.123, 0.384)	0.293	0.08	(0.148, 0.449)
π^C_2	0.795	0.044	(0.705, 0.874)	0.758	0.07	(0.616, 0.877)	0.707	0.08	(0.551, 0.852)
τ^2	0.208	0.046	(0.133, 0.309)	0.154	0.059	(0.065, 0.29)	0.134	0.072	(0.03, 0.319)
θ_0 (intercept)	0.305	0.995	(-1.631, 2.329)	0.279	0.972	(-1.637, 2.268)	0.012	0.985	(-1.886, 1.979)
θ_1 (log(BMI))	0.068	0.277	(-0.501, 0.607)	0.039	0.264	(-0.497, 0.573)	0.101	0.273	(-0.449, 0.627)
θ_2 (smoking)	0.386	0.17	(0.052, 0.717)	0.37	0.17	(0.039, 0.708)	0.371	0.171	(0.036, 0.708)
θ_3 (variance)	1.576	0.565	(0.498, 2.703)	1.887	0.747	(0.451, 3.394)	1.96	0.723	(0.579, 3.403)
$\tilde{\theta}_{01}$	-0.144	0.6	(-1.453, 1.125)	-0.104	0.554	(-1.3, 1.096)	-0.098	0.555	(-1.303, 1.138)
$\tilde{\theta}_{02}$	-0.158	0.624	(-1.4, 1.055)	-0.113	0.579	(-1.223, 0.945)	-0.11	0.565	(-1.198, 0.981)
$\tilde{\theta}_{03}$	-0.175	0.487	(-1.141, 0.624)	-0.117	0.413	(-0.888, 0.644)	-0.113	0.417	(-0.899, 0.642)
$\tilde{\theta}_{04}$	-0.139	0.144	(-0.444, 0.069)	-0.1	0.117	(-0.332, 0.086)	-0.101	0.118	(-0.329, 0.088)
$\tilde{\theta}_{05}$	0.037	0.187	(-0.268, 0.453)	0.009	0.159	(-0.29, 0.326)	0.007	0.162	(-0.297, 0.321)
$\tilde{\theta}_{06}$	0.085	0.317	(-0.542, 0.709)	0.072	0.278	(-0.428, 0.671)	0.078	0.286	(-0.412, 0.697)
$\tilde{\theta}_{07}$	0.09	0.668	(-1.185, 1.304)	0.082	0.59	(-0.905, 1.235)	0.08	0.582	(-0.942, 1.219)

CHAPTER IV

Bayesian Model Assessments in Evaluating Mixtures of Longitudinal Trajectories and Their Associations with Cross-Sectional Health Outcomes

4.1 Introduction

There is a growing body of literature that models information from longitudinal data to predict risks of health outcome of interest (Taylor et al., 2005; Yu et al., 2008; Proust-Lima and Taylor, 2009; Rizopoulos, 2011; Proust-Lima et al.,2012; Elliott et al., 2012; Taylor et al., 2013; Chapter II and III). An attractive feature of such predictions is that they are individualized. However, when several candidate models are available, the derived outcomes can be greatly affected by the use of different models. For example, different numbers of mixture components as well as the assumed association structure (e.g., multiple shared random effect vs. latent class structures) to link the longitudinal and primary outcome submodels can affect the target individualized predictions, as discussed in Chapter II. In general, researchers also tend to use the resulting latent classes to interpret the class-specific association findings. Various model selection criteria can be adopted to guide the selection of the proper number of components and the association structure. However, the performances of these model selection criteria have not been well studied in the joint modeling

framework.

Choosing the number of mixture components in a finite mixture setting is non-trivial. The difficulty arises mainly because the parameter estimation in finite mixture models is not a regular problem but a singular problem; hence log-likelihood function is not well approximated by a quadratic function, and maximum likelihood estimates are not asymptotically normal. See McLachlan and Peel (2000, section 6), Frühwirth-Schnatter (2006, section 4), Steele and Raftery (2010) as well as the references therein for thorough discussions of parameter estimation in finite mixture models.

Here we consider six model selection criteria. They are Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwartz, 1978), integrated classification likelihood criterion (ICL) (Biernacki et al., 1998), the deviance information criterion (DIC) (Spiegelhalter et al., 2002), the logarithm of the pseudo-marginal likelihood (LPML) (Geisser and Eddy 1979) and the widely applicable information criterion (WAIC) (Watanabe, 2010). AIC and BIC are long-standing and most commonly used information based model selection criteria in general; ICL is closely related to BIC with focus on the classification likelihood and entropy, whereas LPML has been most widely used in Bayesian model assessment. More details, including certain necessary modifications to accommodate the joint modeling framework are provided in Sections 4.3.1.3 to 4.3.1.5. DIC is often viewed as a Bayesian version of AIC with prior information on model parameters and is equivalent to AIC for non-hierarchical models with non-informative or flat priors. Many authors have proposed alternative versions of DIC. For example, Plummer (2002, 2006 and 2008) and Gelman et al. (2003) proposed alternative definitions of model complexity, while Celeux et al. (2006) proposed eight variations of DIC for "missing data" problems, including hierarchical models with latent variables.

WAIC, a recently proposed approach, was derived based on singular learning the-

ory (Watanabe, 2009) as an asymptotically unbiased approximation to the out-of-sample prediction error, and is a generalization of AIC that is applicable for both regular and singular statistical models. It is straightforward to compute based on the posterior draws, even for complex hierarchical models. Gelman et al. (2013) discussed the construction of AIC, DIC and WAIC from a Bayesian perspective using some simple examples and concluded that WAIC is a “very fast and computationally-convenient alternative" to their most preferred but often computationally expensive cross-validation approach to choose among several candidate models. However, its properties have not been studied in the setting of choosing numbers of components for finite mixture distributions.

Model assessment for the joint models consisting of mixture distributions as considered in Chapter II is even more challenging due to various reasons. First of all, the variables that are assumed to have mixture distributions are unobserved latent features of the longitudinal trajectories. AIC, BIC or DIC based on the observed data likelihood is not available in closed form. Secondly, the evaluation of model goodness-of-fit has to take into account of the model fits of longitudinal submodel and the primary outcome model jointly; this can be problematic, as the relatively larger gain in the fit of the primary outcome model, which contains a larger number of components, may dominate the overall model fit. This phenomenon could lead to favoring larger class models. Thirdly, when data are generated from a multiple shared random effect (MSRE) model, incorrectly assuming a latent class (LC) structure to link the longitudinal submodel and the primary outcome model has a high chance of creating an “outcome-informed artifact” as reported previously in Chapter II. When the primary outcome is binary and the information about the mixture components from the longitudinal data is weak, artificial mixture-components are created to match the two outcome groups of 0 or 1 under the assumed LC structure, which could lead to a false sense of accurate prediction than when the correct model is

fit. Despite the artificial enhancement in the predictive performance, such outcome-informed mixture-components are completely due to this artifact. Therefore, some naturally relevant and important questions arise: (i) whether such artifacts due to the assumed association structure in the primary outcome model would play any role in the performances of model selection criteria in choosing the numbers of mixture components; and (ii) whether model selection criteria would favor the assumed LC structure, which leads to seemingly better prediction over the MSRE structure in the presence of outcome-informed artifacts.

To address these issues, we conduct numerical studies to compare and contrast the performances of several commonly used model selection criteria. We consider WAIC and other modified criteria based on Bayesian principles in the joint modeling setting considered in Chapter II. Our main goals include understanding the performances of these commonly used criteria under different scenarios, including when the data-generating scheme differ from the assumed structure; gaining insights on similarity of selection performances of different criteria; and uncovering the model predictive performances based on out-of-sample validation, where the performances of the selected models are further linked with model selection criteria.

The remainder of this article is organized as follows. In Section 4.2, we describe the joint LC and MSRE models with mixture distributions for the mean and residual variances profiles of the longitudinal trajectories. In Section 4.3.1, we briefly review the Bayesian model assessment criteria as well as the overall model predictive performance measure to be included in our simulation. In Section 4.4, we describe our simulation study and report the outcomes. In Section 4.5, we describe the procedures used to validate prediction performances of the selected models by different criteria using newly generated independent samples. In Section 4.6, we study the performances of the selection criteria for the joint modeling of the follicle stimulating hormone trajectories and severity of hot flash for a group of middle-aged women

from the Penn ovarian aging study. Concluding remarks are given in Section 4.7.

4.2 Joint LC and MSRE models

Mixture modeling is commonly used to identify unique and distinct feature subgroups (i.e., latent classes) in longitudinal trajectories, e.g., the proposal of growth mixture models (GMMs) in Muthén and Shedden (1999). Chapter II considered two classes of joint models for normally distributed longitudinal data and a binary health outcome data. Both models used a generalized GMM for the longitudinal data. GMMs assume latent classes for the subject-level mean profiles. Our extension in Chapter II considered latent classes for not only the mean profiles but also the residual variabilities of the longitudinal trajectories. Specifically, the longitudinal submodel has the form

$$y_{ij}|\mathbf{b}_i, \sigma_i^2 \sim \mathrm{N}\{\mu(\mathbf{b}_i; t_{ij}), \sigma_i^2\}, \tag{4.1}$$

where, y_{ij} denotes the longitudinal covariate for the i^{th} subject at time t_{ij}, $j = 1,...,n_i$, $i = 1,...,n$, $\mathbf{b}_i$ is the vector of r random effects that reflects the subject-level mean profile/trajectory patterns and σ_i^2 is the subject-level residual variance. D_i and C_i define the latent class memberships for the individual mean profile and variance respectively:

$$\begin{aligned} D_i &\sim \text{Multinomial}(\pi_1^D, ..., \pi_{K_D}^D); \\ \mathbf{b}_i|D_i = d &\sim \mathrm{N}(\boldsymbol{\beta}_d, \Sigma_d), d = 1, ..., K_D \\ C_i &\sim \text{Multinomial}(\pi_1^C, ..., \pi_{K_C}^C); \\ \sigma_i^2|C_i = c &\sim \text{log-N}(\mu_c, \tau^2), c = 1, ..., K_C \end{aligned} \tag{4.2}$$

We consider two commonly used association structure to link longitudinal trajectory features with the binary outcome of interest: first, an MSRE structure, where the random effects $\mathbf{b}_i$, σ_i^2 and their interactions $\mathbf{b}_i \times \sigma_i^2$ are included as linear predictors in the primary outcome model and second, an LC structure, where the main and

interaction effects of latent classes D_i and C_i are included in the primary outcome model. In both cases, the primary outcome model can be written as

$$\Phi^{-1}(\mathrm{P}(o_i = 1)) = \mathbf{Z}_i'\boldsymbol{\eta}, \tag{4.3}$$

where o_i denotes the binary indicator of outcome, $\mathbf{Z}_i$ denotes a vector of the covariates in the probit model for subject i, $i = 1, \cdots, n$ and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. For the LC model, $\mathbf{Z}_i$ contains indicators for the latent classes D_i and C_i; for the MSRE model, $\mathbf{Z}_i$ contains the subject-specific random effect $\mathbf{b}_i$ and residual variance σ_i^2; other fully observed baseline covariates of interest can be included in $\mathbf{Z}_i$ in either model as well.

4.2.1 Likelihood specification

Let $\boldsymbol{\phi} = (\{\boldsymbol{\beta}_d\}_{d=1}^{K_D}, \{\Sigma_d\}_{d=1}^{K_D}, \{\pi_d^D\}_{d=1}^{K_D}, \{\mu_c\}_{c=1}^{K_C}, \{\pi_c^C\}_{c=1}^{K_C}, \tau^2, \boldsymbol{\eta})'$. We assume each component of $\boldsymbol{\phi}$ has an independent prior distribution with the joint prior distribution denoted by $\pi(\boldsymbol{\phi})$. All unobserved latent variables are denoted by $\mathbf{z}$, $\mathbf{z} = (\mathbf{b}, \boldsymbol{\sigma}, \mathbf{C}, \mathbf{D})'$. The observed data $\mathbf{x}$ consists of the longitudinal profiles $\mathbf{y}_1, ..., \mathbf{y}_n$ and the observed outcomes $o_1, ..., o_n$. Then the complete data likelihood of $\boldsymbol{\phi}$ based on the complete data $(\mathbf{x}, \mathbf{z})$ is given by

$$\begin{aligned} f(\mathbf{x}, \mathbf{z}|\boldsymbol{\phi}) \propto \Bigg\{ \prod_{i=1}^{n} \Bigg[& \prod_{d} \left[\pi_d (2\pi)^{-\frac{r}{2}} |\Sigma_d|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\mathbf{b}_i - \boldsymbol{\beta}_d\right)' \Sigma_d \left(\mathbf{b}_i - \boldsymbol{\beta}_d\right) \right\} \right]^{\mathrm{I}(D_i = d)} \\ & \times \prod_{c} \left[\pi_c (2\pi\tau^2)^{-\frac{1}{2}} \sigma_i^{-2} \exp\left\{ -\frac{1}{2\tau^2} \left(\log\sigma_i^2 - \mu_c\right)^2 \right\} \right]^{\mathrm{I}(C_i = c)} \\ & \times \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2} \left\{ y_{ij} - f(\mathbf{b}_i; t_{ij}) \right\}^2 \right] \\ & \times \Phi(\mathbf{Z}_i'\boldsymbol{\eta})^{o_i} \left\{ 1 - \Phi(\mathbf{Z}_i'\boldsymbol{\eta}) \right\}^{1-o_i} \Bigg] \Bigg\} \pi(\boldsymbol{\phi}). \end{aligned} \tag{4.4}$$

4.2.2 Prior specification and posterior computation

We consider the same prior distributions as considered in Chapter II, where certain empirical data-driven priors were considered for some parameters to avoid either improper posterior or existence of empty classes during the iterations of MCMC sampling. We found that our considered model assessment criteria were not sensitive to these choices.

For the mixture normal distribution of the random effects, we let $\boldsymbol{\beta}_d \sim \text{N}(\mathbf{0}, \mathbf{V})$, $\mathbf{V} = n\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the linear regression estimator of $\mathbf{y}$ on the design matrix of $\mathbf{t}$ defined by $f(\cdot; t_{ij})$. This corresponds to a prior "single observation" data-driven inflated covariance structure centered at a null model, and avoids improper posteriors resulting from the possibility that some of the latent classes are not represented in the data (Elliott et al. 2005). For the the variance-covariance matrix for the random effects Σ_d, we use the prior proposed by Kass and Natarajan (2006): $\Sigma_d \sim$ Inverse-Wishart(df $= r, \Lambda$), where $\Lambda = r\left(\sum_{i=1}^{n} \widehat{\text{Cov}}(\tilde{\mathbf{b}}_i)^{-1}/n\right)^{-1}$, $\tilde{\mathbf{b}}_i$ is given by OLS estimator of $\mathbf{b}_i$ for subject i and r is the dimension of $\mathbf{b}_i$. This prior restrains the eigenvalues of the variance-covariance matrices away from 0, avoiding the improper posterior that can result from unbounded likelihoods when the variance-covariance matrix is unrestricted in normal mixture models (Day, 1969).

For the mixture log normal distribution for the residual variances, we used diffuse priors: $\mu_c \sim \text{N}(0, \upsilon)$, $\tau^{-2} \sim \text{Gamma}(a, b)$ with $\upsilon = 1000$ and $a = b = .001$. For the class membership probabilities, we assume conjugate Dirichilet$(4, ..., 4)$ on both $\pi^C = (\pi_1^C, ..., \pi_{K_C}^C)$ and $\pi^D = (\pi_1^D, ..., \pi_{K_D}^D)$ (Frühwirth-Schnatter, 2006); this is equivalent to assuming *a priori* 4 observations in each class, avoiding the existence of empty classes. Lastly, we let $\theta \sim \text{N}(\mathbf{0}, (9/4)\mathbf{I})$ in the probit regression, where $(9/4)\mathbf{I}$ is chosen to bound the estimated outcome probabilities to be away from 0 and 1 (Garrett and Zeger, 2000; Elliott et al., 2007, and Neelon et al., 2011).

Gibbs sampling is used to obtain draws from the corresponding posterior distribu-

tions. For $(\boldsymbol{\eta} \mid \mathbf{C}, \mathbf{D}, \mathbf{O})$ we use the Albert and Chib (1993) data augmentation method for probit regression models. The draws of $(\sigma_i^2 | C_i, \mu_c, \tau^2, \mathbf{b}_i, o_i, \{y_{ij}\}_j)$ for $i = 1, ..., n$ are obtained the inverse cumulative distribution method. The detailed MCMC sampling procedures are provided in the Appendix A.

4.3 Model selection and assessment criteria

4.3.1 Model selection criteria

We consider several commonly used model selection criteria that are both computationally feasible and stable for our joint models. For a comprehensive review of Bayesian model selection criteria, please refer to Vehtari and Ojanen (2012).

4.3.1.1 Log-pseudo marginal likelihood criterion

The log-pseudo marginal likelihood (LPML) (Geisser and Eddy, 1979) corresponds to a Bayesian cross-validation measure, defined as $\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i)$, where $\text{CPO}_i = f(\mathbf{y}_i, o_i | \mathbf{y}_{(-i)}, \mathbf{o}_{(-i)})$ represents a cross-validated posterior predictive density for $\mathbf{x}_i = (\mathbf{y}_i, o_i)$ given the data excluding $(\mathbf{y}_i, o_i)$ (denoted by $(\mathbf{y}_{(-i)}, \mathbf{o}_{(-i)}) = \mathbf{x}_{(-i)}$). The model with higher value of LPML provides better fit to the data (Ibrahim et al., 2001). Details of the LPML computation are provided in the Appendix A.

4.3.1.2 Deviance information criterion

DIC (Spiegelhalter et al., 2002) is a Bayesian analog of the original AIC (Akaike, 1974), but DIC uses the discrepancy between the posterior mean of the deviance $\overline{D(\boldsymbol{\phi})} = E_{\boldsymbol{\phi}} \left\{ -2 \log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x} \right\}$ and the deviance evaluated at the posterior mean $D(\overline{\boldsymbol{\phi}}) = -2 \log f \left\{ \mathbf{x} \mid E(\boldsymbol{\phi} \mid \mathbf{x}) \right\}$ to estimate the effective number of degrees of freedom in the

model p_D:

$$\begin{aligned}
\mathrm{DIC}(\mathbf{x}) &= \overline{D(\boldsymbol{\phi})} + p_D \\
&= 2\overline{D(\boldsymbol{\phi})} - D(\overline{\boldsymbol{\phi}}) \\
&= D(\overline{\boldsymbol{\phi}}) + 2p_D \\
&= -4\mathrm{E}_{\boldsymbol{\phi}}\left\{\log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2\log f\left\{\mathbf{x} \mid \mathrm{E}(\boldsymbol{\phi} \mid \mathbf{x})\right\}
\end{aligned} \tag{4.5}$$

In our setting, the observed data likelihood $f(\mathbf{x} \mid \boldsymbol{\phi})$ is not available in closed form, where $\mathbf{x} = (\mathbf{y}, \mathbf{o})'$; instead we use the approach outlined in Celeux et al. (2006) to obtain $\mathrm{E}_{\mathbf{z}}\{\mathrm{DIC}(\mathbf{x},\mathbf{z})\} = -4\mathrm{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\log f(\mathbf{x},\mathbf{z} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2\mathrm{E}_{\mathbf{z}}\left[\log f\left\{\mathbf{x},\mathbf{z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x},\mathbf{z})\right\} \mid \mathbf{x}\right]$, where the complete data likelihood $f(\mathbf{x},\mathbf{z} \mid \boldsymbol{\phi})$ has a closed form as specified in (4.4), and $\mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x},\mathbf{z})$ is obtained via numerical methods. The detail of DIC computation is provided in the Appendix A.

4.3.1.3 Modified AIC

Although the original AIC proposed by Akaike (1974) is developed for "regular" models and hence is not directly defined for Bayesian hierarchical model, we consider AIC modified based on the complete data likelihood using Bayesian principle. Specifically, the modified AIC is defined using the deviance based on the complete data likelihood with a penalty term to account for the number of model parameters as follows:

- $\mathrm{AIC}_1 = \overline{D(\boldsymbol{\phi})} + 2p = -2\mathrm{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\log f(\mathbf{x},\mathbf{z} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2p$
- $\mathrm{AIC}_2 = D(\overline{\boldsymbol{\phi}}) + 2p = -2\mathrm{E}_{\mathbf{z}}\left[\log f\left\{\mathbf{x},\mathbf{z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x},\mathbf{z})\right\} \mid \mathbf{x}\right] + 2p$

where for joint LC model, $p = K_D + rK_D + r^2K_D + K_C + K_C + K_CK_D + 1$ and for joint MSRE model, $p = K_D + rK_D + r(r+1)K_D/2 + K_C + K_C + 2(r+1) + 1$, where r is the dimension of random effect $\mathbf{b}_i$. For both models, there are K_D parameters for π_d, rK_D for μ_d, $r(r+1)K_D/2$ for Σ_d in the mean profile; there are K_C parameters for π_c, K_C for μ_c, 1 for τ^2 in the variance profile. For the LC structure, since we consider a

saturated model with all possible main and interaction effect between the mean and the variance profiles, there are $K_C K_D$ parameters in the primary outcome model; for the MSRE structure, since we consider all possible main and interaction effect between random effects $\mathbf{b}_i$ and variances σ_i^2, there are $2(r+1)$ parameters in the primary outcome model.

4.3.1.4 Modified BIC and ICL

Accordingly, we consider the following modified BIC's that correspond to the above definition of AIC's:

- $\text{BIC}_1 = \overline{D(\boldsymbol{\phi})} + p\log(n) = -2\text{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\log f(\mathbf{x},\mathbf{z} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + p\log(n)$
- $\text{BIC}_2 = D(\overline{\boldsymbol{\phi}}) + p\log(n) = -2\text{E}_{\mathbf{z}}\left[\log f\left\{\mathbf{x},\mathbf{z} \mid \text{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x},\mathbf{z})\right\} \mid \mathbf{x}\right] + p\log(n)$

To identify the correct number of components for finite mixture distributions, Biernacki et al. (1998) also suggested an integrated classification likelihood criterion (ICL), which was shown by McLachlan and Peel (2006, page 216) to be approximately equal to BIC plus two times the entropy of classification probability into assumed number of clusters. Here, we adopt this approximated version of ICL. Further, given that we have two mixture distributions for the random effects and the residual variances, respectively, we have the following two forms of ICL,

- $\text{ICL}_1 = \text{BIC}_1 + 2\text{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\text{EN}(\boldsymbol{\phi},\mathbf{z}) \mid \mathbf{x}\right\}$
- $\text{ICL}_2 = \text{BIC}_2 + 2\text{E}_{\mathbf{z}}\left[\text{EN}\left\{\text{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x},\mathbf{z}),\mathbf{z}\right\} \mid \mathbf{x}\right]$

where

$$\begin{aligned}\text{EN}(\boldsymbol{\phi},\mathbf{z}) = &-\sum_{d=1}^{K_D}\sum_{i=1}^{n}\text{P}(D_i = d \mid \boldsymbol{\phi},\mathbf{z})\log\text{P}(D_i = d \mid \boldsymbol{\phi},\mathbf{z}) \\ &-\sum_{c=1}^{K_C}\sum_{i=1}^{n}\text{P}(C_i = c \mid \boldsymbol{\phi},\mathbf{z})\log\text{P}(C_i = c \mid \boldsymbol{\phi},\mathbf{z})\end{aligned} \tag{4.6}$$

where, the expressions of $\text{P}(D_i = d \mid \boldsymbol{\phi},\mathbf{z})$ and $\text{P}(D_i = d \mid \boldsymbol{\phi},\mathbf{z})$ differ for LC and MSRE models and are given in the Appendix C.

4.3.1.5 WAIC

Following Gelman et al. (2013), we consider the following two forms of WAIC, defined based on the conditional data likelihood $f(\mathbf{x}_i \mid \mathbf{z}, \boldsymbol{\phi})$:

$$\text{WAIC}_i = -2 \sum_{i=1}^{n} \log \left[\text{E}_{\mathbf{z},\boldsymbol{\phi}} \left\{ f(\mathbf{x}_i \mid \mathbf{z}, \boldsymbol{\phi}) \mid \mathbf{x} \right\} \right] + 2 p_{\text{WAIC}i} \tag{4.7}$$

- $p_{\text{WAIC}1} = 2 \sum_{i=1}^{n} \left[\log \left[\text{E}_{\mathbf{z},\boldsymbol{\phi}} \left\{ f(\mathbf{x}_i \mid \mathbf{z}, \boldsymbol{\phi}) \mid \mathbf{x} \right\} \right] - \text{E}_{\mathbf{z},\boldsymbol{\phi}} \left[\log f(\mathbf{x}_i \mid \mathbf{z}, \boldsymbol{\phi}) \mid \mathbf{x} \right] \right]$
- $p_{\text{WAIC}2} = \sum_{i=1}^{n} \text{Var}_{\mathbf{z},\boldsymbol{\phi}} \left[\log f(\mathbf{x}_i \mid \mathbf{z}, \boldsymbol{\phi}) \mid \mathbf{x} \right]$

where $f(\mathbf{x}_i \mid \mathbf{z}, \boldsymbol{\phi}) = \left[o_i \Phi(\mathbf{Z}_i^T \boldsymbol{\eta}) + (1 - o_i)(1 - \Phi(\mathbf{Z}_i^T \boldsymbol{\eta})) \right] \prod_{j=1}^{n_i} \phi(y_{ij} | \mu(\mathbf{b}_i; t_{ij}), \sigma_i^2)$, and $\phi(y_{ij} | \mu(\mathbf{b}_i; t_{ij}), \sigma_i^2)$ represents the normal density with mean $\mu(\mathbf{b}_i; t_{ij})$ and variance σ_i^2 evaluated at y_{ij}.

4.3.2 Overall model performance measure

For each model considered, we also summarize its predictive performance to link with different model selection criteria. There are many choices to quantify the performance of a predictive model for binary prediction (Taylor et al., 2008 and Steyerberg et al., 2010). Here, we consider the widely-used the area under the curve (AUC) based on the receiver operating characteristic (ROC) curve to assess the overall model discrimination ability averaged across all predictive cutoffs. For out-of-sample prediction validation in Section 4.5, we consider the Brier-score based posterior predictive mean squared error as an additional performance measurement; details are provided therein.

Briefly, the ROC curve plots true positive rate (TP) versus false positive rate (FP) for all possible cutoffs based on predicted $\text{P}(o_i = 1) = \Phi(\mathbf{Z}_i' \boldsymbol{\eta})$ obtained from (4.3). The ROC curve and AUC were computed at each MCMC iteration using the ROCR package in R (Sing et al. 2005). The ROC curve is computed by ordering

the observations $(i) = 1,...,n$ so that $\hat{\mathrm{P}}(o_{(i)} = 1) \geq \hat{\mathrm{P}}(o_{(i+1)} = 1)$, computing change-points $c = 2,...,n_c$, $n_c \leq n$ where the observations change from positive to negative (i.e., $o_{(c-1)} = 1$, $o_{(c)} = 0$), and plotting $\sum_{(i)=1}^{c}(1-o_{(i)})/\sum_{(i)=1}^{n}(1-o_{(i)})$ on the x-axis versus $\sum_{(i)=1}^{c} o_{(i)}/\sum_{(i)=1}^{n} o_{(i)}$ on the y-axis. The AUC is then computed using a trapezoidal approximation. The reported AUC is calculated as the posterior mean AUC averaged across all MCMC iterations.

4.4 Simulation Study

In this section, we conduct several simulation studies to evaluate the performances of these model selection criteria under both scenarios that the data generating schemes, LC or MSRE, and the fitted model coincide with each other and when they do not. In what follows, we refer to the model where the observations are generated from as the "true model". Two representative data-generating structures, $K_D = K_C = 2$ and $K_D = K_C = 1$, are considered. The former, with different combinations of relative mixture locations, represents a simple but informative mixture structure; while the latter represents the null model. We report the number of components selected by each criterion and under each scenario. We also report the within and out-of samples prediction performances.

4.4.1 Simulation Setup

We specify a combination of sub-structures for our simulation studies below.

4.4.1.1 $K_D = K_C = 2$

For the longitudinal observations, we generate data from the following models with two components within both the mean and the variance profiles,

$$\begin{aligned} y_{ij}|\mathbf{b}_i, \sigma_i^2 &\sim \mathrm{N}(b_{0i} + b_{1i}t_{ij}, \sigma_i^2), \\ \mathbf{b}_i &\sim \pi\mathrm{N}(\beta_1, \Sigma_1) + (1-\pi)\mathrm{N}(\beta_2, \Sigma_2), \\ \log\sigma_i^2 &\sim \pi\mathrm{N}(\mu_1, \tau^2) + (1-\pi)\mathrm{N}(\mu_2, \tau^2), \end{aligned} \tag{4.8}$$

where $i = 1, ..., 200$ and $t_{ij} = 0, 1, 2, ..., n_i$ with $n_i \equiv 20$. For $k = 1, 2$, we denote $\beta_k = (\beta_{k1}, \beta_{k2})'$ and $\Sigma_k = \begin{pmatrix} \omega^2 & \rho_k\omega^2 \\ \rho_k\omega^2 & \omega^2 \end{pmatrix}$.

We let $\beta_1 = (0,0)'$ and $\beta_2 = (2\sqrt{2}, 2\sqrt{2})'$, $\rho_1 = 0$ and $\rho_2 = -0.6$, $\mu_1 = -2$ and $\mu_2 = -.5$. Thus the mean of the two bivariate normals differs by 4 throughout, while the mean log of the variances are separated by 1.5. In our investigation, we consider the cases of "overlapped" versus "separated" mixture components, crossed with "balanced" 50:50 versus "unbalanced" 20:80 mixing proportions for both the mean and the variance profiles. Besides the separation in mixture components, we anticipate the mixing proportion of 50:50 to yield more difficult to separate latent classes, since the populations provide no information about class memberships. Figure 4.1 shows the corresponding 95% contours and density plots of the two "overlapped" versus "separated" components for the mean and the variance profile, respectively. Finally, our eight longitudinal model scenarios are defined by (ω^2, τ^2, π) = (2,.25,.5), (1,.25,.5), (2,.06,.5), (1,.06,.5), (2,.25,.2), (1,.25,.2), (2,.06,.2), and (1,.06,0.2), respectively.

For each scenario, we simulate 100 data sets and report the models (i.e., K_D and K_C) selected by each selection criteria. The setups considered include the combinations where the assumed and fitted structures from the choices of LC and MSRE models match or dis-match each other, with the assumed numbers of components

being $K_D = 1,2,3$ and $K_C = 1,2,3$.

For each of the simulation scenarios proposed for the longitudinal observations, the following two underlying probit models are considered for the health outcome:

1. **latent class (LC) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \theta_0 + \theta_1 \mathrm{I}(D_i = 2) + \theta_2 \mathrm{I}(C_i = 2) + \theta_3 \mathrm{I}(D_i = 2, C_i = 2); \tag{4.9}$$

2. **multiple shared random effect (MSRE) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \gamma_0 + \gamma_1 b_{i0} + \gamma_2 b_{i1} + \gamma_3 \sigma_i^2 + \gamma_4 b_{i0}\sigma_i^2 + \gamma_5 b_{i1}\sigma_i^2, \tag{4.10}$$

where $D_i = 1$ corresponds to the mean component $\mathrm{N}((\mathbf{0},\mathbf{0})', \Sigma_1)$, and $C_i = 1$ corresponds to the variance component $\mathrm{N}(-2, \tau^2)$ in the longitudinal submodel (4.8). We replace $\boldsymbol{\eta}$ in the general models (4.3) by $\boldsymbol{\theta}$ for the LC and by $\boldsymbol{\gamma}$ for the MSRE probit primary models to simplify the task of presentation; we let $\boldsymbol{\theta} = (-0.8, 1.8, -.2, -.3)$ and $\boldsymbol{\gamma} = (-1,1,-1,2,-2,2)'$ for each scenario so that the outcome prevalence is approximately 50 percent.

4.4.1.2 $K_D = K_C = 1$

Further, we consider the null case when there does not exist mixture/latent class for either the mean or the variance profiles by dropping the corresponding second mixture components where $D_i = C_i = 0$. Consequently, the primary probit models for the health outcomes are:

1. **latent class (LC) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \theta_0; \tag{4.11}$$

2. **multiple shared random effect (MSRE) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \gamma_0 + \gamma_1 b_{i0} + \gamma_2 b_{i1} + \gamma_3 \sigma_i^2 + \gamma_4 b_{i0}\sigma_i^2 + \gamma_5 b_{i1}\sigma_i^2. \tag{4.12}$$

For each scenario, we simulate 100 data sets and report the models (i.e., K_D and K_C) selected by each selection criteria under variour scenarios, equivalently to those in Section 4.4.1.1, but only consider the fitted models with $K_D = 1,2$ and $K_C = 1,2$.

4.4.2 Simulation Results

4.4.2.1 $K_D = K_C = 2$

Among the 100 simulated data sets, in Tables 4.1 to 4.4, we report the number of times each model, indicated by particular numbers of mixture components (K_D, K_C), is selected by one of the criteria given in the first column. Table 4.1 and 4.2 show the results when the true data-generating model has LC structure while Table 4.3 and 4.4 reports those under the MSRE structure. In general, separations in mixture components play an important role in the performances of these criteria in identifying the correct number of components: when there is a sufficiently large degree of separation in either the mean or the variance profiles, it is generally easier to choose the correct number of components. Scenario (a)-(d) represent different levels of separation of mean (or variance) components, as indicated in the Tables. Mixing proportion might also play a role in selecting correct numbers of mixture components. We use an unbalanced 20% vs. 80% mixing design to create some asymmetry in the mixture density. In our study, all criteria seem to perform slightly better in the cases of the unbalanced design. Incorrectly assuming the outcome structure has some impacts on the performances of these criteria, the degrees of impact depend on the criteria and hence the goal in model selection, as well as the true association structure in the outcome model. In particular, the outcome-informed artifacts due to

fitting an LC structure to the data generated under a true MSRE model, reported in Chapter II, also have some connection with the performances of these criteria, which we re-visit later in this section.

Overall, under the correctly assumed association structure in the primary outcome model, the modified AIC, BIC and ICL perform very well in selecting the correct numbers of components even for the harder to separate mixture scenarios. When the true structure in the primary outcome model is LC, the modified AIC, BIC and ICL still perform equally well regardless of fitting LC or MSRE models. However, when the true structure is MSRE, fitting an LC structure can affect the performances of the modified AIC, BIC and ICL. This phenomenon is most prominent for scenario (a) in Table 4.3. The reported results corresponding to the two different versions of AIC, BIC, ICL or WAIC differ sometimes, but not to a noteworthy level, and therefore we do not differentiate the summary according to the versions used.

In contrast, DIC, LPML and WAIC tend to choose too many components for both the mean and the variance in all scenarios. In particular, the numbers of mixture components selected by WAIC and LPML tend to agree regardless of the fitted model structure used in the primary outcome model. More interestingly to note is that, when fitting with a joint LC model, both WAIC and LPML tend to always select the numbers of components that lead to a higher AUC value. When fitting with a joint MSRE models, WAIC and LPML still tend to select models with more mixture components for both the mean and the variance, but the model based on such a selection does not have a higher value of the AUC. In fact, the AUC values vary little under different fitted models.

As we can see from the AUC values given in Tables 4.1 and 4.2, even correctly assuming the LC structure can lead to either (i) lower or (ii) higher AUC values than the AUC values by the true models. We believe that (i) is due to the difficulty to separate the mixture components of the means and, as a consequence, leads to

high potential for misclassification. An extreme case is when almost all subjects are assigned to one mixture component and the prediction of the outcome is solely determined by the variance profile, which results in worse prediction ability than the true model. On the other hand, (ii) is likely due to the outcome-informed artifact, where subjects are assigned to spurious mixture components to generate predictions that are more accurate than those obtained with the true model, as discussed in Chapter II. Meanwhile, fitting joint MSRE models when LC is the true structure only leads to slightly loss of the predictive power relative to the true model, when the mixture components within the mean profile overlap; otherwise the AUC values obtained by the MSRE model are similar to the true model.

On the other hand, as shown in Tables 4.1 and 4.2, when MSRE is the true data-generating mechanism, the AUC values obtained by fitting the joint MSRE models are always close to the AUC values by the true models in all scenarios we consider, indicating that the predictive ability under such scenarios is not affected much by any potential misclassification due to the overlapping mixture components. However, when MSRE is the true structure, fitting LC models would lead to either an increase in predictive power, indicated by extremely high AUC values due to artificially recognized new components, or loss of predictive power with low AUC values, originated from replacing a set of continuous variables (i.e., MSRE) by a discretized version (i.e., LC) in the primary outcome model. Because of these mentioned potentials when fitting joint LC and MSRE models, all criteria suggest that it is difficult in differentiating the LC and MSRE models under the scenarios with overlapping components. Table 4.5, focusing on the performances of all criteria in distinguishing between LC and MSRE assumed structure with the true numbers of mixture components, further illustrates this phenomenon. Each criterion could be in favor of the structure other than the true data-generating one, notably when the mixture components overlap. The performances of LPML and WAIC consistently reflect such

difficulties by frequently favoring the models with higher AUC values regardless of the truth.

4.4.2.2 $K_D = K_C = 1$

As shown in Table 4.6, under the true LC model and for all fitting structures, the modified AIC, BIC and ICL all perform very well when distinct mixture components do not exist for both the mean and the variance. The AICs behave slightly different from BIC and ICL, in favor of more complex structures at times. DIC, LPML and WAIC have the tendency to select models with too many mixture components, where LPML and WAIC tend to select the number of components that lead to higher AUC values. This once again suggests that WAIC and LPML tend to select models with high prediction accuracy regardless of true association structure.

Under this null structure, we clearly see that the joint LC model has the ability to create additional outcome-informed components that lead to misleadingly high AUC values. The highest spurious AUC value is 0.97 as shown in Table 4.6 when the true model is MSRE, giving the impression of almost perfect prediction when there is no mixture at all in either the mean or the variance profiles. In this case, all criteria under the assumed LC model tend to choose more numbers of components than that of the data-generating scheme, reaching a better goodness-of-fit.

4.5 Validation of the models selected by different criteria

It is well known that using the predictive model built on the same data set where the prediction is conducted would lead to optimistically biased prediction evaluation. In this section, we conduct evaluations of different model-selection criteria using newly generated independent samples to obtain a fair assessment of the predictive performances. With Tables 1 to 4.6 showing that the key over-fitting phenomenon reflected by AUC is preserved in the the simplest data-generation scheme of $K_D =$

$K_C = 1$, we focus on this setup and again allow the fitted models to have 1 or 2 components. We choose this particular scenario as an extreme case of completely overlapped mixture components to amplify the effects of potential outcome-informed artifacts by fitting LC models, since the true LC model is essentially a null model with AUC=0.5. For this simplest scenario, we observe that the joint LC models with $K_D > 1$ or $K_C > 1$ could lead to exceedingly high AUC values relative to the true AUC, and that such joint LC models are frequently favored by LPML and WAIC. On the positive side, when the observations are generated from the MSRE model but fitted with LC ones, the models selected by LPML, WAIC and DIC do result in better prediction performances on the validation samples. The obtained outcomes also allow us to see whether such high predictive accuracy is real or an artifact.

For each data set $\{(\mathbf{y}_i^{(s)}, o_i^{(s)})\}_{i=1}^n$, $s = 1, \cdots, 100$, generated from the given true joint LC and MSRE models, we generate an additional new validation data set $\{(\tilde{\mathbf{y}}_i^{(s)}, \tilde{o}_i^{(s)})\}_{i=1}^{\tilde{n}}$, $\tilde{n} = 50$. We use $\mathrm{H}_a^{(s)}$ to denote the model selected by model selection criteria $a \in \{\mathrm{DIC}, \mathrm{LPML}, \mathrm{AIC}_1, \mathrm{BIC}_1, \mathrm{ICL}_1, \mathrm{WAIC}_1\}$ for the data set $\{(\mathbf{y}_i^{(s)}, o_i^{(s)})\}_{i=1}^n$ and then each $\mathrm{H}_a^{(s)}$ has unique values of K_D and K_C.

When introducing our target prediction, we drop the superscript (s) in the notations for notational simplicity. We also split the model parameter vector $\boldsymbol{\phi}$ into two components: 1) $\boldsymbol{\phi}_{\text{long}} = \left(\{\boldsymbol{\beta}_d\}_{d=1}^{K_D}, \{\Sigma_d\}_{d=1}^{K_D}, \{\pi_d^D\}_{d=1}^{K_D}, \{\mu_c\}_{c=1}^{K_C}, \tau^2, \{\pi_c^C\}_{c=1}^{K_C}\right)^T$, including all the population level parameters in the longitudinal submodel, and 2) $\boldsymbol{\eta}$, the vector of coefficients in the primary outcome model. We let $\tilde{\mathbf{z}} = \left(\tilde{\mathbf{D}}, \tilde{\mathbf{C}}, \tilde{\mathbf{b}}, \tilde{\sigma}\right)^T$ include all individual level latent variables for the new validation sample and, compatibly, $\mathbf{z} = (\mathbf{D}, \mathbf{C}, \mathbf{b}, \sigma)^T$ includes all such latent variables for the original sample set used to obtain the fitted model. The prediction of the primary outcome for new validation

sample is then based on the following quantity,

$$
\begin{aligned}
& p(\tilde{\mathbf{o}}|\tilde{\mathbf{y}},\mathbf{y},\mathbf{o},\mathrm{H}_a) \\
&= \int p\left(\tilde{\mathbf{o}}|\tilde{\mathbf{z}},\boldsymbol{\eta},\mathrm{H}_a\right) p\left(\tilde{\mathbf{z}},\mathbf{z},\boldsymbol{\phi}_{\text{long}},\boldsymbol{\eta}|\tilde{\mathbf{y}},\mathbf{y},\mathbf{o},\mathrm{H}_a\right) d\boldsymbol{\phi}_{\text{long}} d\boldsymbol{\eta} d\tilde{\mathbf{z}} d\mathbf{z} \\
&\simeq \frac{1}{M}\sum_{m=1}^{M} p\left(\tilde{\mathbf{o}}|\tilde{\mathbf{z}}^{(m)},\boldsymbol{\eta}^{(m)},\mathrm{H}_a\right) \qquad (4.13)
\end{aligned}
$$

where, $\tilde{\mathbf{z}}^{(m)},\mathbf{z}^{(m)},\boldsymbol{\phi}_{\text{long}}^{(m)}$ and $\boldsymbol{\eta}^{(m)}$, for $m = 1,\cdots,M$ are drawn from the posterior distribution $p(\tilde{\mathbf{z}},\mathbf{z},\boldsymbol{\phi}_{\text{long}},\boldsymbol{\eta}|\tilde{\mathbf{y}},\mathbf{y},\mathbf{o},\mathrm{H}_a)$. Details of the MCMC sampling algorithm are given in the Appendix. Further, $\tilde{p}_i^{(m)} := p(\tilde{o}_i = 1|\boldsymbol{\eta}^{(m)},\tilde{\mathbf{z}}_i^{(m)})$ can be obtained from $\Phi((\boldsymbol{\eta}^{(m)})^T\mathbf{Z}_i^{(m)})$ as described in (4.3).

Then, for each validation data set, we focus on two performance measures: 1) the posterior predictive mean squared error (PMSE): PMSE $= M^{-1}\sum_{m=1}^{M}\text{PMSE}^{(m)}$, where $\text{PMSE}^{(m)} = \tilde{n}^{-1}\sum_{i=1}^{\tilde{n}}(\tilde{o}_i^{\text{pred},(m)} - \tilde{o}_i)^2$ with $\tilde{o}_i^{\text{pred},(m)}$ as a draw of a Bernoulli random variable with success probability $\tilde{p}_i^{(m)}$; 2) the area under the ROC curve AUC $= M^{-1}\sum_{m=1}^{M}\text{AUC}^{(m)}$ (i.e., test AUC), where $\text{AUC}^{(m)}$ is obtained based on $\tilde{p}_i^{(m)}$, $i = 1,\cdots,\tilde{n}$, using the approach as described in Section 4.

For each criteria a, we calculate the posterior PMSE and test AUC for the validation sample $\{(\tilde{\mathbf{y}}_i^{(s)},\tilde{o}_i^{(s)})\}_{i=1}^{\tilde{n}}$ fitted by model $\mathrm{H}_a^{(s)}$, $s = 1,\cdots,100$, and report the posterior mean PMSE and AUC as well as the 95% credible intervals based on 100 simulations. As a comparison, we also summarize the training AUC for the sample $\{(\mathbf{y}_i^{(s)},o_i^{(s)})\}_{i=1}^{\tilde{n}}$ that are reported in Table 4.6 for each criteria. We then repeat this procedure for the two simulation scenarios as described in previous section: when the data is generated from the joint LC and the joint MSRE model, respectively.

Table 4.7 shows the PMSE and test AUC for the new validation test samples, along with the AUC for the training samples, based on the selected models by DIC, LPML, modified AIC_1, BIC_1 and ICL_1, as well as WAIC_1, respectively. When the data is generated from the LC models with $K_D = K_C = 1$, the true AUC is always 0.5.

Fitting both LC and MSRE models leads to comparable predictive performances on the test sample, with the values of PMSE and test AUC varying little among different model selection criteria, and the estimates of training AUC all centering around the true value. As expected, the values of test AUC are slightly smaller than those of the training AUC.

When fitting joint MSRE models, the values of PMSE and test AUC vary little among different model selection criteria, with the 95% credible intervals of training AUC always covering the true AUC value for the test sample, regardless of the true structure in the primary outcome model. In particular, when the data are generated from LC models, fitting MSRE models leads to comparable predictive performances on the validation sample in comparison to those obtained by LC models. We note that this is due to choosing the setups of the LC and MSRE models to be compatible to each other. As expected, the values of test AUC are slightly smaller than those of the training AUC.

In contrast, when we study LC fitting, the values of training/test AUC and PMSE differ for different criteria and are affected by the true data generation mechanism. When fitting the LC generated data, DIC, LPML and WAIC frequently select the models that better classify the outcome with higher training AUC values than those chosen by other criteria. However, the values of PMSE and test AUC do not vary much by different criteria, with the training AUC centering around the true value 0.5, indicating that the better predictive performance of the models selected by LPML and WAIC is likely due to potential overfitting. When the data is generated from the MSRE models, the LC models selected by DIC, LPML and WAIC lead to a higher number and potentially outcome-informed mixture components, and consequently to the optimistically-biased training AUC relative to the test AUC. The joint LC models chosen by LPML, WAIC and DIC still lead to higher test AUC and lower MPSE, indicating somewhat more accurate prediction for validation sample than the models

chosen by other criteria. The modified AIC_1, BIC_1 and ICL_1, which tend to select the correct numbers of mixture components frequently, perform as expected for the training versus test samples, suggesting the validity of the compatibly chosen models. Since now the data is generated from the MSRE models, fitting LC models causes inferior predictive performances on the validation sample in comparison to those obtained by MSRE models. However, only focusing on the predictive performance on the training sample leads to an impression that LC models tend to classify the outcome much better.

4.6 Penn ovarian aging study revisited

In this section, we use the knowledge obtained from the simulation study to guide us to identify the plausible models in Penn ovarian aging study with the purpose of linking the longitudinal trajectory of Follicle Stimulating Hormone (FSH) and the occurrence of severe hot flashes during the study period. In our analysis, a total of 4,244 FSH values were observed for the final sample of 245 women, with a minimum of 3 and a maximum of 26 observations per woman. Of the 245 women without severe hot flash symptoms at baseline, 118 (48.2%) had experienced the outcome of interest, an indicator variable for experiencing severe hot flashes at least once during the study. We fit both joint LC and MSRE models, as described in Section 2, to the FSH trajectories and severe hot flash outcome data, adjusting for baseline log(BMI) and smoking indicator in the primary outcome model (2).

Table 4.8 reports the model selection statistics for the joint LC and MSRE model for the analysis of Penn ovarian aging study, with K_D and $K_C \in \{1, 2, 3\}$, respectively. For both joint LC and MSRE models, DIC, the modified AIC, BIC and ICL choose $K_D = 1$ and $K_C = 2$, while LPML and WAIC prefer models with more mixture components. Under the joint MSRE model, the AUC in the primary outcome model vary little, with the values being likely close to the truth. The AUC for the

joint LC model is elevated in models with more latent classes and WAIC and LPML tend to favor such LC models, likely due to their higher AUC values. This overall finding is not surprising and reflects some typical behaviors of these criteria as we have observed in the simulation study. In particular, the outcomes in our simulation study suggest that the notable difference between the AUC under the joint LC model chosen by WAIC, with $K_D = K_C = 3$, and the AUC under the MSRE model with any values of K_D and K_C implies potential over-fitting of the larger model. As the second best choice for both LC and MSRE models, the model with $K_D = K_C = 2$ is favored by DIC, the modified AIC and BIC. The true model is likely to be a model of $K_C = 2$ and $K_D = 1$ or, alternatively, $K_D = 2$ but with the two components closely overlapped with each other. However, as already indicated by AUC values, assuming $K_D = 2$ instead of $K_D = 1$ when $K_C = 2$ had very little impact on the predictive power for both joint LC and MSRE models; Chapter II also reported that the effect of mean profile is not significantly associated with the risk of severe hot flash in the primary outcome model using both models. Therefore, $K_D = 1$ and $K_C = 2$ is the most parsimonious choice for both joint LC and MSRE models.

Finally, in terms of choosing between LC and MSRE models assuming $K_D = 1$ and $K_C = 2$, DIC, the modified AIC, BIC and ICL do not choose the same model although these statistics from the two models are very similar, indicating similar fit to the FSH trajectories and severe hot flash outcome. The LC and MSRE models also share similar overall prediction performance; with ΔAUC being .04 and the corresponding credible set covering 0, $(-0.04, 0.11)$, as reported in Chapter II. The advantages of studying this data set using both modeling approaches with different evaluation criteria lie on a higher level of confidence that suitable models are used and that the results are not heavily determined by the assumed model.

4.7 Discussion

In this article, we studied several commonly-used model selection criteria in terms of choosing the numbers of mixture components in a joint modeling context, when both correctly and incorrectly assuming the association structure to link the longitudinal submodel and the primary outcome model. These criteria are all built upon Bayesian principles in the sense that they are evaluated over the entire posterior distribution rather than conditional on single point estimates. In particular, DIC and the modified AIC, BIC, ICL are based on deviance, while LPML is based on leave-one-out cross validated predictive density, which is shown by Watanabe (2010) to be asymptotically equivalent to WAIC.

In terms of choosing the numbers of mixture components, the performances of the modified AIC, BIC and ICL appear to be more reliable and predictable than other criteria when fitting joint LC and MSRE models with correctly assumed structure in the primary outcome model in the sense that, when the mixture components are easily separated, they frequently identify the correct numbers of mixture components while when the mixture components are fairly overlapped and hence difficult to separate, they frequently choose one instead of multiple mixture components. On the contrary, the numbers of mixture components chosen by DIC, LPML and WAIC are often more than the truth for the purpose of reaching improved prediction. In particular, WAIC and LPML tend to select the same models with higher AUC values relative to the models selected by other criteria.

For joint MSRE models, assuming different numbers of mixture components is not crucial in deciding the predictive performance as assessed by AUC values; however, for joint LC models, the predictive performance is closely related to the assumed numbers of mixture components. In particular, when the mixture components are difficult to separate and the true structure is MSRE, joint LC models tend to have high chance of artificially creating spurious mixture components to enhance AUC values

for the sample that is used to derive the model, giving the impression of much better prediction power by LC models than MSRE models. This phenomenon could cause some of our considered criteria to frequently choose incorrect numbers of mixture components and favor specific LC structure.

When this happens, our simulation studies suggest that new independent sample validation can be helpful to confirm whether the chosen models are suitable for the desired purposes. We find that the test AUC values for the validation sample based on the models chosen by WAIC and LPML also tend to be higher than the test AUC values based on the models chosen by other criteria. One needs to be cautious, though, that the overly optimistic training AUC values relative to the test AUC values contradicts the validity of the chosen models and therefore suggests that the seemingly high predictive power of the chosen LC models is unlikely to be the truth.

Based on our experience in the simulation study and the data example, we suggest fitting both LC and MSRE models and comparing the two sets of results with the same numbers of mixture components. When the outcome-informed artifact is present, the inference about the mixture components from the two models usually do not match and the AUC value obtained by LC model is often much higher than that by MSRE model. Otherwise, the two models assuming the same numbers of mixture components tend to lead to similar inference results, including similar scientific interpretation in the primary outcome and similar predictive power assessed by AUC. In addition to comparing the results by different model selection criteria, these rules can be helpful to guide us to choose the suitable models the best we can in practice.

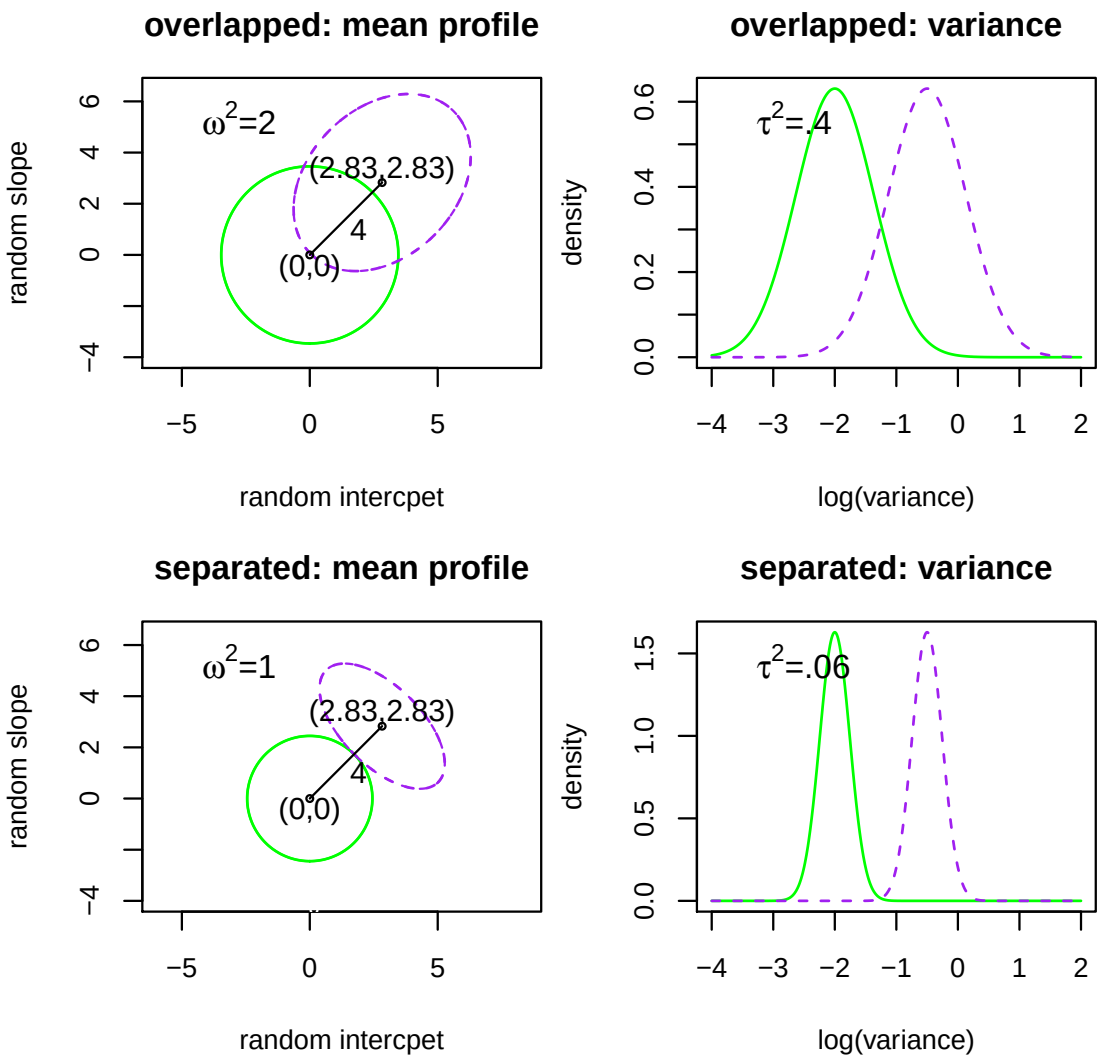

Figure 4.1: Simulation setup for the mean and variance profiles representing low versus high levels of separation. Left panels: 95% contour plots of the two components for the mean profiles; right panels: density plots of the two components for the variance profiles.

Table 4.1: Number of times each specified fitted model are selected by using the 10 criteria given in the first column of each sub-table. Observations are generated as described in Section 4.4.1 under the joint LC model with $K_D = K_C = 2$ and the (T_D, T_C) mixture structure for the mean and the variance profiles, respectively, where $T_D, T_C \in$ {separated, overlapped}. Scenarios (a)-(d) specify the data-generating mechanism. The fitted models consist of both LC and MSRE structures with $K_D = 1,2,3$ and $K_C = 1,2,3$. The mixing proportions are 50-50. The corresponding values of AUC were reported at the end of the table for each scenario. The "true AUC" is the AUC obtained when predictions are generated using the correct outcome model (true parameters and random effects/latent classes).

	fitting LC model (K_D,K_C)									**fitting MSRE model** (K_D,K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)
(a) true model=LC; (separated, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	0	26	0	0	73	1	0	0	0	0	36	1	0	62	1
LPML	0	0	0	0	8	47	0	5	40	0	11	11	0	15	33	0	7	23
AIC_1	0	0	0	0	90	1	0	9	0	0	0	0	0	90	1	0	8	1
AIC_2	0	0	0	0	88	1	0	11	0	0	0	0	0	85	1	0	13	1
BIC_1	0	0	0	0	99	1	0	0	0	0	0	0	0	98	1	0	1	0
BIC_2	0	0	0	0	99	1	0	0	0	0	0	0	0	97	1	0	2	0
ICL_1	0	0	0	1	98	1	0	0	0	0	0	0	0	98	1	0	1	0
ICL_2	0	0	0	1	98	1	0	0	0	0	0	0	0	98	1	0	1	0
$WAIC_1$	0	0	0	0	0	49	0	0	51	0	14	51	0	4	26	0	0	5
$WAIC_2$	0	0	0	0	1	68	0	1	30	0	31	24	0	16	22	0	2	5
Area under the ROC curves																		
true AUC=0.81	0.5	0.54	0.66	0.79	0.81	0.85	0.79	0.82	0.86	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
(b) true model=LC; (overlapped, overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	56	0	0	2	0	0	42	0	0	27	0	0	17	0	0	56	0	0
LPML	0	0	1	3	13	38	2	15	28	9	14	22	10	13	12	5	8	7
AIC_1	76	0	0	19	0	0	5	0	0	80	0	0	17	0	0	3	0	0
AIC_2	70	0	0	21	0	0	9	0	0	72	0	0	23	0	0	5	0	0
BIC_1	94	0	0	6	0	0	0	0	0	100	0	0	0	0	0	0	0	0
BIC_2	92	0	0	8	0	0	0	0	0	99	0	0	1	0	0	0	0	0
ICL_1	100	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
ICL_2	100	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
$WAIC_1$	0	0	0	0	2	60	1	0	37	32	19	27	9	8	1	0	1	3
$WAIC_2$	0	0	2	0	2	65	1	0	30	16	25	36	4	8	5	1	0	5
Area under the ROC curves																		
true AUC=0.81	0.5	0.63	0.72	0.78	0.85	0.88	0.78	0.85	0.88	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
(c) true model=LC; (overlapped, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	57	0	0	3	2	0	37	1	0	41	1	0	12	1	0	45	0
LPML	0	0	0	0	13	43	0	11	33	0	9	28	0	16	19	0	13	15
AIC_1	0	80	0	0	17	1	0	2	0	0	81	1	0	14	1	0	3	0
AIC_2	0	72	0	0	24	1	0	3	0	0	76	1	0	19	1	0	3	0
BIC_1	0	97	1	0	2	0	0	0	0	0	98	2	0	0	0	0	0	0
BIC_2	0	94	0	0	5	1	0	0	0	0	98	2	0	0	0	0	0	0
ICL_1	0	99	1	0	0	0	0	0	0	0	99	1	0	0	0	0	0	0
ICL_2	0	99	1	0	0	0	0	0	0	0	99	1	0	0	0	0	0	0
$WAIC_1$	0	0	1	0	2	66	0	0	31	0	22	45	0	0	20	0	3	10
$WAIC_2$	0	0	0	0	2	65	0	1	32	0	39	19	0	16	7	0	14	5
Area under the ROC curves																		
true AUC=0.82	0.5	0.55	0.66	0.80	0.83	0.87	0.80	0.83	0.87	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
(d) true model=LC; (separated, overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	11	0	0	88	1	0	1	0	0	16	0	0	82	1	0
LPML	0	0	0	0	17	37	2	11	33	6	8	2	13	14	16	8	16	17
AIC_1	0	0	0	85	0	0	15	0	0	1	0	0	77	1	0	21	0	0
AIC_2	0	0	0	73	1	0	26	0	0	1	0	0	71	0	0	27	1	0
BIC_1	0	0	0	100	0	0	0	0	0	1	0	0	98	0	0	1	0	0
BIC_2	0	0	0	100	0	0	0	0	0	1	0	0	98	0	0	1	0	0
ICL_1	0	0	0	100	0	0	0	0	0	2	0	0	98	0	0	0	0	0
ICL_2	0	0	0	100	0	0	0	0	0	1	0	0	98	0	0	1	0	0
$WAIC_1$	0	0	0	0	0	59	0	0	41	31	16	28	6	7	9	2	1	0
$WAIC_2$	0	0	0	0	0	74	0	0	26	13	17	29	6	10	17	1	2	5
Area under the ROC curves																		
true AUC=0.82	0.5	0.65	0.73	0.79	0.85	0.88	0.79	0.85	0.88	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81

Table 4.2: As in Table 4.1 but the mixing proportions are 80-20.

	fitting LC model (K_D,K_C)									**fitting MSRE model** (K_D,K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)
(a) true model=LC; (separated, separated) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	0	27	0	0	73	0	0	0	0	0	41	0	0	59	0
LPML	0	0	0	0	3	51	0	2	44	0	6	4	0	16	36	0	13	25
AIC_1	0	0	0	0	96	0	0	4	0	0	0	0	0	93	0	0	7	0
AIC_2	0	0	0	0	95	0	0	5	0	0	0	0	0	83	0	0	17	0
BIC_1	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
BIC_2	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
ICL_1	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
ICL_2	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
$WAIC_1$	0	0	0	0	0	60	0	0	40	0	18	67	0	4	7	0	0	4
$WAIC_2$	0	0	0	0	0	67	0	0	33	0	24	34	0	12	15	0	3	12
Area under the ROC curves																		
true AUC=0.72	0.5	0.55	0.70	0.68	0.72	0.81	0.69	0.72	0.81	0.73	0.74	0.74	0.73	0.74	0.74	0.73	0.74	0.74
(b) true model=LC; (overlapped, overlapped) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	12	1	0	17	0	0	70	0	0	19	0	0	22	1	1	57	0	0
LPML	0	3	15	1	5	38	1	8	29	18	17	15	13	7	10	12	4	4
AIC_1	64	1	0	28	0	0	7	0	0	75	1	1	21	0	0	2	0	0
AIC_2	54	1	0	35	0	0	10	0	0	71	1	1	25	0	0	2	0	0
BIC_1	95	1	0	4	0	0	0	0	0	97	1	1	1	0	0	0	0	0
BIC_2	93	1	0	6	0	0	0	0	0	97	1	1	1	0	0	0	0	0
ICL_1	100	0	0	0	0	0	0	0	0	99	0	0	1	0	0	0	0	0
ICL_2	100	0	0	0	0	0	0	0	0	99	0	0	1	0	0	0	0	0
$WAIC_1$	0	0	9	0	2	39	0	2	48	11	29	32	3	5	9	1	4	6
$WAIC_2$	0	1	17	0	2	44	0	1	35	7	31	30	6	7	10	1	2	6
Area under the ROC curves																		
true AUC=0.72	0.5	0.63	0.73	0.59	0.70	0.78	0.60	0.71	0.78	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
(c) true model=LC; (overlapped, separated) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	0	23	0	0	10	0	0	67	0	0	26	0	0	18	0	0	56	0
LPML	0	0	11	0	5	44	0	4	36	0	12	27	0	10	21	0	12	18
AIC_1	0	72	1	0	23	0	0	4	0	0	78	0	0	19	0	0	3	0
AIC_2	0	58	1	0	34	0	0	7	0	0	76	0	0	19	0	0	5	0
BIC_1	0	94	0	0	6	0	0	0	0	0	98	0	0	2	0	0	0	0
BIC_2	0	91	0	0	9	0	0	0	0	0	97	0	0	3	0	0	0	0
ICL_1	0	99	0	0	1	0	0	0	0	0	99	0	0	1	0	0	0	0
ICL_2	0	99	0	0	1	0	0	0	0	0	99	0	0	1	0	0	0	0
$WAIC_1$	0	0	6	0	0	43	0	0	51	0	16	46	0	3	23	0	4	8
$WAIC_2$	0	0	8	0	2	50	0	0	40	0	30	27	0	13	11	0	12	7
Area under the ROC curves																		
true AUC=0.72	0.5	0.54	0.70	0.60	0.64	0.76	0.60	0.65	0.77	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
(d) true model=LC; (separated, overlapped) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	23	0	0	76	1	0	0	0	0	24	0	0	74	2	0
LPML	0	0	0	2	7	40	1	6	44	0	1	2	23	18	18	20	14	4
AIC_1	0	0	0	90	1	0	9	0	0	0	0	0	89	0	0	11	0	0
AIC_2	0	0	0	87	1	0	12	0	0	0	0	0	85	2	0	13	0	0
BIC_1	0	0	0	99	0	0	1	0	0	0	0	0	100	0	0	0	0	0
BIC_2	0	0	0	99	0	0	1	0	0	0	0	0	99	0	0	1	0	0
ICL_1	0	0	0	99	0	0	1	0	0	0	0	0	100	0	0	0	0	0
ICL_2	0	0	0	99	0	0	1	0	0	0	0	0	100	0	0	0	0	0
$WAIC_1$	0	1	0	0	0	56	0	0	43	11	30	34	4	5	13	0	2	1
$WAIC_2$	0	0	0	0	0	68	0	1	31	9	20	24	13	9	17	3	4	1
Area under the ROC curves																		
true AUC=0.72	0.5	0.63	0.74	0.69	0.77	0.83	0.69	0.77	0.83	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72

Table 4.3: As in Table 4.1 but the observations are generated under the joint MSER model.

	fitting LC model (K_D,K_C)									**fitting MSRE model** (K_D,K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)
(a) true model=MSRE; (separated, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	1	0	0	25	4	0	65	5	0	0	0	0	37	1	0	60	2
LPML	0	3	14	0	8	47	0	3	25	0	6	7	0	21	32	0	11	23
AIC_1	0	1	0	0	57	5	0	34	3	0	0	0	0	88	3	0	9	0
AIC_2	0	1	0	0	55	5	0	36	3	0	0	0	0	81	3	0	16	0
BIC_1	0	3	0	0	73	4	0	19	1	0	0	0	0	94	3	0	3	0
BIC_2	0	2	0	0	72	4	0	20	2	0	0	0	0	94	3	0	3	0
ICL_1	0	5	0	0	74	2	0	19	0	0	1	0	0	97	1	0	1	0
ICL_2	0	5	0	0	71	3	0	21	0	0	1	0	0	96	1	0	2	0
$WAIC_1$	0	0	18	0	7	45	0	1	29	0	24	45	0	8	18	0	1	4
$WAIC_2$	0	0	31	0	8	48	0	1	12	0	37	21	0	19	13	0	4	6
Area under the ROC curves																		
true AUC=0.82	0.5	0.64	0.72	0.52	0.70	0.80	0.54	0.70	0.77	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
(b) true model=MSRE; (overlapped, overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	15	0	0	11	3	0	56	13	2	23	0	0	17	0	0	60	0	0
LPML	0	1	2	0	32	30	0	20	15	13	24	17	8	7	11	5	6	9
AIC_1	74	0	0	15	8	1	2	0	0	84	0	0	13	0	0	3	0	0
AIC_2	63	0	0	18	13	1	5	0	0	73	0	0	21	0	0	6	0	0
BIC_1	98	0	0	0	2	0	0	0	0	100	0	0	0	0	0	0	0	0
BIC_2	95	0	0	1	4	0	0	0	0	99	0	0	1	0	0	0	0	0
ICL_1	99	0	0	0	1	0	0	0	0	100	0	0	0	0	0	0	0	0
ICL_2	99	0	0	0	1	0	0	0	0	100	0	0	0	0	0	0	0	0
$WAIC_1$	0	1	3	0	50	18	0	20	8	23	32	26	3	3	8	3	0	2
$WAIC_2$	0	1	4	0	39	27	0	17	12	7	30	38	1	5	14	1	1	3
Area under the ROC curves																		
true AUC=0.85	0.5	0.74	0.79	0.52	0.91	0.92	0.53	0.92	0.93	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
(c) true model=MSRE; (overlapped, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	65	3	0	3	2	0	18	9	0	38	2	0	16	2	0	42	0
LPML	0	0	3	0	10	44	0	8	35	0	19	31	0	7	21	0	8	14
AIC_1	0	83	3	0	7	6	0	0	1	0	81	4	0	10	0	0	5	0
AIC_2	0	77	3	0	10	4	0	1	5	0	79	4	0	12	0	0	5	0
BIC_1	0	95	2	0	0	3	0	0	0	0	97	3	0	0	0	0	0	0
BIC_2	0	94	2	0	0	4	0	0	0	0	96	4	0	0	0	0	0	0
ICL_1	0	98	1	0	0	1	0	0	0	0	98	2	0	0	0	0	0	0
ICL_2	0	98	1	0	0	1	0	0	0	0	97	3	0	0	0	0	0	0
$WAIC_1$	0	0	0	0	0	54	0	3	43	0	26	54	0	3	9	0	2	6
$WAIC_2$	0	0	0	0	2	55	0	4	39	0	42	27	0	9	10	0	6	6
Area under the ROC curves																		
true AUC=0.82	0.5	0.63	0.71	0.52	0.83	0.88	0.53	0.84	0.89	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
(d) true model=MSRE; (separated,overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	16	0	0	83	1	0	0	0	0	19	1	0	80	0	0
LPML	0	2	8	0	30	32	0	9	19	5	8	10	12	18	21	7	7	12
AIC_1	0	0	0	79	0	0	21	0	0	0	0	0	78	1	0	21	0	0
AIC_2	0	0	0	71	0	0	29	0	0	0	0	0	68	1	0	31	0	0
BIC_1	2	0	0	94	1	0	3	0	0	1	0	0	91	1	0	7	0	0
BIC_2	1	0	0	94	1	0	4	0	0	1	0	0	91	1	0	7	0	0
ICL_1	2	0	0	95	0	0	3	0	0	2	0	0	91	0	0	7	0	0
ICL_2	2	0	0	95	0	0	3	0	0	2	0	0	91	0	0	7	0	0
$WAIC_1$	0	4	10	0	37	22	0	7	20	13	21	41	8	9	5	0	3	0
$WAIC_2$	0	2	19	0	33	23	0	6	17	7	17	36	9	12	14	0	3	2
Area under the ROC curves																		
true AUC=0.85	0.5	0.75	0.80	0.53	0.85	0.88	0.54	0.81	0.86	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85

Table 4.4: As in Table 4.2 but the observations are generated under the joint MSER model.

	fitting LC model (K_D, K_C)									**fitting MSRE model** (K_D, K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)
(a) true model=MSRE; (separated, separated) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	0	18	0	0	82	0	0	0	0	0	25	0	0	74	1
LPML	0	1	12	0	0	36	0	1	50	0	2	5	0	16	37	0	10	30
AIC_1	0	0	0	0	95	0	0	5	0	0	0	0	0	89	1	0	10	0
AIC_2	0	0	0	0	86	0	0	14	0	0	0	0	0	87	1	0	12	0
BIC_1	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
BIC_2	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
ICL_1	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
ICL_2	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0
$WAIC_1$	0	0	28	0	0	37	0	0	35	0	25	53	0	5	9	0	1	7
$WAIC_2$	0	1	47	0	0	29	0	0	23	0	25	28	0	13	15	0	8	11
Area under the ROC curves																		
true AUC=0.77	0.5	0.59	0.74	0.52	0.60	0.77	0.52	0.61	0.77	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
(b) true model=MSRE; (overlapped, overlapped) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	18	1	0	21	0	0	57	3	0	21	0	0	22	0	0	55	2	0
LPML	0	3	17	0	13	23	0	16	28	16	17	7	10	5	13	14	14	4
AIC_1	83	2	0	12	1	0	2	0	0	78	2	0	18	0	0	2	0	0
AIC_2	82	1	0	12	2	0	3	0	0	76	2	0	20	0	0	2	0	0
BIC_1	99	1	0	0	0	0	0	0	0	98	1	0	1	0	0	0	0	0
BIC_2	99	1	0	0	0	0	0	0	0	98	1	0	1	0	0	0	0	0
ICL_1	100	0	0	0	0	0	0	0	0	99	0	0	1	0	0	0	0	0
ICL_2	100	0	0	0	0	0	0	0	0	99	0	0	1	0	0	0	0	0
$WAIC_1$	0	5	12	0	19	21	0	17	26	10	21	41	0	7	14	1	1	5
$WAIC_2$	0	4	22	0	8	29	0	10	27	16	23	33	1	4	10	3	5	5
Area under the ROC curves																		
true AUC=0.82	0.5	0.77	0.82	0.51	0.81	0.85	0.51	0.82	0.85	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
(c) true model=MSRE; (overlapped, separated) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	0	19	1	0	28	0	0	51	1	0	27	0	0	23	0	0	50	0
LPML	0	1	39	0	2	27	0	0	31	0	11	35	0	6	22	0	6	20
AIC_1	0	85	2	0	9	1	0	3	0	0	80	0	0	15	0	0	5	0
AIC_2	0	80	2	0	13	1	0	4	0	0	73	0	0	20	0	0	7	0
BIC_1	0	98	2	0	0	0	0	0	0	0	99	0	0	1	0	0	0	0
BIC_2	0	98	2	0	0	0	0	0	0	0	97	0	0	3	0	0	0	0
ICL_1	0	100	0	0	0	0	0	0	0	0	99	0	0	1	0	0	0	0
ICL_2	0	100	0	0	0	0	0	0	0	0	99	0	0	1	0	0	0	0
$WAIC_1$	0	0	27	0	0	30	0	0	43	0	19	51	0	6	13	0	4	7
$WAIC_2$	0	0	38	0	0	30	0	0	32	0	22	36	0	11	16	0	5	10
Area under the ROC curves																		
true AUC=0.75	0.5	0.59	0.74	0.51	0.61	0.76	0.51	0.62	0.76	0.75	0.74	0.75	0.75	0.74	0.75	0.75	0.74	0.75
(d) true model=MSRE; (separated,overlapped) mixtures; 20:80 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	22	0	0	77	1	0	0	0	0	25	0	0	75	0	0
LPML	0	4	12	0	11	38	0	5	30	2	2	1	20	21	20	11	14	9
AIC_1	0	0	0	92	1	0	7	0	0	0	0	0	90	0	0	10	0	0
AIC_2	0	0	0	92	1	0	7	0	0	0	0	0	84	0	0	16	0	0
BIC_1	0	0	0	99	1	0	0	0	0	0	0	0	100	0	0	0	0	0
BIC_2	0	0	0	99	1	0	0	0	0	0	0	0	100	0	0	0	0	0
ICL_1	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0	0
ICL_2	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0	0	0
$WAIC_1$	0	10	48	0	2	16	0	1	23	4	22	42	4	9	13	0	2	4
$WAIC_2$	0	7	53	0	2	24	0	2	12	5	13	24	7	12	23	1	7	8
Area under the ROC curves																		
true AUC=0.84	0.5	0.74	0.80	0.52	0.75	0.82	0.53	0.76	0.82	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84

Table 4.5: Number of times correct association structure identified by model selection criteria when comparing joint LC versus joint MSRE model with $K_D = K_C = 2$. Observations are generated as described in Section 4.4.1 under both the joint LC and MSRE model with $K_D = K_C = 2$ and the (T_D, T_C) mixture structure for the mean and the variance profiles, respectively, where $T_D, T_C \in$ {separated, overlapped}. Scenarios (a)-(d) specify the data-generating mechanism.

(a) true model=LC

AUC			No. of times LC model selected from 100 simulations									
Truth	LC	MSRE	DIC	LPML	AIC_1	AIC_2	BIC_1	BIC_2	ICL_1	ICL_2	$WAIC_1$	$WAIC_2$
(separated, separated) mixtures; 50:50 proportions												
0.81	0.81	0.81	65	81	76	73	94	92	97	96	91	89
(overlapped, overlapped) mixtures; 50:50 proportions												
0.81	0.85	0.77	12	90	12	12	15	15	17	19	98	98
(overlapped, separated) mixtures; 50:50 proportions												
0.82	0.83	0.78	10	88	12	12	14	14	13	13	96	90
(separated,overlapped) mixtures; 50:50 proportions												
0.82	0.85	0.81	64	97	67	66	71	70	77	77	98	96
(separated, separated) mixtures; 20:80 proportions												
0.72	0.72	0.74	29	67	40	37	68	68	70	68	90	85
(overlapped, overlapped) mixtures; 20:80 proportions												
0.72	0.70	0.69	33	74	39	38	45	47	51	51	81	69
(overlapped, separated) mixtures; 20:80 proportions												
0.72	0.64	0.70	2	47	4	4	10	11	9	10	58	52
(separated,overlapped) mixtures; 20:80 proportions												
0.72	0.77	0.72	60	92	63	62	74	74	74	72	98	97

(b) true model=MSRE

AUC			No. of times MSRE model selected from 100 simulations									
Truth	LC	MSRE	DIC	LPML	AIC_1	AIC_2	BIC_1	BIC_2	ICL_1	ICL_2	$WAIC_1$	$WAIC_2$
(separated, separated) mixtures; 50:50 proportions												
0.82	0.70	0.81	100	87	100	100	100	100	100	100	83	90
(overlapped, overlapped) mixtures; 50:50 proportions												
0.85	0.91	0.85	39	34	37	34	30	29	13	13	18	30
(overlapped, separated) mixtures; 50:50 proportions												
0.82	0.83	0.82	97	46	97	97	93	93	91	92	35	50
(separated,overlapped) mixtures; 50:50 proportions												
0.85	0.85	0.85	82	67	80	80	73	73	29	29	60	68
(separated, separated) mixtures; 20:80 proportions												
0.77	0.60	0.76	97	99	95	95	93	93	92	93	100	100
(overlapped, overlapped) mixtures; 20:80 proportions												
0.82	0.81	0.82	58	86	53	54	51	51	30	30	50	86
(overlapped, separated) mixtures; 20:80 proportions												
0.75	0.61	0.74	96	95	96	96	91	93	91	92	98	97
(separated,overlapped) mixtures; 20:80 proportions												
0.84	0.75	0.84	81	100	78	79	76	76	50	50	91	100

Table 4.6: Number of times each specified fitted model are selected by using the 10 criteria given in the first column of each sub-table. Observations are generated as described in Section 4.4.1 under the joint LC model for scenario (a), and MSRE model for scenario (b), with $K_D = K_C = 1$. The fitted models consist of both LC and MSRE strutures with $K_D = 1,2$, and $K_C = 1,2$. The mixing proportions are 50-50. The corresponding values of AUC were reported at the end of the table for each scenario.

	fitting LC model (K_D, K_C)				**fitting MSRE model** (K_D, K_C)			
	(1,1)	(1,2)	(2,1)	(2,2)	(1,1)	(1,2)	(2,1)	(2,2)
				(a) True model=LC				
				No. of times selected				
DIC	9	0	91	0	10	0	90	0
LPML	0	58	0	42	30	23	22	25
AIC_1	92	0	8	0	81	0	19	0
AIC_2	76	0	24	0	72	0	28	0
BIC_1	100	0	0	0	100	0	0	0
BIC_2	100	0	0	0	100	0	0	0
ICL_1	100	0	0	0	100	0	0	0
ICL_2	100	0	0	0	100	0	0	0
$WAIC_1$	0	46	0	54	43	41	9	7
$WAIC_2$	0	56	0	44	27	49	10	14
				Area under the ROC curves				
true AUC=0.5	0.5	0.68	0.51	0.69	0.57	0.57	0.57	0.57
				(b) True model=MSRE				
				No. of times selected				
DIC	7	0	92	1	14	0	86	0
LPML	0	0	50	50	31	28	19	22
AIC_1	15	0	84	1	84	0	16	0
AIC_2	10	0	89	1	73	0	27	0
BIC_1	42	0	58	0	100	0	0	0
BIC_2	34	0	66	0	100	0	0	0
ICL_1	69	0	31	0	100	0	0	0
ICL_2	59	0	41	0	100	0	0	0
$WAIC_1$	0	1	66	33	40	33	11	16
$WAIC_2$	0	1	58	41	35	42	6	17
				Area under the ROC curves				
true AUC=0.88	0.5	0.69	**0**.95	**0**.97	0.88	0.88	0.88	0.88

Table 4.7: Values of AUC for independent validation sample ($\tilde{n} = 50$) based on the models selected by DIC, LPML, AIC_1, BIC_1, ICL_1 and $WAIC_1$. Both the new validation sample and the original sample to build the models are generated as described in Section 4.4.1 under the the joint LC model, for scenario (a), and MSRE model, for scenario (b), with $K_D = K_C = 1$.

	fitting joint LC			**fitting joint MSRE**		
	PMSE (95% CI)	test AUC (95% CI)	training AUC (95% CI)	PMSE (95% CI)	test AUC (95% CI)	training AUC (95% CI)
	(a) true model=LC (true test AUC=0.50)					
DIC	0.50 (0.48,0.51)	0.50 (0.49,0.52)	0.51 (0.50,0.53)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.63)
LPML	0.50 (0.48,0.51)	0.50 (0.49,0.52)	0.68 (0.64,0.74)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.62)
AIC_1	0.50 (0.48,0.51)	0.50 (0.50,0.50)	0.50 (0.50,0.53)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.62)
BIC_1	0.50 (0.48,0.51)	0.50 (0.50,0.50)	0.50 (0.50,0.50)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.63)
ICL_1	0.50 (0.48,0.51)	0.50 (0.50,0.50)	0.50 (0.50,0.50)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.63)
$WAIC_1$	0.50 (0.48,0.51)	0.50 (0.48,0.52)	0.69 (0.65,0.74)	0.5 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.62)
	(b) true model=MSRE (true test AUC=0.88)					
DIC	0.31 (0.24,0.44)	0.68 (0.50,0.76)	0.93 (0.50,0.98)	0.28 (0.22,0.35)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
LPML	0.3 (0.24,0.36)	0.70 (0.63,0.76)	0.96 (0.93,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
AIC_1	0.32 (0.24,0.47)	0.66 (0.50,0.76)	0.89 (0.50,0.98)	0.28 (0.22,0.35)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
BIC_1	0.36 (0.24,0.49)	0.61 (0.50,0.75)	0.77 (0.50,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
ICL_1	0.4 (0.25,0.50)	0.56 (0.50,0.75)	0.65 (0.50,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
$WAIC_1$	0.3 (0.24,0.37)	0.69 (0.62,0.76)	0.96 (0.93,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)

Table 4.8: Model comparison statistics from different joint models for the analysis of Penn ovarian aging data. The ten selection criteria are given in the first column of the table. The top and bottom panels correspond to the scenarios of fitting with the joint LC and MSRE models, respectively. Best fit model is given by boldface.

	joint model (K_D, K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
					fitting joint LC model				
DIC	6930.1	**6857.9**	7024.4	6924.5	6860.5	7020.9	6973.7	6899.9	7076.3
LPML	−3794.1	−3779.6	−3772.9	−3779.8	−3763.1	−3757.3	−3780.3	−3761.5	**−3747.3**
AIC_1	6941.5	**6875.7**	7043.6	6943.7	6882.6	7051.2	7001.4	6934.9	7118.5
AIC_2	6932.9	**6864.9**	7030.9	6928.9	6865.3	7031.5	6981.1	6911.9	7092.7
BIC_1	6976.5	**6921.3**	7099.7	7003.3	6956.1	7138.7	7085.4	7036.4	7237.5
BIC_2	6967.9	**6910.5**	7086.9	6988.5	6938.8	7119.0	7065.1	7013.4	7211.8
ICL_1	6976.5	**6949.5**	7261.9	7131.7	7102.3	7392.5	7265.8	7222.4	7546.7
ICL_2	6967.9	**6938.4**	7249.6	7117.3	7084.9	7372.8	7247.0	7200.0	7522.2
$WAIC_1$	7078.8	7025.2	6986.4	7054.3	6997.1	6948.9	7045.9	6964.9	**6938.7**
$WAIC_2$	7343.8	7305.0	7279.1	7322.6	7287.9	7258.8	7322.0	7268.3	**7251.3**
AUC	0.55	0.65	0.78	0.59	0.67	0.83	0.69	0.79	0.86
					fitting joint MSRE model				
DIC	6907.9	**6852.9**	7044.0	6910.8	6862.9	7060.6	6991.4	6938.4	7128.3
LPML	−3788.2	−3784.0	−3776.9	−3773.2	−3768.8	**−3764.4**	−3778.3	−3769.9	−3768.3
AIC_1	6927.1	**6874.2**	7067.5	6936.5	6890.3	7090.3	7023.7	6972.4	7164.9
AIC_2	6916.3	**6861.5**	7052.9	6920.2	6871.6	7070.0	7001.9	6948.4	7139.5
BIC_1	6979.6	**6933.7**	7134.0	7010.0	6970.8	7177.8	7118.2	7073.9	7273.4
BIC_2	6968.8	**6921.0**	7119.4	6993.7	6952.2	7157.5	7096.4	7049.9	7248.0
ICL_1	6979.6	**6964.3**	7355.7	7143.2	7135.8	7538.7	7334.1	7318.6	7711.7
ICL_2	6968.8	**6951.3**	7342.0	7127.2	7117.1	7519.9	7314.9	7296.6	7689.6
$WAIC_1$	7062.6	7032.6	7033.7	7043.9	7012.4	**7011.6**	7046.1	7015.6	7015.8
$WAIC_2$	7328.6	7309.6	7310.5	7309.1	7293.3	**7291.5**	7314.5	7301.5	7298.9
AUC	0.68	0.69	0.69	0.67	0.68	0.68	0.67	0.68	0.68

CHAPTER V

Discussion

In this dissertation, we have extended the joint modeling framework to incorporate both mean and variability features from longitudinal data to predict cross-sectional outcomes. In particular, we focus on studying the association between FSH trajectories and severity of menopausal hot flashes using the data from the Penn Ovarian Aging study.

In Chapter II, we use latent growth curve features derived from each individual's linear deviation from a population trend as predictors of a binary health outcome, i.e., occurrence of moderate-to-severe hot flashes. Moderate to severe hot flashes are defined as those rated 2 or 3 by the participants, when asked whether hot flashes or night sweats occurred in the past month and the severity, rated as 0 (none), 1 (mild), 2 (moderate), 3 (severe) at each follow-up. A key contribution from this work is to make contrast of pros and cons of the use of latent classes and multiple shared random effects in joint mixture modeling when the data-generating mechanism both matches and differs from the fitting one. In Chapter III, we consider a robust semi-parametric model that uses Bayesian penalized B splines, mixture distributions and Student t-error assumptions to improve model flexibility and robustness to potential outlying observations in the FSH hormone trajectories to predict an ordinal menopausal symptom severity, 0 (if self-rated severity score is < 2 throughout the follow-up), 1

(self-rated severity score is once 2 or once 3 but that occurred before age 40) and 2 (if self-rated severity score is once 3 after age 40).

Although the two proposed methodologies seek trajectory features from two different perspectives, both indicate the heterogeneous nature in the mean and the residual variability of the FSH trajectories for the women in the study, revealing four distinct patterns of change in the FSH profiles. Specifically, the mean FSH profile has two patterns featuring 1) women who tend to have an earlier increasing time frame in the FSH trajectories along with higher FSH values in comparison to 2) women who tend to have increasing FSH levels at later ages and hence lower FSH values; the residual variation in the FSH trajectories can also be clustered into low versus high categories.

However, since the two mean profile patterns are mainly due to relatively early or late ages when women experience elevated FSH levels, considering FSH patterns in relation to the Final Menstrual Period (FMP) might eliminate such differences due to different timing of FSH change. To further compare and contrast the differences in the FSH trajectory clustering between the use of years relative to FMP versus that of chronological age, we refit the model considered in Chapter III for a subgroup of 152 women for whom the FMP was observed, using the FSH levels in relation to both chronological age and the years prior to or post FMP to predict the ordinal severity of hot flashes.

5.1 FSH patterns in relation to Final Menstrual Period (FMP)

Figure 5.1 (a) and (b) show the estimated population-mean trends of FSH levels in relation to the years relative to FMP and chronological age, respectively. Both plots show similar patterns in the population trend, reflecting the typical FSH change patterns for women in the transition to menopause. FSH is relatively flat prior to the menopause transition, and then has an increasing period during the menopause

transition with acceleration and deceleration patterns (Sowers et al., 2008). As shown in Figure 5.2, when clustering the mean FSH trajectories in relation to FMP, the majority of the subjects (93% of trajectories) follow the typical change pattern in FSH levels as observed in the population trend, while 7% of trajectories form a separate cluster with relatively constant FSH levels over the observation period. That is, there is a notable reduction on the distinctions between the two clusters of subjects due to the early versus late ages in experiencing elevated FSH levels. This finding is not surprising since all the women included in the analysis experience FMP and such typical patterns of change in their FSH levels during the transition period are anticipated. Note that the estimated decreasing trend in the FSH trajectories 5 to 10 years following FMP is likely due to the sparsity of observations. Figure 5.1 (a) and (b) both show evidence of the existence of subgroups of "low" versus "high" residual variation; however, the large sampling variability in estimating the means of the variance class is likely due to the dropping of 35% women without experiencing FMP.

To obtain additional insight into the trajectory patterns in each of the mean profile class, we look at the individual level fitted FSH curves. Figure 5.3 (a) and (b) present several chosen fitted FSH trajectories by fitting the models using the FMP and chronological age, respectively. Note the trajectories of the same colors in both (a) and (b) belong to the same women and all the class membership assignments are based on the maximum of the posterior mean assignment probabilities of cluster membership. In particular, the green curves shown in Figure 5.3 (a) are the total six curves that are assigned to the minor mean class when fitting the model using the FMP; they all show relatively constant trends over the years around FMP, in comparison to the curves assigned to the major class that show the typical increasing and slightly plateau pattern over the years around FMP. In contrast, when using chronological age, as shown in Figure 5.3 (b), these six green curves are all assigned to the major class that features increasing FSH period at relatively late ages with rel-

atively stable FSH levels during the period when women in the minor class already experience elevated FSH levels. For the randomly selected curves shown in both red and blue colors in Figure 5.3 (a), they show typical FSH change patterns in that their FSH levels started to rise 5 to 10 years prior to the FMP and started to plateau around 2 years following the FMP; however, when using chronological age, since they differ in the ages of experiencing elevated FSH levels, they are assigned into the "late rising" versus "early rising" FSH mean cluster, respectively. Finally, Figure 5.3 (a) and (b) suggest that although these trajectories are clustered differently because of different timing of reaching FMP, considering either FMP or chronological age leads to almost identical trajectory fitting except some horizontal shift in either the axis of years in relation to FMP or chronological age.

However, when it comes to estimating the time-varying effect of cumulative changes in the mean profiles, $\theta_1(t)$, considering FSH levels in relation to FMP versus chronological age leads to different results as shown in Figure 5.4. When focusing on the time window between 7 years prior to and 3 years following FMP, it suggests that higher values of change in the mean FSH levels, $\mu_i'(t)$, in the window of 1 to 3 years following FMP seem to be associated with the levels of severity of hot flash; however, such signal is masked by considering the FSH levels in relation to chronological age in the range of $[45,55]$. Nevertheless, since our approaches allow the flexibility in choosing the study time period, equivalent findings, as that of using FMP adjusted age, were identified in Figure 3.5 (b), when we focus on studying the patterns in relation to chronological age in the range of $[50,55]$. Obviously, one should consider FMP adjusted age when the relevant measurements become available. However, the FMP ages are not always observed and the reasons why they are not available usually include hysterectomy prior to natural menopause, drop-out and using hormone therapy for hot flash.

In this sub-study, 35% of the subjects without FMP were tentatively excluded.

A more scientifically sound and meaningful investigation may be conducted when treating the FMP as a variable with missing observations and considering the unobserved FMP in women using hormone therapy for hot flash as informative missing values. The analysis accounting for this feature no doubt will bring in additional variation. Additional efforts would also need to be made in the future to carefully classify the reasons behind missing FMP. Based on the currently available information, our proposed approaches using chronological age focusing on an alternative age period suggest the potential that the changing-rate of FSH trajectories in the years immediately prior to and post menopause could be of interest, at least for a sub-group of women.

5.2 Future work

There are several directions of future investigation. Since our proposed methodologies incorporate both mean and variability features in the longitudinal trajectories, the extension to model time-to-event outcomes may help improve the accuracy of predictions that are only based on the mean features. Extending the idea of using latent features in longitudinal trajectories as functional predictors to predict the risk for an event may also be desirable. For example, we can relate the latent features $g(t)$ ($g(t)$ can be $\mu(t)$ or $\mu'(t)$) of the longitudinal trajectories to the hazard of an event at time t through $h(t) = h_0(t)\exp\left(\gamma^T \mathbf{z} + \int_0^t g(t)\theta(t)dt\right)$, where $\mathbf{z}$ includes baseline covariates. In contrast to the standard Cox models with time-dependent covariates, this extension allows different "weighting" assigned to all past values of the longitudinal features and is therefore more desirable because depending on the latent period of the event of interest, the values within the time window of $[0,t]$ may contribute to the event risks differently. Given the prevalence of missing data in longitudinal studies, another direction of future work will be to extend the proposed models to incorporate not missing at random (NMAR) missing mechanisms and assess their impact on

statistical inferences. One approach is to add another hierarchical layer to our joint models by modeling the probability of missingness to depend on unobserved data. Further, we can consider modeling the within-subject heteroscedasticity, that is, to allow subject-specific residual variances change over time. Extensions to incorporate time-varying variance structures will have practical applications in certain settings. For example, some scientific evidence suggests that the residual variabilities of reproductive hormone trajectories relative to the underlying mean trends closer to menopause tend to be of higher degree. Developing joint models that simultaneously consider the time-varying mean and residual variabilities of longitudinal trajectories would provide tools for our collaborators to address this important research hypothesis. In particular, the time-varying residual variance can also be considered as a potential functional predictor to relate to an outcome of interest. Finally, it would be of interests to investigate how to use multiple longitudinal trajectories to relate to either one or multiple health outcomes of interest in joint modeling settings to borrow information from either correlated longitudinal trajectories or correlated multiple health outcomes to eventually improve the prediction accuracy of the outcome of interest.

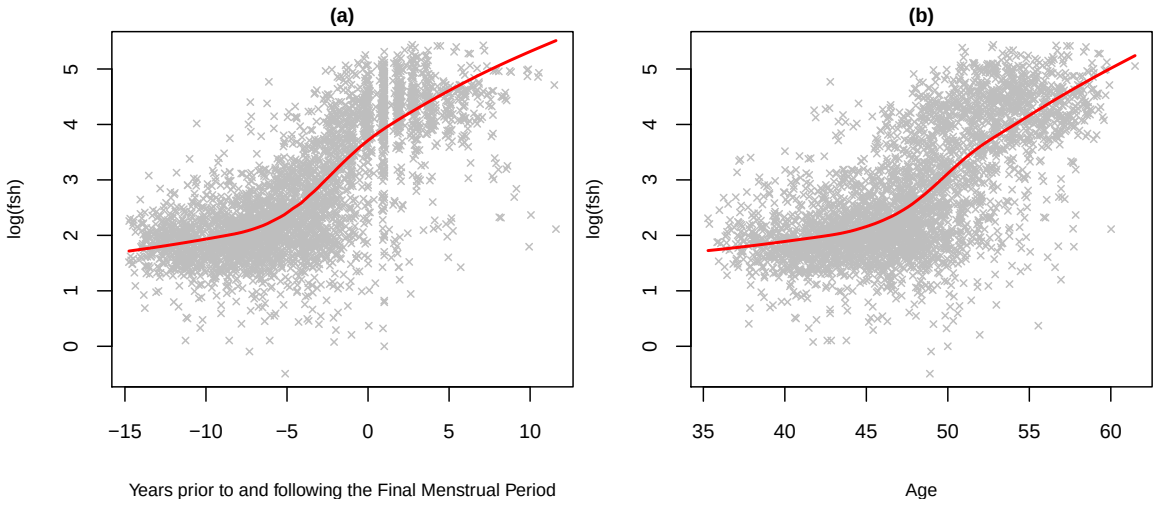

Figure 5.1: Estimated population longitudinal trend by lowess: (a) in relation to years before and after the Final Menstrual Period (FMP) and (b) according to chronological age.

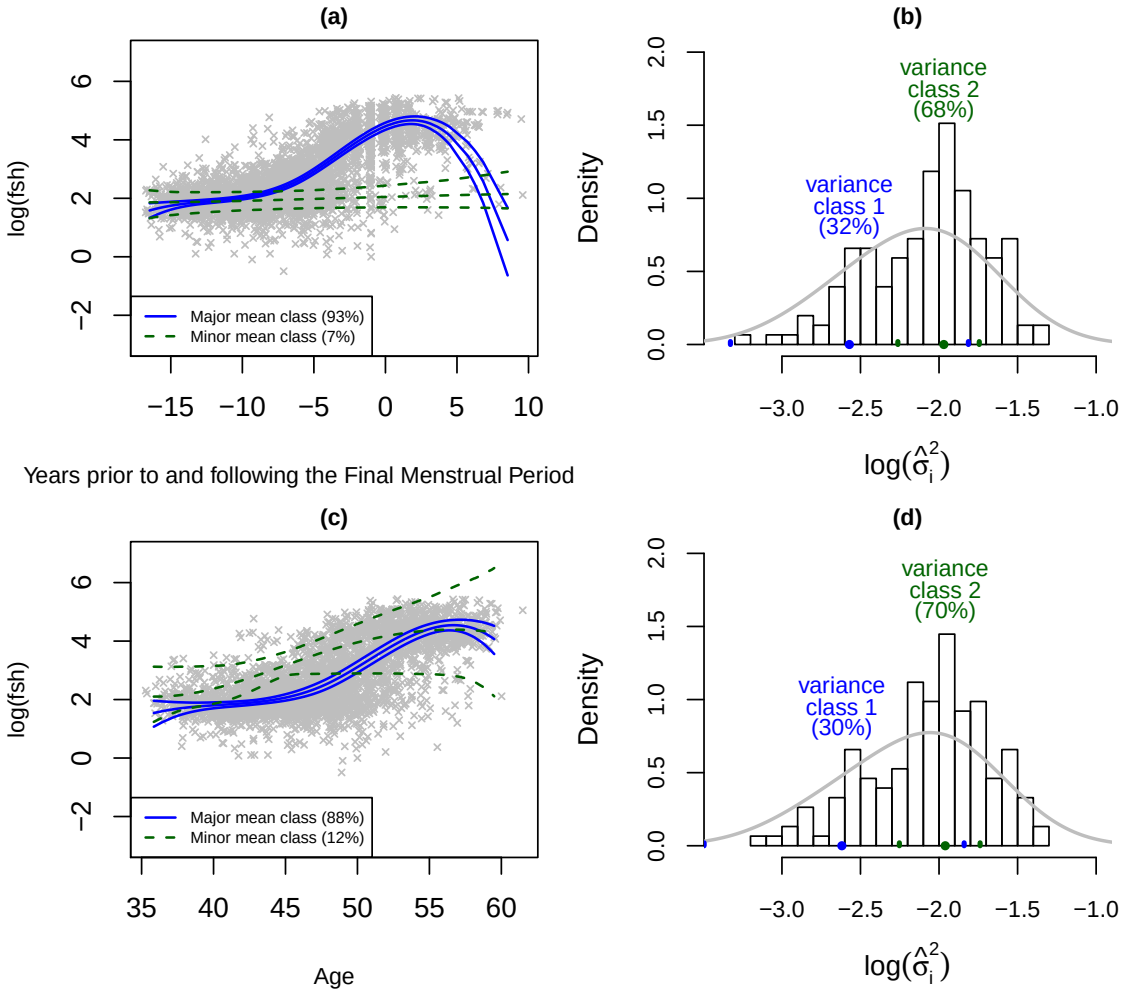

Figure 5.2: **longitudinal profiles:** (a) and (b) in relation to years before and after the Final Menstrual Period (FMP); (c) and (d) according to chronological age.

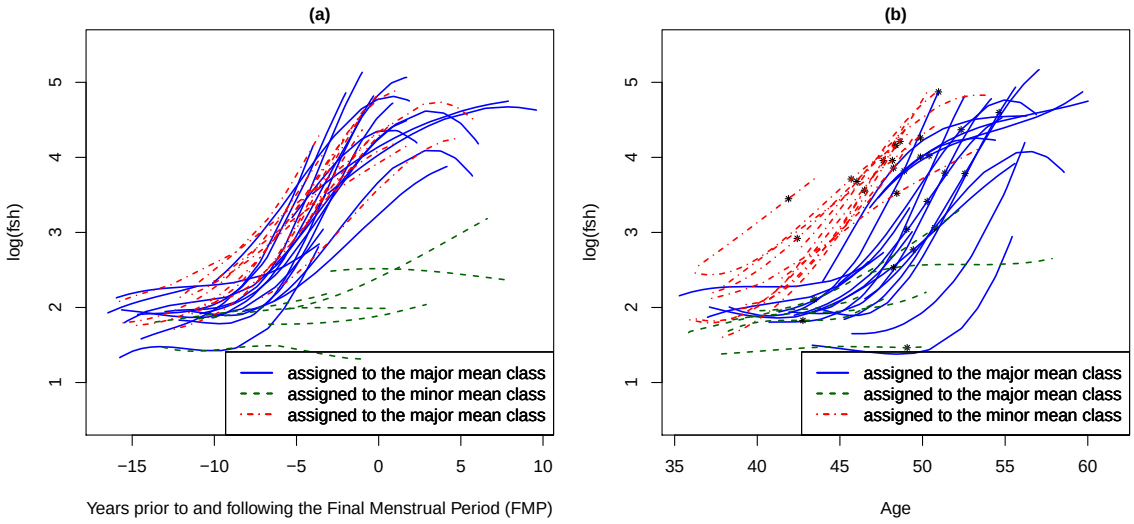

Figure 5.3: Individual FSH trajectories that are assigned to the minor and major mean profile classes when fitting the models using (a) years before and after the Final Menstrual Period (FMP) and (b) chronological age, respectively. The curves of the same colors in (a) and (b) belong to the same women. The black points in (b) indicate the fitted FSH values at the ages of FMP. The red trajectories are the predicted FSH for women who were in the minor mean classification according to chronological age but in the major mean classification in relation to time to FMP, while the green trajectories are the predicted FSH for women who were in the major mean classification according to chronological age but in the minor mean classification in relation to time to the FMP.

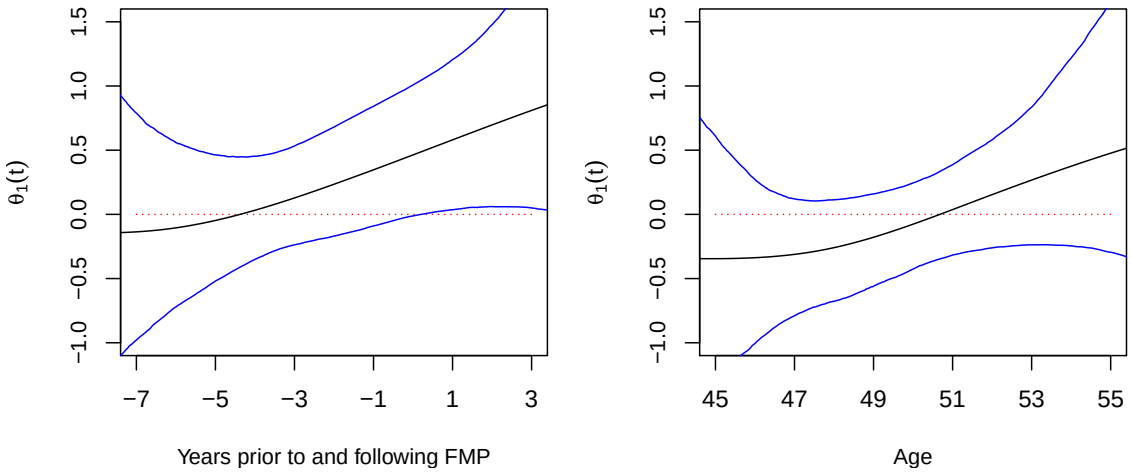

Figure 5.4: **Functional coefficient function** $\theta_1(t)$ when $\mu_i'(t)$ is functional predictor with when (a) in relation to years before and after the Final Menstrual Period (FMP) and time window is $[-7,3]$, and (b) according to chronological age and time wondow is $[45,55]$.

APPENDICES

APPENDIX A

Computation Details in Chapter II

Posterior computations for the joint LC model

(1) update for longitudinal submodel

- **update** the mean profile class memberships D_i, $i = 1,...,n$: the full conditional posterior distribution $[D_i|\cdot] \sim \text{Multinomial}(\tilde{\pi}^D_{i1},...,\tilde{\pi}^D_{iK_D})$, where

$$\tilde{\pi}^D_{id} = \Pr(D_i = d|.) = \frac{\pi^D_d |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)'\Sigma_d^{-1}(\mathbf{b}_i - \boldsymbol{\beta}_d) - \frac{1}{2}(W_i - \theta_{C_i,d})^2\right\}}{\sum_{d=1}^{K_D} \pi^D_d |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)'\Sigma_d^{-1}(\mathbf{b}_i - \boldsymbol{\beta}_d) - \frac{1}{2}(W_i - \theta_{C_i,d})^2\right\}}$$

  $\theta_{C_i,d} = \mathbf{Z}_i^T \boldsymbol{\eta}$ in the latent class probit submodel given $D_i = d$ and C_i as well as other covariates.

- **update** the mean profile class parameters:

  - **update** $\boldsymbol{\beta}_d$: Assuming the prior for $\boldsymbol{\beta}_d \overset{ind}{\sim} \text{MVN}(\boldsymbol{\nu}, \mathbf{V})$, then the full conditional posterior density for $\boldsymbol{\beta}_d$ for $d = 1,...,K_D$ is $[\boldsymbol{\beta}_d|.] \sim \text{MVN}(\tilde{\boldsymbol{\nu}}_d, \tilde{\mathbf{V}}_d)$

where

$$\tilde{\nu}_d = \left\{ \mathbf{V}^{-1} + \Sigma_d^{-1} \sum_{i=1}^{n} \mathrm{I}(D_i = d) \right\}^{-1} \left\{ \mathbf{V}^{-1}\boldsymbol{\nu} + \Sigma_d^{-1} \sum_{i=1}^{n} \mathrm{I}(D_i = d)\mathbf{b}_i \right\}$$

$$\tilde{\mathbf{V}}_d = \left\{ \mathbf{V}^{-1} + \Sigma_d^{-1} \sum_{i=1}^{n} \mathrm{I}(D_i = d) \right\}^{-1}$$

– **update** Σ_d: Assuming the prior for $\Sigma_d \overset{ind}{\sim}$ Inverse-Wishart(m, Λ), then the full conditional posterior density is, $[\Sigma_d|\cdot] \sim$ Inverse-Wishart$\left(\tilde{m}_d, \tilde{\Lambda}_d\right)$ where

$$\tilde{m}_d = m + \sum_{i=1}^{n} \mathrm{I}(D_i = d)$$

$$\tilde{\Lambda}_d = \left\{ \Lambda^{-1} + \sum_{i=1}^{n} \mathrm{I}(D_i = d) \left(\mathbf{b}_i - \boldsymbol{\beta}_d\right) \left(\mathbf{b}_i - \boldsymbol{\beta}_d\right)' \right\}^{-1}$$

- **update** the mixing proportion $\{\pi_d^D\}_d$: assuming $\{\pi_d^D\}_d \sim$ Dirichlet$(e_1^D, ..., e_{K_D}^D)$ then the full conditional posterior distribution is $\left[\{\pi_d^D\}_d|\cdot\right] \sim$ Dirichlet$(\{e_d^D + \sum_{i=1}^{n} \mathrm{I}(D_i = d)\}_d)$.

- **update** the variance class memberships C_i, $i = 1, ..., n$: the full conditional posterior distribution $[C_i|\cdot] \sim$ Multinomial$(\tilde{\pi}_{i1}^C, ..., \tilde{\pi}_{iK_C}^C)$ where

$$\tilde{\pi}_{i1}^C = \Pr(C_i = c|.) = \frac{\pi_c^C \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma_i^2 - \mu_c\right)^2 - \frac{1}{2}(W_i - \theta_{c,D_i})^2\right\}}{\sum_{c=1}^{K_C} \pi_c^C \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma_i^2 - \mu_c\right)^2 - \frac{1}{2}(W_i - \theta_{c,D_i})^2\right\}}$$

$\theta_{c,D_i} = \mathbf{Z}_i\boldsymbol{\eta}$ in the probit latent class submodel given $C_i = c$ and D_i as well as other covariates.

- **update** the variance class parameters:

– **update** μ_c: assuming the prior for $\mu_c \overset{ind}{\sim} \mathrm{N}(a,b)$, then the full conditional posterior distribution is, $[\mu_c|\cdot] \sim \mathrm{N}\left(\tilde{a},\tilde{b}\right)$ where

$$\tilde{a} = \frac{\sum_{i=1}^{n} \mathrm{I}(C_i = c)\log\sigma_i^2/\tau^2 + a/b}{1/b + \sum_{i=1}^{n} \mathrm{I}(C_i = c)/\tau^2}$$

$$\tilde{b} = \left\{1/b + \sum_{i=1}^{n} \mathrm{I}(C_i = c)/\tau^2\right\}^{-1}$$

– **update** τ^2: assuming $\tau^2 \sim$ Inverse-Gamma(v,e), then the full conditional posterior distribution is

$$[\tau^2|\cdot] \sim \text{Inverse-Gamma}\left\{v + \frac{n}{2}, e + \sum_{i=1}^{n}\sum_{c=1}^{K_C} \frac{1}{2}\mathrm{I}(C_i = c)\left(\log\sigma_i^2 - \mu_c\right)^2\right\}.$$

- **update** the mixing proportions $\{\pi_c^C\}_c$: assuming $\{\pi_c^C\}_c \sim$ Dirichlet$(e_1^C, ..., e_{K_C}^C)$ then the full conditional posterior distribution is

$$\left[\{\pi_c^C\}_c|\cdot\right] \sim \text{Dirichlet}\left(\{e_c^C + \sum_{i=1}^{n} \mathrm{I}(C_i = c)\}_c\right).$$

- **update** the random effects $\mathbf{b}_i$, $i = 1,...,n$ the full conditional posterior distribution is $\mathbf{b}_i$ $[\mathbf{b}_i|.] \sim \mathrm{MVN}\left(\tilde{\beta}_i, \tilde{\Sigma}_i\right)$, where

$$\tilde{\boldsymbol{\beta}}_i = \left(\Sigma_{D_i}^{-1} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n_i} \mathbf{t}_{ij}\mathbf{t}_{ij}'\right)^{-1}\left(\Sigma_{D_i}^{-1}\boldsymbol{\beta}_{D_i} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n_i} y_{ij}\mathbf{t}_{ij}\right)$$
$$\tilde{\Sigma}_{id} = \left(\Sigma_{D_i}^{-1} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n_i} \mathbf{t}_{ij}\mathbf{t}_{ij}'\right)^{-1}$$

$\mathbf{t}_{ij}$ is a vector of functional forms of the observation time t_{ij} for the longitudinal measurement y_{ij} such that $y_{ij} \sim \mathrm{N}\{f(\mathbf{b}_i; t_{ij}), \sigma_i^2\}$ with $f(\mathbf{b}_i; t_{ij}) = \mathbf{b}_i^T\mathbf{t}_{ij}$.

- **update** the variances σ_i^2, $i = 1,...,n$

$$\pi(\sigma_i^2|.) \propto \prod_{c=1}^{K_C} \mathrm{LN}\left(\sigma_i^2;\mu_c,\tau^2\right)^{\mathrm{I}(C_i=c)} \prod_{j=1}^{n_i} \mathrm{N}\left\{y_{ij};f(\mathbf{b}_i;t_{ij}),\sigma_i^2\right\}$$
$$\propto (\sigma_i^2)^{-\frac{n_i}{2}-1} \exp\left[-\sum_{c=1}^{K_C} \mathrm{I}(C_i=c)\frac{\left(\log\sigma_i^2-\mu_c\right)^2}{2\tau^2} - \frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i}\left\{y_{ij}-f(\mathbf{b}_i;t_{ij})\right\}^2\right]$$

$\mathrm{LN}\left(\sigma_i^2;\mu_c,\tau^2\right)$ represents the density of log normal disribution with mean μ_c and variance τ^2 evaluated at σ_i^2 and $\mathrm{N}\left\{y_{ij};f(\mathbf{b}_i;t_{ij}),\sigma_i^2\right\}$ represents the density of normal distribution with mean $f(\mathbf{b}_i;t_{ij})$ and variance σ_i^2 evaluated at y_{ij}. Since there is no closed form of the full conditional posterior density, the draws for σ_i^2, $i = 1,...,n$ at each iteration of the Gibbs sampling are obtained using the inverse cumulative distribution sampling method.

(2) update for LC probit model:

- **update** W_i, $i = 1,...,n$

$$[W_i|o_i = 1,\cdot] \sim \mathrm{N}(\theta_{C_i,D_i},1)I_{(0,\infty)}(\cdot)$$
$$[W_i|o_i = 0,\cdot] \sim \mathrm{N}(\theta_{C_i,D_i},1)I_{(-\infty,0)}(\cdot)$$

where, $\theta_{C_i,D_i} = \mathbf{Z}_i^T\boldsymbol{\eta}$ in the latent class probit submodel given C_i and D_i.

- **update** $\boldsymbol{\eta}$: Assuming the prior for $\boldsymbol{\eta} \sim \mathrm{MVN}(\boldsymbol{\nu}_{\boldsymbol{\eta}},\mathbf{V}_{\boldsymbol{\eta}})$, then the full conditional posterior density for $\boldsymbol{\eta}$ is $\left[\boldsymbol{\eta}|.\right] \sim \mathrm{MVN}\left(\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}},\tilde{\mathbf{V}}_{\boldsymbol{\eta}}\right)$ where

$$\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}} = \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i'\right)^{-1}\left(\mathbf{V}_{\boldsymbol{\eta}}^{-1}\boldsymbol{\nu}_{\boldsymbol{\eta}} + \sum_{i=1}^{n}W_i\mathbf{Z}_i\right)$$
$$\tilde{\mathbf{V}}_{\boldsymbol{\eta}} = \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i'\right)^{-1}$$

$\mathbf{Z}_i$ is the i^{th} row of the design matrix in the probit submodel given D_i and C_i as well as other covariates.

Posterior computations for the joint MSRE model

(1) update for longitudinal submodel

- **update** the mean profile class memberships D_i, $i = 1,...,n$: the full conditional posterior distribution $[D_i|\cdot] \sim \text{Multinomial}(\tilde{\pi}^D_{i1},...,\tilde{\pi}^D_{iK_D})$, where

$$\tilde{\pi}^D_{id} = \Pr(D_i = d|.) = \frac{\pi^D_d |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)'\Sigma_d^{-1}(\mathbf{b}_i - \boldsymbol{\beta}_d)\right\}}{\sum_{d=1}^{K_D} \pi^D_d |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)'\Sigma_d^{-1}(\mathbf{b}_i - \boldsymbol{\beta}_d)\right\}}$$

- **update** the mean profile class parameters:

  - **update** $\boldsymbol{\beta}_d$: Assuming the prior for $\boldsymbol{\beta}_d \overset{ind}{\sim} \text{MVN}(\boldsymbol{\nu}, \mathbf{V})$, then the full conditional posterior density for $\boldsymbol{\beta}_d$ for $d = 1,...,K_D$ is $\left[\boldsymbol{\beta}_d|.\right] \sim \text{MVN}\left(\tilde{\boldsymbol{\nu}}_d, \tilde{\mathbf{V}}_d\right)$ where

$$\tilde{\nu}_d = \left\{\mathbf{V}^{-1} + \Sigma_d^{-1}\sum_{i=1}^{n} \text{I}(D_i = d)\right\}^{-1}\left\{\mathbf{V}^{-1}\boldsymbol{\nu} + \Sigma_d^{-1}\sum_{i=1}^{n} \text{I}(D_i = d)\mathbf{b}_i\right\}$$
$$\tilde{\mathbf{V}}_d = \left\{\mathbf{V}^{-1} + \Sigma_d^{-1}\sum_{i=1}^{n} \text{I}(D_i = d)\right\}^{-1}$$

  - **update** Σ_d: Assuming the prior for $\Sigma_d \overset{ind}{\sim}$ Inverse-Wishart(m, Λ), then the full conditional posterior density is, $[\Sigma_d|\cdot] \sim$ Inverse-Wishart$\left(\tilde{m}_d, \tilde{\Lambda}_d\right)$ where

$$\tilde{m}_d = m + \sum_{i=1}^{n} \text{I}(D_i = d)$$
$$\tilde{\Lambda}_d = \left\{\Lambda^{-1} + \sum_{i=1}^{n} \text{I}(D_i = d)\left(\mathbf{b}_i - \boldsymbol{\beta}_d\right)\left(\mathbf{b}_i - \boldsymbol{\beta}_d\right)'\right\}^{-1}$$

- **update** the mixing proportion $\{\pi^D_d\}_d$: assuming $\{\pi^D_d\}_d \sim \text{Dirichlet}(e^D_1,....,e^D_{K_D})$ then the full conditional posterior distribution is $\left[\{\pi^D_d\}_d|\cdot\right] \sim \text{Dirichlet}(\{e^D_d + \sum_{i=1}^{n} \text{I}(D_i = d)\}_d)$.

- **update** the variance class memberships C_i, $i = 1,...,n$: the full conditional posterior distribution $[C_i|\cdot] \sim \text{Multinomial}(\tilde{\pi}^C_{i1},...,\tilde{\pi}^C_{iK_C})$ where

$$\tilde{\pi}^C_{i1} = \Pr(C_i = c|.) = \frac{\pi^C_c \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma^2_i - \mu_c\right)^2\right\}}{\sum_{c=1}^{K_C} \pi^C_c \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma^2_i - \mu_c\right)^2\right\}}$$

- **update** the variance class parameters:

  - **update** μ_c: assuming the prior for $\mu_c \overset{ind}{\sim} \text{N}(a,b)$, then the full conditional posterior distribution is, $\left[\mu_c|\cdot\right] \sim \text{N}\left(\tilde{a},\tilde{b}\right)$ where

$$\tilde{a} = \frac{\sum_{i=1}^{n} \text{I}(C_i = c)\log\sigma^2_i/\tau^2 + a/b}{1/b + \sum_{i=1}^{n} \text{I}(C_i = c)/\tau^2}$$

$$\tilde{b} = \left\{1/b + \sum_{i=1}^{n} \text{I}(C_i = c)/\tau^2\right\}^{-1}$$

  - **update** τ^2: assuming $\tau^2 \sim \text{Inverse-Gamma}(v,e)$, then the full conditional posterior distribution is

$$\left[\tau^2|\cdot\right] \sim \text{Inverse-Gamma}\left\{v + \frac{n}{2}, e + \sum_{i=1}^{n}\sum_{c=1}^{K_C} \frac{1}{2}\text{I}(C_i = c)\left(\log\sigma^2_i - \mu_c\right)^2\right\}.$$

- **update** the mixing proportions $\{\pi^C_c\}_c$: assuming $\{\pi^C_c\}_c \sim \text{Dirichlet}(e^C_1,....,e^C_{K_C})$ then the full conditional posterior distribution is

$$\left[\{\pi^C_c\}_c|\cdot\right] \sim \text{Dirichlet}\left(\{e^C_c + \sum_{i=1}^{n} \text{I}(C_i = c)\}_c\right).$$

- **update** the random effects $\mathbf{b}_i$, $i = 1, ..., n$ the full conditional posterior distribution is $\mathbf{b}_i$ $[\mathbf{b}_i|.] \sim \text{MVN}\left(\tilde{\boldsymbol{\beta}}_i, \tilde{\Sigma}_i\right)$, where

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\Sigma}_{id}\left\{\Sigma_{D_i}^{-1}\boldsymbol{\beta}_{D_i} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n_i} y_{ij}\mathbf{t}_{ij} + (\mathbf{Z}_i^T\boldsymbol{\eta} - \tilde{g}(\boldsymbol{\eta}, \sigma_i^2))\mathbf{g}(\boldsymbol{\eta}, \sigma_i^2)\right\}$$

$$\tilde{\Sigma}_{id} = \left\{\Sigma_{D_i}^{-1} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n_i}\mathbf{t}_{ij}\mathbf{t}_{ij}^T + \mathbf{g}(\boldsymbol{\eta}, \sigma_i^2)\mathbf{g}(\boldsymbol{\eta}, \sigma_i^2)^T\right\}^{-1}$$

$\mathbf{t}_{ij}$ is a functional form of the time t_{ij} for the longitudinal measurement y_{ij} such that $y_{ij} \sim \text{N}\{f(\mathbf{b}_i; t_{ij}), \sigma_i^2\}$ with $f(\mathbf{b}_i; t_{ij}) = \mathbf{b}_i^T\mathbf{t}_{ij}$. $\mathbf{g}(\eta, \sigma_i^2)$ is a vector such that $\mathbf{Z}_i^T\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\eta}, \sigma_i^2)'\mathbf{b}_i + \tilde{g}(\boldsymbol{\eta}, \sigma_i^2)$ in the shared random effects and variances model.

- **update** the variances σ_i^2, $i = 1, ..., n$

$$\begin{aligned}\pi(\sigma_i^2|.) &\propto \prod_{c=1}^{K_C}\text{LN}\left(\sigma_i^2; \mu_c, \tau^2\right)^{\text{I}(C_i=c)}\prod_{j=1}^{n_i}\text{N}\{y_{ij}; f(\mathbf{b}_i; t_{ij}), \sigma_i^2\}\,\text{N}(W_i; \mathbf{Z}_i^T\boldsymbol{\eta}, 1)\\ &\propto (\sigma_i^2)^{-\frac{n_i}{2}-1}\exp\left\{-\sum_{c=1}^{K_C}\text{I}(C_i = c)\frac{\left(\log\sigma_i^2 - \mu_c\right)^2}{2\tau^2}\right\}\\ &\quad\times\exp\left[-\frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i}\left\{y_{ij} - f(\mathbf{b}_i; t_{ij})\right\}^2 - \frac{1}{2}(W_i - \mathbf{Z}_i^T\boldsymbol{\eta})^2\right]\end{aligned}$$

$\text{LN}\left(\sigma_i^2; \mu_c, \tau^2\right)$ represents the density of log normal disribution with mean μ_c and variance τ^2 evaluated at σ_i^2; $\text{N}\{y_{ij}; f(\mathbf{b}_i; t_{ij}), \sigma_i^2\}$ represents the density of normal distribution with mean $f(\mathbf{b}_i; t_{ij})$ and variance σ_i^2 evaluated at y_{ij} and similarly $\text{N}(W_i; \mathbf{Z}_i^T\boldsymbol{\eta}, 1)$ represents the density of normal distribution with mean $\mathbf{Z}_i^T\boldsymbol{\eta}$ and variance 1 evaluated at W_i. Since there is no closed form of the full conditional posterior density, the draws for σ_i^2, $i = 1, ..., n$ at each iteration of the Gibbs sampling are obtained using the inverse cumulative distribution sampling method.

(2) update for MSRE probit model:

- **update** W_i, $i = 1, \ldots, m$

$$[W_i | o_i = 1, \cdot] \sim \mathrm{N}(\mathbf{Z}_i^T \boldsymbol{\eta}, 1) I_{(0,\infty)}(\cdot)$$
$$[W_i | o_i = 0, \cdot] \sim \mathrm{N}(\mathbf{Z}_i^T \boldsymbol{\eta}, 1) I_{(-\infty,0)}(\cdot)$$

  where, $\mathbf{Z}_i$ is the i^{th} row of the design matrix in the MSRE probit submodel.

- **update** $\boldsymbol{\eta}$: Assuming the prior for $\boldsymbol{\eta} \sim \mathrm{MVN}(\boldsymbol{\nu}_{\boldsymbol{\eta}}, \mathbf{V}_{\boldsymbol{\eta}})$, then the full conditional posterior density for $\boldsymbol{\eta}$ is $[\boldsymbol{\eta}|.] \sim \mathrm{MVN}\left(\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}}, \tilde{\mathbf{V}}_{\boldsymbol{\eta}}\right)$ where

$$\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}} = \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i'\right)^{-1} \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} \boldsymbol{\nu}_{\boldsymbol{\eta}} + \sum_{i=1}^{n} W_i \mathbf{Z}_i\right)$$
$$\tilde{\mathbf{V}}_{\boldsymbol{\eta}} = \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i'\right)^{-1}$$

  $\mathbf{Z}_i$ is the i^{th} row of the design matrix in the MSRE probit submodel.

Computation of DIC

DIC is given by

$$\mathrm{DIC}(\mathbf{x}) = -4\mathrm{E}_{\boldsymbol{\phi}}\left\{\log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x}\right\} + 2\log f\left\{\mathbf{x} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x})\right\}$$

Celeux et al. (2006) suggests that, when the model has missing data or latent variables, the appropriate DIC measure is obtained by first considering the DIC measure in the "complete data" setting, where $\mathbf{x}$ indicates the fully observed data, and $\mathbf{z}$ the unobserved (typically latent) data:

$$\mathrm{DIC}(\mathbf{x}, \mathbf{z}) = -4\mathrm{E}_{\boldsymbol{\phi}}\left\{\log f(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\phi}) \mid \mathbf{x}, \mathbf{z}\right\} + 2\log f\left\{\mathbf{x}, \mathbf{z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z})\right\}$$

Integrating out over the unobserved data yields

$$\mathrm{DIC}(\mathbf{x}) = \mathrm{E}_{\mathbf{z}}\left[-4\mathrm{E}_{\boldsymbol{\phi}}\left\{\log f(\mathbf{x},\mathbf{z}\mid\boldsymbol{\phi})\mid\mathbf{x}\right\} + 2\log f\left\{\mathbf{x},\mathbf{z}\mid\ \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi}\mid\mathbf{x},\mathbf{z})\right\}\right]$$

$$= -4\mathrm{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\log f(\mathbf{x},\mathbf{z}\mid\boldsymbol{\phi})\mid\mathbf{x}\right\} + 2\mathrm{E}_{\mathbf{z}}\left[\log f\left\{\mathbf{x},\mathbf{z}\mid E_{\boldsymbol{\phi}}(\boldsymbol{\phi}\mid\mathbf{x},\mathbf{z})\right\}\mid\mathbf{x}\right]$$

To obtain DIC for our MSRE model, let $\boldsymbol{\phi}$ denote the model parameter and $\mathbf{z}_i$ the latent variables $(D_i, C_i, \mathbf{b}_i, \sigma_i^2, W_i)'$ for the ith subject. The data $\mathbf{x}_i'$, $i = 1,...,n$ correspond to the longitudinal data $(y_{i1},...,y_{in_i})'$ and the outcome o_i. The complete data log-likelihood (ignoring normalizing constants) for the latent class model is given by

$$\begin{aligned}
&\log f(\mathbf{x},\mathbf{z}\mid\boldsymbol{\phi}) \\
&= \sum_{i=1}^{n}\left[\sum_d \mathrm{I}(D_i=d)\left\{\log\pi_d - \frac{1}{2}\log\mid\Sigma_{\mathbf{d}}\mid - \frac{1}{2}(\mathbf{b}_i-\boldsymbol{\beta}_d)'\Sigma_d^{-1}(\mathbf{b}_i-\boldsymbol{\beta}_d)\right\}\right. \\
&+ \sum_c \mathrm{I}(C_i=c)\left\{\log\pi_c - \log\tau - \log\sigma_i^2 - \frac{1}{2\tau^2}(\log\sigma_i^2-\mu_c)^2\right\} \\
&\left.+ \sum_{j=1}^{n_i}\left\{\log\sigma_i - \frac{(y_{ij}-f(\mathbf{b}_i;t_{ij}))^2}{2\sigma_i^2}\right\} - \frac{1}{2}(W_i-\mathbf{Z}_i'\boldsymbol{\eta})^2\right]
\end{aligned}$$

Dividing $\mathbf{z}_i$ into $\mathbf{z}_{i1} = (D_i, C_i)$ and $\mathbf{z}_{i2} = (\mathbf{b}_i, \sigma_i^2, W_i)$, we have

$$\begin{aligned}
&\mathrm{E}_{\mathbf{z},\boldsymbol{\phi}}\left\{\log f(\mathbf{x},\mathbf{z}\mid\boldsymbol{\phi})\mid\mathbf{x}\right\} = \\
&E_{\mathbf{z}_2,\boldsymbol{\phi}}\left[\sum_{i=1}^{n}\left[\sum_d \mathrm{P}(D_i=d\mid\mathbf{x},\boldsymbol{\phi},\mathbf{z}_2)\left\{\log\pi_d - \tfrac{1}{2}\log\mid\Sigma_{\mathbf{d}}\mid - \tfrac{1}{2}(\mathbf{b}_i-\boldsymbol{\beta}_d)'\Sigma_d^{-1}(\mathbf{b}_i-\boldsymbol{\beta}_d)\right\}\right.\right. \\
&+ \sum_c \mathrm{P}(C_i=c\mid\mathbf{x},\boldsymbol{\phi},\mathbf{z}_2)\left\{\log\pi_c - \log\tau - \log\sigma_i^2 - \frac{1}{2\tau^2}(\log\sigma_i^2-\mu_c)^2\right\} \\
&\left.\left.+ \sum_{j=1}^{n_i}\left\{\log\sigma_i - \frac{(y_{ij}-f(\mathbf{b}_i;t_{ij}))^2}{2\sigma_i^2}\right\} - \frac{1}{2}(W_i-\mathbf{Z}_i'\boldsymbol{\eta})^2\right]\mid\mathbf{x}\right]
\end{aligned}$$

where

$$\mathrm{P}(D_i = d \mid \mathbf{x}, \boldsymbol{\phi}, \mathbf{z}_2) = \frac{\pi_d^D |\Sigma_{\mathbf{d}}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)' \Sigma_d^{-1} (\mathbf{b}_i - \boldsymbol{\beta}_d)\right]}{\sum_{d=1}^{K_D} \pi_d^D |\Sigma_{\mathbf{d}}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)' \Sigma_d^{-1} (\mathbf{b}_i - \boldsymbol{\beta}_d)\right]}$$

$$\mathrm{P}(C_i = c \mid \mathbf{x}, \boldsymbol{\phi}, \mathbf{z}_2) = \frac{\pi_c^C \exp\left[-\frac{1}{2}\left(\log\sigma_i^2 - \mu_c\right)^2 / \tau^2\right]}{\sum_{c=1}^{K_C} \pi_c^C \exp\left[-\frac{1}{2}\left(\log\sigma_i^2 - \mu_c\right)^2 / \tau^2\right]}$$

This expectation can then be approximated from M MCMC draws:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{z},\boldsymbol{\phi}}\left[\{\log f(\mathbf{x} \mid \boldsymbol{\phi}) \mid \mathbf{x}, \mathbf{z}\} \mid \mathbf{x}\right] = \\
&M^{-1} \sum_{m=1}^{M} \left[\sum_{i=1}^{n} \left[\sum_d \mathrm{P}(D_i = d \mid \mathbf{x}, \boldsymbol{\phi}, \mathbf{z}_2)^{(m)} \left\{ \log \pi_d^{(m)} - \frac{1}{2} \log \mid \Sigma_d^{(m)} \mid \right.\right.\right. \\
&\left. -\frac{1}{2}(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})' (\Sigma_d^{-1})^{(m)} (b_i^{(m)} - \boldsymbol{\beta}_d^{(m)}) \right\} \\
&+ \sum_c \mathrm{P}(C_i = c \mid \mathbf{x}, \boldsymbol{\phi}, \mathbf{z}_2)^{(m)} \left\{ \log \pi_c^{(m)} - \log \tau^{(m)} - \log(\sigma_i^2)^{(m)} - \frac{(\log(\sigma_i^2)^{(m)} - \mu_c^{(m)})^2}{2(\tau^2)^{(m)}} \right\} \\
&\left.\left. + \sum_{j=1}^{n_i} \left\{ \log \sigma_i^{(m)} - \frac{(y_{ij} - (b_i^{(m)})' t_{ij})^2}{2(\sigma_i^2)^{(m)}} \right\} - \frac{1}{2}(W_i^{(m)} - \mathbf{Z}_i' \boldsymbol{\eta}^{(m)})^2 \right] \right]
\end{aligned}$$

Obtaining $E_{\mathbf{z}}\left[\log f\left\{\mathbf{x}, \mathbf{z} \mid E_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z})\right\} \mid \mathbf{x}\right]$ requires a bit more effort. We can broadly use the same approach of averaging over the MCMC draws to integrate with respect to $\mathbf{z}$, but instead of using the draws of the model parameters directly, we need to obtain their expectation conditional on the current draw of $\mathbf{z}$. So

$$\begin{aligned}
&\mathrm{E}_{\mathbf{z}}\left[\log f\left\{\mathbf{x}, \mathbf{z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z})\right\} \mid \mathbf{x}\right] = M^{-1} \sum_{m=1}^{M} \left[\sum_{i=1}^{n} \left[\sum_d I(D_i = d)^{(m)} \right.\right. \\
&\left\{ \log \hat{\pi}_d^{(m)} - \frac{1}{2} \log \mid \hat{\Sigma}_d^{(m)} \mid - \frac{1}{2}(\mathbf{b}_i^{(m)} - \hat{\boldsymbol{\beta}}_d^{(m)})' (\hat{\Sigma}_d^{-1})^{(m)} (b_i^{(m)} - \hat{\boldsymbol{\beta}}_d^{(m)}) \right\} \\
&+ \sum_c I(C_i = c)^{(m)} \left\{ \log \hat{\pi}_c^{(m)} - \log \hat{\tau}^{(m)} - \log(\sigma_i^2)^{(m)} - \frac{(\log(\sigma_i^2)^{(m)} - \hat{\mu}_c^{(m)})^2}{2(\hat{\tau}^2)^{(m)}} \right\} \\
&\left.\left. + \sum_{j=1}^{n_i} \left\{ \log \sigma_i^{(m)} - \frac{(y_{ij} - (b_i^{(m)})' t_{ij})^2}{2(\hat{\sigma}_i^2)^{(m)}} \right\} - \frac{1}{2}(W_i^{(m)} - -\mathbf{Z}_i' \boldsymbol{\eta}^{(m)})^2 \right] \right]
\end{aligned}$$

where $\hat{\boldsymbol{\phi}}^{(m)} = E_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z}^{(m)})$.

Some components of $\hat{\boldsymbol{\phi}}^{(m)}$ have closed form solutions:

$$\hat{\pi}_d^{(m)} = \frac{e_d^D + \sum_{i=1}^n \mathrm{I}(D_i = d)^{(m)}}{\sum_d e_d^D + n}$$
$$\hat{\pi}_c^{(m)} = \frac{e_c^C + \sum_{i=1}^n \mathrm{I}(C_i = c)^{(m)}}{\sum_c e_c^C + n}$$
$$\hat{\boldsymbol{\eta}}^{(m)} = \left(V_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^n \mathbf{Z}_i^{(m)} \mathbf{Z}_i^{(m)'} \right)^{-1} \left(V_{\boldsymbol{\eta}}^{-1} \nu_{\boldsymbol{\eta}} + \sum_{i=1}^n W_i^{(m)} \mathbf{Z}_i^{(m)} \right)$$

where $\mathbf{Z}_i^{(m)}$ is the i^{th} row of the design matrix in the probit submodel for the mth MCMC draw and e_d^D, e_c^C, $V_{\boldsymbol{\eta}}$, and $\boldsymbol{\nu}_{\boldsymbol{\eta}}$ are specified hyperprior values.

The other components of $\hat{\boldsymbol{\phi}}^{(m)}$ will have to be obtaining by running small MCMC chains *for each of the main MCMC interations* to get the marginal expectations: $\hat{\boldsymbol{\beta}}_d^{(m)} = (M^*)^{-1} \sum_{m*} \boldsymbol{\beta}_d^{(m,m^*)}$ and $\hat{\Sigma}_d^{(m)} = (M^*)^{-1} \sum_{m*} \Sigma_d^{(m,m^*)}$, where $\boldsymbol{\beta}_d^{(m,m^*)}$ and $\Sigma_d^{(m,m^*)}$ are obtained by alternating draws from the following distributions with known hyperparameters V, ν, m, and Λ:

$\boldsymbol{\beta}_d^{(m,m^*)} \sim \mathbf{MVN}\left(\tilde{\boldsymbol{\nu}}_d^{(m,m^*)}, \tilde{\mathbf{V}}_d^{(m,m^*)} \right)$, where

$$\tilde{\mathbf{V}}_d^{(m,m^*)} = \left\{ \mathbf{V}^{-1} + (\Sigma_d^{(m,m^*-1)})^{-1} \sum_{i=1}^n \mathrm{I}(D_i = d)^{(m)} \right\}^{-1}$$

$$\tilde{\nu}_d^{(m,m^*)} = \tilde{\mathbf{V}}_d^{(m,m^*)} \left\{ \mathbf{V}^{-1} \nu + (\Sigma_d^{(m,m^*-1)})^{-1} \sum_{i=1}^n \mathrm{I}(D_i = d)^{(m)} \mathbf{b}_i^{(m)} \right\}.$$

$\Sigma_d^{(m,m^*)} \sim \mathbf{Inverse\text{-}Wishart}\left(\tilde{m}_d^{(m)}, \tilde{\Lambda}_d^{(m,m^*)} \right)$, where

$$\tilde{m}_d^{(m)} = m + \sum_{i=1}^n \mathrm{I}(D_i = d)^{(m)}$$

$$\tilde{\Lambda}_d^{(m,m^*)} = \left\{ \Lambda^{-1} + \sum_{i=1}^n \mathrm{I}(D_i = d)^{(m)} \left(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m,m^*)} \right) \left(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m,m^*)} \right)' \right\}^{-1}$$

Similarly, $\hat{\mu}_c^{(m)} = (M^*)^{-1}\sum_{m*}\mu_c^{(m,m^*)}$ and $(\hat{\tau}^2)^{(m)} = (M^*)^{-1}\sum_{m*}(\tau^2)^{(m,m^*)}$, where $\mu_c^{(m,m^*)}$ and $(\tau^2)^{(m,m^*)}$ are obtained by alternating draws from the following distributions with known hyperparameters a, b, e, and f:

$\mu_c^{(m,m^*)} \sim \mathbf{N}(\tilde{a}^{(m,m^*)}, \tilde{b}^{(m,m^*)})$, where

$$\tilde{a}^{(m,m^*)} = \frac{\sum_{i=1}^{n} \mathrm{I}(C_i = c)^{(m)} \log(\sigma_i^2)^{(m)}/(\tau^2)^{(m,m^*-1)} + a/b}{1/b + \sum_{i=1}^{n} \mathrm{I}(C_i = c)^{(m)}/(\tau^2)^{(m,m^*-1)}}$$

$$\tilde{b}^{(m,m^*)} = \left\{ 1/b + \sum_{i=1}^{n} \mathrm{I}(C_i = c)^{(m)}/(\tau^2)^{(m,m^*-1)} \right\}^{-1}$$

$(\tau^2)^{(m,m^*)} \sim \textbf{Inverse-Gamma}(\tilde{v}, \tilde{e}^{(m,m^*)})$, where

$$\tilde{v} = v + \frac{n}{2}$$

$$\tilde{e}^{(m,m^*)} = e + \sum_{i=1}^{n}\sum_{c=1}^{K_C} \frac{1}{2}\mathrm{I}(C_i = c)^{(m)} \left\{ \log(\sigma_i^2)^{(m)} - \mu_c^{(m,m^*)} \right\}^2$$

Because we are conditioning on $\mathbf{z}$ and only need the posterior expectation, a modest number of drawn (here we use $M^* = 250$) is found to be sufficient to obtain an accurate approximation.

Similarly, we can obtain DIC for our LC model.

Computation of LPML

For our model, we have

$$\begin{aligned}
\text{CPO}_i^{-1} &= \frac{f(\mathbf{y}_{(-i)},\mathbf{o}_{(-i)}|\mathbf{v})}{f(\mathbf{y},\mathbf{o}|\mathbf{v})} \\
&= \int \frac{f(\mathbf{y}_{(-i)},\mathbf{o}_{(-i)}|\mathbf{C},\mathbf{D},\mathbf{b},\boldsymbol{\sigma}^2,\boldsymbol{\phi},\mathbf{v})f(\mathbf{C},\mathbf{D},\mathbf{b},\boldsymbol{\sigma}^2,\boldsymbol{\phi})}{f(\mathbf{y},\mathbf{o}|\mathbf{v})} d\mathbf{b}d\sigma^2 d\mathbf{C}d\mathbf{D}d\boldsymbol{\phi} \\
&= \int \frac{f(\mathbf{y},\mathbf{o}|\mathbf{C},\mathbf{D},\mathbf{b},\boldsymbol{\sigma}^2,\boldsymbol{\phi},\mathbf{v})f(\mathbf{C},\mathbf{D},\mathbf{b},\sigma^2,\boldsymbol{\phi})}{f(\mathbf{y},\mathbf{o}|\mathbf{v}_i)f(\mathbf{y}_i,o_i|\mathbf{C},\mathbf{D},\mathbf{b},\sigma^2,\boldsymbol{\phi},\mathbf{v})} d\mathbf{b}d\boldsymbol{\sigma}^2 d\mathbf{C}d\mathbf{D}d\boldsymbol{\phi} \\
&= \int \frac{1}{f(\mathbf{y}_i,o_i|\mathbf{C},\mathbf{D},\mathbf{b},\sigma^2,\boldsymbol{\phi},\mathbf{v}_i)} f(\mathbf{C},\mathbf{D},\mathbf{b},\boldsymbol{\sigma}^2,\boldsymbol{\phi}|\mathbf{y},\mathbf{o},\mathbf{v}) d\mathbf{b}d\boldsymbol{\sigma}^2 d\mathbf{C}d\mathbf{D}d\boldsymbol{\phi} \\
&= \int \frac{f(\mathbf{C},\mathbf{D},\mathbf{b},\boldsymbol{\sigma}^2,\boldsymbol{\phi}|\mathbf{y},\mathbf{o},\mathbf{v})}{f(\mathbf{y}_i|\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi},\mathbf{v}_i)f(o_i|C_i,D_i,\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi},\mathbf{v}_i)} d\mathbf{b}d\boldsymbol{\sigma}^2 d\mathbf{C}d\mathbf{D}d\boldsymbol{\phi}
\end{aligned}$$

where $\boldsymbol{\phi}$ is the vector of model parameters which does not include the unobserved random effects and unknown residual variances. $\mathbf{v} = (\mathbf{v}_1,...,\mathbf{v}_n)^T$ include all observed variables including obervation time t_{ij}, $i = 1,...,n_i$, $j = 1,...,n_i$ and baseline covariates of interest. $f(o_i|C_i,D_i,\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi},\mathbf{v}_i)$ can be reduced to $f(o_i|C_i,D_i,\phi,\mathbf{v}_i)$ in the case of LC probit submodel and $f(o_i|\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi},\mathbf{v}_i)$ in the case of MSRE probit submodel. Using the MCMC posterior draws, we estimate CPO_i^{-1} by

$$\frac{1}{B}\sum_{s=1}^{B} \frac{1}{f(\mathbf{y}_i|\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi}^{(s)},\mathbf{v}_i)f(o_i|C_i,D_i,\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi}^{(s)},\mathbf{v}_i)}$$

where B is the number of MCMC posterior draws and $\boldsymbol{\phi}^{(s)}$ is the vector of the posterior draws of all model parameters at the s^{th} iteration. We have,

$$f(\mathbf{y}_i|\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi},\mathbf{v}_i) = \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\{y_{ij} - f(\mathbf{b}_i;t_{ij})\}^2}{2\sigma_i^2}\right]$$

$$f(o_i|C_i,D_i,\boldsymbol{\phi},\mathbf{v}_i) = \Phi(\mathbf{Z}_i^T\boldsymbol{\theta})^{o_i}\left[1-\Phi(\mathbf{Z}_i^T\boldsymbol{\theta})\right]^{1-o_i} \text{ for LC probit submodel}$$

$$f(o_i|\mathbf{b}_i,\sigma_i^2,\boldsymbol{\phi},\mathbf{v}_i) = \Phi(\mathbf{Z}_i^T\gamma)^{o_i}\left[1-\Phi(\mathbf{Z}_i^T\boldsymbol{\gamma})\right]^{1-o_i} \text{ for MSRE probit submodel}$$

where, $\Phi(\cdot)$ is the cdf for standard normal distribution. $\mathbf{Z}_i$ denotes the corresponding i^{th} row in the design matrix for either LC or MSRE probit submodel given $\mathbf{v}_i$.

We retain every 5^{th} of the 100,000 posterior draws after the chains converge and divide these posterior draws into 20 blocks of length 1000 draws. To obtain stable LPML measures, we calculate the CPO's and LPML based on each of the 20 blocks of draws and then report the median LPML. We found this approach would lead to relatively stable LPML results in our simulations.

APPENDIX B

Computation Details in Chapter III

Details of posterior computations when assuming t model for longitudinal observations

note: in the case of assuming normal model for longitudinal observations, we are not updating m_{ij} but let m_{ij} be constant 1; and the missing longitudinal observation y_{ij} is drawn from $\mathrm{N}\left(\mu(t_{ij}), \sigma_i^2\right)$.

(1) update for longitudinal submodel

- **update** the mean profile class memberships D_i, $i = 1, ..., n$: the full conditional posterior distribution $[D_i|\cdot] \sim \text{Multinomial}(\tilde{\pi}_{i1}^D, ..., \tilde{\pi}_{iK_D}^D)$, where

$$\tilde{\pi}_{id}^D = \Pr(D_i = d|.) = \frac{\pi_d^D |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{b}_i - \beta_d)' \Sigma_d^{-1} (\mathbf{b}_i - \beta_d)\right]}{\sum_{d=1}^{K_D} \pi_d^D |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{b}_i - \beta_d)' \Sigma_d^{-1} (\mathbf{b}_i - \beta_d)\right]}.$$

- **update** the mean profile class parameters:

  – **update** $\beta_d = (\beta_{d1}, ..., \beta_{dL})$:

  Assuming the prior for $\beta_{d1} \sim \mathrm{N}(0, v)$ and first order random walk prior

$\beta_{dl} \sim \mathrm{N}(\beta_{d,l-1}, \tau^2_{\beta d})$, $l = 2,...,L$, then the prior for β_d can be written as: $\pi(\beta_d) = \left(\frac{1}{\sqrt{2\pi}\tau_{\beta d}}\right)^{L-1} \exp\{-\frac{1}{2}\beta_d^T \mathbf{V} \beta_d\}$, where $\mathbf{V} = \begin{pmatrix} v^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \mathbf{P}^T\mathbf{P}/\tau^2_{\beta d}$ and $\mathbf{P} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}$ is the $(L-1)\times L$ penalty matrix. Then the full conditional posterior density for β_d for $d = 1,...,K_D$ is $\left[\beta_d|.\right] \sim \mathrm{MVN}\left(\tilde{\nu}_d, \tilde{\mathbf{V}}_d\right)$

$$\tilde{\nu}_d = \left[\mathbf{V} + \Sigma^{-1}\sum_{i=1}^{n} \mathrm{I}(D_i = d)\right]^{-1}\left[\Sigma^{-1}\sum_{i=1}^{n} \mathrm{I}(D_i = d)\mathbf{b}_i\right]$$
$$\tilde{\mathbf{V}}_d = \left[\mathbf{V} + \Sigma^{-1}\sum_{i=1}^{n} \mathrm{I}(D_i = d)\right]^{-1}.$$

- **update** Σ: Assuming the prior for $\Sigma \overset{ind}{\sim}$ Inverse-Wishart(m, Λ), where m and Λ are the degrees of freedom and scale matrix, respectively, then the full conditional posterior density is, $[\Sigma|\cdot] \sim$ Inverse-Wishart$\left(\tilde{m}_d, \tilde{\Lambda}_d\right)$ where

$$\tilde{m}_d = m + n$$
$$\tilde{\Lambda}_d = \left[\Lambda + \sum_{i=1}^{n}\left(\mathbf{b}_i - \beta_{D_i}\right)\left(\mathbf{b}_i - \beta_{D_i}\right)'\right].$$

- **update** $\tau^2_{\beta d}$: assuming $\tau^2_{\beta d} \sim$ Inverse-Gamma(v, e), where v and e are the shape and rate parameters, then the full conditional posterior distribution is

$$\left[\tau^2_{\beta d}|\cdot\right] \sim \text{Inverse-Gamma}\left(v + \frac{L-1}{2}, e + \frac{1}{2}\beta_d^T\mathbf{P}\beta_d\right),$$

where L is the number of B spline basis functions.

- **update** the mixing proportion $\{\pi^D_d\}_d$: assuming $\left[\{\pi^D_d\}_d\right] \sim$ Dirichlet$(e^D_1, ..., e^D_{K_D})$

then the full conditional posterior distribution is

$$[\{\pi_d^D\}_d|\cdot] \sim \text{Dirichlet}(\{e_d^D + \sum_{i=1}^{n} \text{I}(D_i = d)\}_d).$$

- **update** the variance class memberships C_i, $i = 1,...,n$: the full conditional posterior distribution $[C_i|\cdot] \sim \text{Multinomial}(\tilde{\pi}_{i1}^C,...,\tilde{\pi}_{iK_C}^C)$ where

$$\tilde{\pi}_{ic}^C = \Pr(C_i = c|.) = \frac{\pi_c^C \exp\left[-\frac{1}{2}\left(\log\sigma_i^2 - \mu_c\right)^2/\tau^2\right]}{\sum_{c=1}^{K_C} \pi_c^C \exp\left[-\frac{1}{2}\left(\log\sigma_i^2 - \mu_c\right)^2/\tau^2\right]}.$$

- **update** the variance class parameters:

  - **update** μ_c: assuming the prior for $\mu_c \overset{ind}{\sim} \text{N}(a,b)$, then the full conditional posterior distribution is, $\left[\mu_c|\cdot\right] \sim \text{N}\left(\tilde{a},\tilde{b}\right)$ where

$$\tilde{a} = \frac{\sum_{i=1}^{n} \text{I}(C_i = c)\log\sigma_i^2/\tau^2 + a/b}{1/b + \sum_{i=1}^{n} \text{I}(C_i = c)/\tau^2}$$

$$\tilde{b} = \left(1/b + \sum_{i=1}^{n} \text{I}(C_i = c)/\tau^2\right)^{-1}$$

  - **update** τ^2: assuming $\tau^2 \sim$ Inverse-Gamma(v,e), then the full conditional posterior distribution is

$$\left[\tau^2|\cdot\right] \sim \text{Inverse-Gamma}\left(v + \frac{n}{2}, e + \sum_{i=1}^{n}\sum_{c=1}^{K_C} \frac{1}{2}\text{I}(C_i = c)\left(\log\sigma_i^2 - \mu_c\right)^2\right).$$

- **update** the mixing proportions $\{\pi_c^C\}_c$: assuming $\left[\{\pi_c^C\}_c\right] \sim \text{Dirichlet}(e_1^C,....,e_{K_C}^C)$

then the full conditional posterior distribution is

$$[\{\pi_c^C\}_c|\cdot] \sim \text{Dirichlet}\left(\{e_c^C + \sum_{i=1}^{n} \text{I}(C_i = c)\}_c\right).$$

- **update** the random effects $\mathbf{b}_i$, $i = 1,...,n$ the full conditional posterior distribution is $\mathbf{b}_i$ $[\mathbf{b}_i|.] \sim \text{MVN}\left(\tilde{\beta}_i, \tilde{\Sigma}_i\right)$, where

$$\tilde{\Sigma}_{id} = \left[\Sigma^{-1} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n_i} m_{ij}\phi_{ij}\phi_{ij}' + \mathbf{M}_{bi}\mathbf{M}_{bi}^T\right]^{-1}$$
$$\tilde{\beta}_i = \tilde{\Sigma}_{id}\left[\Sigma^{-1}\beta_{D_i} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n_i} y_{ij}m_{ij}\phi_{ij} + (W_i - \alpha_1 - \mathbf{x}_i^T\lambda_1)\mathbf{M}_{bi}\right],$$

where, $\phi_{ij} = \left(\phi_1(t_{ij}), \cdots, \phi_L(t_{ij})\right)^T$ a vector of B spline basis functions evaluated at time t_{ij} such that $\mu_i(t_{ij}) = \mathbf{b}_i'\phi_{ij}$ where $y_{ij} \sim \text{N}\left(\mu_i(t_{ij}), \sigma_i^2/m_{ij}\right)$ and $\mathbf{M}_{bi}$ is defined such that $\int_T \mu_i(t)\theta_0(t)dt = \int_T \mathbf{b}_i'\phi(t)\psi^0(t)^T\tilde{\theta}dt = \mathbf{b}_i^T\mathbf{G}_T^0\tilde{\theta}_0 = \mathbf{b}_i^T\mathbf{M}_{bi}$.

- **update** the variances σ_i^2, $i = 1,...,n$

$$\pi(\sigma_i^2|.) \propto (\sigma_i^2)^{-\frac{n_i}{2}-1}\exp\left[-\sum_{c=1}^{K_C}\text{I}(C_i = c)\frac{\left(\log\sigma_i^2 - \mu_c\right)^2}{2\tau^2} - \frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i} m_{ij}\left(y_{ij} - \mathbf{b}_i'\phi_{ij}\right)^2\right]$$
$$\times \exp\left[-\frac{1}{2}\left(W_i - \alpha_0 - \mathbf{x}_i^T\lambda_0 - \int_T \mu_i(t)\theta_0(t)dt\right)^2\right]$$

where, $\mathbf{x}_i$ is a vector of baseline covariate including subject specific residual variance $\frac{v_i}{v_i-2}\sigma_i^2$. Since there is no closed form of the full conditional posterior density, the draws for σ_i^2, $i = 1,...,n$ at each iteration of the Gibbs sampling are obtained using the inverse cumulative distribution sampling method.

- **update** m_{ij}, $j = 1,...,n_i$, given that $m_{ij} \sim \text{gamma}(v/2, v/2)$ where $v/2$ and $v/2$ are the shape and rate parameter in gamma distribution, then the full conditional posterior distribution for m_{ij} is $m_{ij} \sim \text{gamma}\left(\frac{v+1}{2}, \frac{1}{2}\left(\frac{(y_{ij}-\mu(t_{ij}))^2}{\sigma_i^2} + v\right)\right)$.

- **update** missing data: missing y_{ij} at time t_{ij} is drawn from $\mathrm{t}\left(\mu(t_{ij}), \sigma_i^2, v\right)$.

(2) update for outcome probit submodel:

- **update** W_i, $i = 1, ..., m$

$$[W_i | o_i = 0, \cdot] \sim \mathrm{N}(\eta_i^W, 1) I_{(-\infty, 0)}(\cdot)$$
$$[W_i | o_i = 1, \cdot] \sim \mathrm{N}(\eta_i^W, 1) I_{(0, \gamma_1)}(\cdot)$$
$$[W_i | o_i = 2, \cdot] \sim \mathrm{N}(\eta_i^W, 1) I_{(\gamma_1, \infty)}(\cdot)$$

where, $\eta_i^W = \alpha_0 - \mathbf{x}_i^T \lambda_0 - \int_T \mu_i(t)\theta_0(t)dt$ and γ_1 is the cutoff.

- **update** cutoff γ_1: assuming flat prior on γ_1, then the full conditional posterior density for γ_1 is $\mathrm{Unif}(\mathrm{Max}_{O_i=1} W_i, \mathrm{Min}_{O_i=2} W_i)$.

- **update** $(\alpha, \lambda)'$: Assuming independent prior for $(\alpha, \lambda)' \sim \mathrm{MVN}(\nu_{\alpha\lambda}, \mathbf{V}_{\alpha\lambda})$, then the full conditional posterior density for $\alpha\lambda$ is $[\alpha\lambda|.] \sim \mathrm{MVN}\left(\tilde{\nu}_{\alpha\lambda}, \tilde{\mathbf{V}}_{\alpha\lambda}\right)$ where

$$\tilde{\nu}_\lambda = \left[\mathbf{V}_{\alpha\lambda}^{-1} + \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i'\right]^{-1} \left[\mathbf{V}_{\alpha\lambda}^{-1} \nu_{\alpha\lambda} + \sum_{i=1}^n \left(W_i - \int_T \mu_i(t)\theta_0(t)dt\right) \mathbf{z}_i\right]$$
$$\tilde{\mathbf{V}}_{\alpha\lambda} = \left[\mathbf{V}_{\alpha\lambda}^{-1} + \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i'\right]^{-1},$$

where $\mathbf{z}_i = (1, \mathbf{x}_i)'$ is a vector of constant 1, baseline covariates and residual variance $\frac{v}{v-2}\sigma_i^2$

- **update** $\tilde{\theta}_\mathbf{0} = (\tilde{\theta}_{01}, ..., \tilde{\theta}_{0K})'$: Assuming the prior for $\tilde{\theta}_{01} \sim \mathrm{N}(0, v_{\theta_0})$ and first order random walk prior $\tilde{\theta}_{0k} \sim \mathrm{N}(\tilde{\theta}_{k-1}, \tau_{\theta_0}^2)$, $k = 2, ..., K$, then the prior for $\tilde{\theta}_0$ can be written as: $\pi(\tilde{\theta}) = \left(\frac{1}{\sqrt{2\pi}\tau_{\theta_0}}\right)^{L-1} \exp\{-\frac{1}{2}\tilde{\theta}_0^T \mathbf{V}_{\theta_0} \tilde{\theta}_0\}$, where $\mathbf{V}_{\theta_0} = \begin{pmatrix} v_{\theta_0}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} +$

$\mathbf{P}_{\theta_0}^T \mathbf{P}_{\theta_0} / \tau_{\theta_0}^2$ with $\mathbf{P}_{\theta_0} =$

$$\begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}$$

is the $(K-1)\times K$ penalty matrix. Then the full conditional posterior density for $\tilde{\theta}_0$ is $[\tilde{\theta}_0|.] \sim \text{MVN}\left(\tilde{\nu}_{\theta_0}, \tilde{\mathbf{V}}_{\theta_0}\right)$ where

$$\tilde{\nu}_\theta = \left[\mathbf{V}_{\theta_\mathbf{0}}^{-1} + \sum_{i=1}^{n} \mathbf{z}_{\theta i}\mathbf{z}'_{\theta i}\right]^{-1} \left[\mathbf{V}_{\theta_0}^{-1}\nu_{\theta_0} + \sum_{i=1}^{n}\left(W_i - \alpha_0 - \mathbf{x}_i^T\lambda_0\right)\mathbf{z}_{\theta i}\right]$$
$$\tilde{\mathbf{V}}_{\theta_0} = \left[\mathbf{V}_{\theta_0}^{-1} + \sum_{i=1}^{n} \mathbf{z}_{\theta i}\mathbf{z}'_{\theta i}\right]^{-1},$$

where $\mathbf{z}_{\theta i}$ is a vector of K elements defined such that

$$\int_T \mu_i(t)\theta_0(t)dt = \int_T \mathbf{b}_i'\phi(t)\psi^0(t)^T\tilde{\theta}_0 dt = \mathbf{b}_i^T\mathbf{G}_T^0\tilde{\theta}_0 = \tilde{\theta}_0^T\mathbf{z}_{\theta i}.$$

- **update** $\tau_{\theta_0}^2$: assuming $\tau_{\theta_0}^2 \sim \text{IG}(f, g)$, where f and g are the shape and rate parameters, then the full conditional posterior distribution is

$$\left[\tau_{\theta_0}^2|\cdot\right] \sim \text{IG}\left(f + \frac{K-1}{2}, g + \frac{1}{2}\tilde{\theta}_0^T\mathbf{P}_{\theta_0}\tilde{\theta}_0\right),$$

where K is the number of B spline basis functions to express $\theta_0(t)$ in the ordinal probit submodel.

APPENDIX C

Computation Details in Chapter IV

Appendix

Exact expressions of $\mathbf{P}(D_i = d \mid \boldsymbol{\phi}, \mathbf{z}, \mathbf{x})$ and $\mathbf{P}(C_i = c \mid \boldsymbol{\phi}, \mathbf{z}, \mathbf{x})$

MSRE model

$$\mathrm{P}(D_i = d \mid \boldsymbol{\phi}, \mathbf{z}) = \frac{\pi_d^D |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)' \Sigma_d^{-1} (\mathbf{b}_i - \boldsymbol{\beta}_d)\right\}}{\sum_{d=1}^{K_D} \pi_d^D |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)' \Sigma_d^{-1} (\mathbf{b}_i - \boldsymbol{\beta}_d)\right\}}$$

$$\mathrm{P}(C_i = c \mid \boldsymbol{\phi}, \mathbf{z}) = \frac{\pi_c^C \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma_i^2 - \mu_c\right)^2\right\}}{\sum_{c=1}^{K_C} \pi_c^C \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma_i^2 - \mu_c\right)^2\right\}}$$

LC model

$$\mathrm{P}(D_i = d \mid \boldsymbol{\phi}, \mathbf{z}) = \frac{\pi_d^D |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)'(\Sigma_d)^{-1}(\mathbf{b}_i - \boldsymbol{\beta}_d) - \frac{1}{2}(W_i - \theta_{C_i,d})^2\right\}}{\sum_{d=1}^{K_D} \pi_d |\Sigma_d|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\beta}_d)'(\Sigma_d)^{-1}(\mathbf{b}_i - \boldsymbol{\beta}_d) - \frac{1}{2}(W_i - \theta_{C_i,d})^2\right\}}$$

$$\mathrm{P}(C_i = c \mid \boldsymbol{\phi}, \mathbf{z}) = \frac{\pi_c^C \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma_i^2 - \mu_c\right)^2 - \frac{1}{2}(W_i - \theta_{c,D_i})^2\right\}}{\sum_{c=1}^{K_C} \pi_c^C \exp\left\{-\frac{1}{2\tau^2}\left(\log\sigma_i^2 - \mu_c\right)^2 - \frac{1}{2}(W_i - \theta_{c,D_i})^2\right\}}$$

where, $\theta_{C_i,d} = \mathbf{Z}_i^T \boldsymbol{\eta}$ in the probit latent class submodel given C_i and $D_i = d$; $\theta_{c,D_i} = \mathbf{Z}_i^T \boldsymbol{\eta}$ in the probit latent class submodel given $C_i = c$ and D_i.

Computation details to predict outcome for new validation sample

In this section, we give the details to draw $\tilde{\mathbf{z}}^{(m)}, \mathbf{z}^{(m)}, \boldsymbol{\phi}_{\text{long}}^{(m)}$ and $\boldsymbol{\eta}^{(m)}$ from the posterior distribution $p(\tilde{\mathbf{z}}, \mathbf{z}, \boldsymbol{\phi}_{\text{long}}, \boldsymbol{\eta} | \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{o}, \mathrm{H}_a)$, $m = 1, \cdots, M$ for some large M.

joint LC model:

After initialize the chain, we repeat the following steps (1) to (5) for $m = 1, \cdots, M$:

(1) update individual level latent variables $\tilde{\mathbf{z}}$ for validation sample (note: this is only conditional on the longitudinal trajectories $\tilde{\mathbf{y}} = \{\tilde{\mathbf{y}}_i\}_{i=1}^{\tilde{n}}$)

- **draw** $\tilde{D}_i^{(m)}$, $i = 1, ..., \tilde{n}$ from Multinomial$(\tilde{\pi}_{i1}^D, ..., \tilde{\pi}_{iK_D}^D)$, where

$$\tilde{\pi}_{id}^D = \frac{\pi_d^{D(m)} |\Sigma_d^{(m)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})\right\}}{\sum_{d=1}^{K_D} \pi_d^{D(m)} |\Sigma_d^{(m)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})\right\}}$$

- **draw** $\tilde{\mathbf{b}}_i^{(m)}$, $i = 1, ..., \tilde{n}$ from $\text{MVN}\left(\tilde{\beta}_i, \tilde{\Sigma}_i\right)$, where

$$\tilde{\Sigma}_i = \left((\Sigma^{(m)}_{\tilde{D}_i^{(m)}})^{-1} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{\tilde{n}_i} \mathbf{t}_{ij} \mathbf{t}'_{ij} \right)^{-1}$$

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\Sigma}_i \left[(\Sigma^{(m)}_{\tilde{D}_i^{(m)}})^{-1} \boldsymbol{\beta}^{(m)}_{\tilde{D}_i^{(m)}} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{n_i} \tilde{y}_{ij} \mathbf{t}_{ij} \right]$$

  $\mathbf{t}_{ij}$ is a vector of functional forms of the observation time t_{ij} for the longitudinal measurement $\tilde{y}_{ij}$ such that $\tilde{y}_{ij} \sim \text{N}\{f(\tilde{\mathbf{b}}_i; t_{ij}), \sigma_i^2\}$ with $f(\tilde{\mathbf{b}}_i; t_{ij}) = \tilde{\mathbf{b}}_i^T \mathbf{t}_{ij}$.

- **draw** the variance class memberships $\tilde{C}_i^{(m)}$, $i = 1, ..., \tilde{n}$ from Multinomial$(\tilde{\pi}^C_{i1}, ..., \tilde{\pi}^C_{iK_C})$, where

$$\tilde{\pi}^C_{ic} = \frac{\pi_c^{C(m)} \exp\left\{ -\frac{1}{2(\tau^{(m)})^2} \left(\log(\tilde{\sigma}_i^{(m)})^2 - \mu_c^{(m)} \right)^2 \right\}}{\sum_{c=1}^{K_C} \pi_c^{C(m)} \exp\left\{ -\frac{1}{2(\tau^{(m)})^2} \left(\log(\tilde{\sigma}_i^{(m)})^2 - \mu_c^{(m)} \right)^2 \right\}}$$

- **draw** the variances $(\tilde{\sigma}_i^{(m)})^2$, $i = 1, ..., \tilde{n}$ from

$$\pi((\tilde{\sigma}_i^{(m)})^2 | .) \propto \prod_{c=1}^{K_C} \text{LN}\left(\tilde{\sigma}_i^{(m)})^2; \mu_c^{(m)}, (\tau^{(m)})^2 \right)^{\text{I}(\tilde{C}_i^{(m)} = c)} \prod_{j=1}^{\tilde{n}_i} \text{N}\left\{ y_{ij}; f(\tilde{\mathbf{b}}_i^{(m)}; t_{ij}), (\tilde{\sigma}_i^{(m)})^2 \right\}$$

  $\text{LN}\left((\tilde{\sigma}_i^{(m)})^2; \mu_c^{(m)}, (\tau^{(m)})^2 \right)$ represents the density of log normal disribution with mean $\mu_c^{(m)}$ and variance $(\tau^{(m)})^2$ evaluated at $(\tilde{\sigma}_i^{(m)})^2$ and $\text{N}\left\{ y_{ij}; f(\tilde{\mathbf{b}}_i^{(m)}; t_{ij}), (\tilde{\sigma}_i^{(m)})^2 \right\}$ represents the density of normal distribution with mean $f(\tilde{\mathbf{b}}_i^{(m)}; t_{ij})$ and variance $(\tilde{\sigma}_i^{(m)})^2$ evaluated at $\tilde{y}_{ij}$.

(2) update individual level latent variables z for old sample (note: this is conditional on both the longitudinal trajectories $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^n$ and the outcome $\mathbf{o} = \{o_i\}_{i=1}^n$)

- **draw** the mean profile class memberships $D_i^{(m)}$, $i = 1,...,n$: from Multinomial$(\tilde{\pi}_{i1}^D,...,\tilde{\pi}_{iK_D}^D)$, where

$$\tilde{\pi}_{id}^D = \frac{\pi_d^{D(m)}|\Sigma_d^{(m)}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{b}_i^{(m)}-\boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\mathbf{b}_i^{(m)}-\boldsymbol{\beta}_d^{(m)})-\frac{1}{2}(W_i^{(m)}-\theta^{(m)}_{C_i^{(m)},d})^2\right\}}{\sum_{d=1}^{K_D}\pi_d^{D(m)}|\Sigma_d^{(m)}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{b}_i^{(m)}-\boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\mathbf{b}_i^{(m)}-\boldsymbol{\beta}_d^{(m)})-\frac{1}{2}(W_i^{(m)}-\theta^{(m)}_{C_i^{(m)},d})^2\right\}}$$

  $\theta^{(m)}_{C_i^{(m)},d} = (\mathbf{Z}_i^{(m)})^T\boldsymbol{\eta}^{(m)}$ in the latent class probit submodel given $D_i^{(m)} = d$ and $C_i^{(m)}$.

- **draw** the random effects $\mathbf{b}_i$, $i = 1,...,n$ from $\sim$ MVN$\left(\tilde{\beta}_i, \tilde{\Sigma}_i\right)$, where

$$\tilde{\Sigma}_{id} = \left[(\Sigma_{D_i}^{(m)})^{-1} + \frac{1}{(\sigma_i^{(m)})^2}\sum_{j=1}^{n_i}\mathbf{t}_{ij}\mathbf{t}_{ij}'\right]^{-1}$$

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\Sigma}_{id}\left[(\Sigma_{D_i}^{(m)})^{-1}\boldsymbol{\beta}_{D_i}^{(m)} + \frac{1}{(\sigma_i^{(m)})^2}\sum_{j=1}^{n_i}y_{ij}\mathbf{t}_{ij}\right]$$

  $\mathbf{t}_{ij}$ is a vector of functional forms of the observation time t_{ij} for the longitudinal measurement y_{ij} such that $y_{ij} \sim \mathrm{N}\{f(\mathbf{b}_i;t_{ij}),\sigma_i^2\}$ with $f(\mathbf{b}_i;t_{ij}) = \mathbf{b}_i^T\mathbf{t}_{ij}$.

- **draw** the variance class memberships $C_i^{(m)}$, $i = 1,...,n$ from Multinomial$(\tilde{\pi}_{i1}^C,...,\tilde{\pi}_{iK_C}^C)$ where

$$\tilde{\pi}_{ic}^C = \frac{\pi_c^{C(m)}\exp\left\{-\frac{1}{2(\tau^{(m)})^2}\left(\log(\sigma_i^{(m)})^2-\mu_c^{(m)}\right)^2-\frac{1}{2}(W_i^{(m)}-\theta^{(m)}_{c,D_i^{(m)}})^2\right\}}{\sum_{c=1}^{K_C}\pi_c^{C(m)}\exp\left\{-\frac{1}{2(\tau^{(m)})^2}\left(\log(\sigma_i^{(m)})^2-\mu_c^{(m)}\right)^2-\frac{1}{2}(W_i^{(m)}-\theta^{(m)}_{c,D_i^{(m)}})^2\right\}}$$

  $\theta^{(m)}_{c,D_i^{(m)}} = (\mathbf{Z}_i^{(m)})^T\boldsymbol{\eta}^{(m)}$ in the probit latent class submodel given $C_i^{(m)} = c$ and $D_i^{(m)}$.

- **update** the variances $(\sigma_i^{(m)})^2$, $i = 1,...,n$

$$\pi((\sigma_i^{(m)})^2|.) \propto \prod_{c=1}^{K_C}\mathrm{LN}\left((\sigma_i^{(m)})^2;\mu_c^{(m)},(\tau^{(m)})^2\right)^{\mathrm{I}(C_i^{(m)}=c)}\prod_{j=1}^{n_i}\mathrm{N}\left\{y_{ij};f(\mathbf{b}_i^{(m)};t_{ij}),(\sigma_i^{(m)})^2\right\}$$

$\text{LN}\left((\sigma_i^{(m)})^2; \mu_c^{(m)}, (\tau^{(m)})^2\right)$ represents the density of log normal disribution with mean $\mu_c^{(m)}$ and variance $(\tau^{(m)})^2$ evaluated at $(\sigma_i^{(m)})^2$ and $\text{N}\left\{y_{ij}; f(\mathbf{b}_i^{(m)}; t_{ij}), (\sigma_i^{(m)})^2\right\}$ represents the density of normal distribution with mean $f(\mathbf{b}_i^{(m)}; t_{ij})$ and variance $(\sigma_i^{(m)})^2$ evaluated at y_{ij}. Since there is no closed form of the full conditional posterior density, the draws for $(\sigma_i^{(m)})^2$, $i = 1, ..., n$ at each iteration of the Gibbs sampling are obtained using the inverse cumulative distribution sampling method.

- **draw** W_i, $i = 1, ..., n$ from:

$$[W_i^{(m)} | o_i = 1, \cdot] \sim \text{N}(\theta_{C_i^{(m)}, D_i^{(m)}}, 1) I_{(0,\infty)}(\cdot)$$
$$[W_i^{(m)} | o_i = 0, \cdot] \sim \text{N}(\theta_{C_i^{(m)}, D_i^{(m)}}, 1) I_{(-\infty,0)}(\cdot)$$

where, $\theta_{C_i^{(m)}, D_i^{(m)}} = (\mathbf{Z}_i^{(m)})^T \boldsymbol{\eta}^{(m)}$ in the latent class probit submodel given $C_i^{(m)}$ and $D_i^{(m)}$.

(3) update population level parameter ϕ_{long} in the longitudinal submodel

- **draw** $\boldsymbol{\beta}_d^{(m)}$, $d = 1, \cdots, K_D$: Assuming the prior for $\boldsymbol{\beta}_d \overset{ind}{\sim} \text{MVN}(\boldsymbol{\nu}, \mathbf{V})$, then the full conditional posterior density for $\boldsymbol{\beta}_d$ for $d = 1, ..., K_D$ is $[\boldsymbol{\beta}_d | .] \sim \text{MVN}\left(\tilde{\boldsymbol{\nu}}_d, \tilde{\mathbf{V}}_d\right)$ where

$$\tilde{\nu}_d = \tilde{\mathbf{V}}_d \left\{ \boldsymbol{\nu} + (\Sigma_d^{(m)})^{-1} \left(\sum_{i=1}^{n} \text{I}(D_i^{(m)} = d) \mathbf{b}_i^{(m)} + \sum_{i=1}^{\tilde{n}} \text{I}(\tilde{D}_i^{(m)} = d) \tilde{\mathbf{b}}_i^{(m)} \right) \right\}$$
$$\tilde{\mathbf{V}}_d = \left\{ \mathbf{V}^{-1} + (\Sigma_d^{(m)})^{-1} \left(\sum_{i=1}^{n} \text{I}(D_i^{(m)} = d) + \sum_{i=1}^{\tilde{n}} \text{I}(\tilde{D}_i^{(m)} = d) \right) \right\}^{-1}$$

- **draw** $\Sigma_d^{(m)}$, $d = 1, \cdots, K_D$: Assuming the prior for $\Sigma_d \overset{ind}{\sim}$ Inverse-Wishart(m, Λ), then the full conditional posterior density is, $[\Sigma_d | \cdot] \sim$ Inverse-Wishart$\left(\tilde{m}_d, \tilde{\Lambda}_d\right)$

where

$$\tilde{m}_d = m + \sum_{i=1}^{n} \mathrm{I}(D_i^{(m)} = d) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{D}_i^{(m)} = d)$$

$$\tilde{\Lambda}_d = \left\{ \Lambda^{-1} + \sum_{i=1}^{n} \mathrm{I}(D_i^{(m)} = d) \left(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right) \left(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right)' + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{D}_i^{(m)} = d) \left(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right) \left(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right)' \right\}^{-1}$$

- **update** the mixing proportion $\{\pi_d^{D(m)}\}_d$: assuming $\{\pi_d^D\}_d \sim \text{Dirichlet}(e_1^D, ..., e_{K_D}^D)$ then the full conditional posterior distribution is

$$\left[\{\pi_d^D\}_d | \cdot\right] \sim \text{Dirichlet}(\{e_d^D + \sum_{i=1}^{n} \mathrm{I}(D_i^{(m)} = d) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{D}_i^{(m)} = d)\}_d).$$

- **update** $\mu_c^{(m)}$: assuming the prior for $\mu_c \overset{ind}{\sim} \mathrm{N}(a, b)$, then the full conditional posterior distribution is, $\left[\mu_c | \cdot\right] \sim \mathrm{N}\left(\tilde{a}, \tilde{b}\right)$ where

$$\tilde{a} = \frac{(\tau^{(m)})^{-2} \left(\sum_{i=1}^{n} \mathrm{I}(C_i^{(m)} = c) \log \sigma_i^2 + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{C}_i^{(m)} = c) \log \tilde{\sigma}_i^2 \right) + a/b}{1/b + (\tau^{(m)})^{-2} \left(\sum_{i=1}^{n} \mathrm{I}(C_i^{(m)} = c) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{C}_i^{(m)} = c) \right)}$$

$$\tilde{b} = \left\{ 1/b + (\tau^{(m)})^{-2} \left(\sum_{i=1}^{n} \mathrm{I}(C_i^{(m)} = c) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{C}_i^{(m)} = c) \right) \right\}^{-1}$$

- **draw** $(\tau^{(m)})^2$: assuming $\tau^2 \sim$ Inverse-Gamma(v, e), then the full conditional posterior distribution is $\left[\tau^2 | \cdot\right] \sim$ Inverse-Gamma$\left\{v + \frac{n+\tilde{n}}{2}, \tilde{e}\right\}$ where $\tilde{e} = e + \sum_{c=1}^{K_C} \left(\sum_{i=1}^{n} \frac{1}{2} \mathrm{I}(C_i^{(m)} = c) \left(\log(\sigma_i^{(m)})^2 - \mu_c^{(m)} \right)^2 + \sum_{i=1}^{\tilde{n}} \frac{1}{2} \mathrm{I}(\tilde{C}_i^{(m)} = c) \left(\log(\tilde{\sigma}^{(m)})_i^2 - \mu_c^{(m)} \right)^2 \right)$.

- **update** the mixing proportion $\{\pi_c^{C(m)}\}_c$: assuming $\{\pi_c^C\}_c \sim \text{Dirichlet}(e_1^C, ..., e_{K_C}^C)$ then the full conditional posterior distribution is

$$\left[\{\pi_c^C\}_c|\cdot\right] \sim \text{Dirichlet}(\{e_c^C + \sum_{i=1}^{n} \text{I}(C_i^{(m)} = c) + \sum_{i=1}^{\tilde{n}} \text{I}(\tilde{C}_i^{(m)} = c)\}_c).$$

(4) update the parameter η in the primary outcome model

- **draw $\boldsymbol{\eta}^{(m)}$**: Assuming the prior for $\boldsymbol{\eta} \sim \text{MVN}(\boldsymbol{\nu}_{\boldsymbol{\eta}}, \mathbf{V}_{\boldsymbol{\eta}})$, then the full conditional posterior density for $\boldsymbol{\eta}^{(m)}$ is $\left[\boldsymbol{\eta}|.\right] \sim \text{MVN}\left(\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}}, \tilde{\mathbf{V}}_{\boldsymbol{\eta}}\right)$ where

$$\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}} = \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n} \mathbf{Z}_i^{(m)} (\mathbf{Z}^{(m)})_i^T\right)^{-1} \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} \boldsymbol{\nu}_{\boldsymbol{\eta}} + \sum_{i=1}^{n} W_i^{(m)} \mathbf{Z}_i^{(m)}\right)$$
$$\tilde{\mathbf{V}}_{\boldsymbol{\eta}} = \left(\mathbf{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n} \mathbf{Z}_i^{(m)} (\mathbf{Z}^{(m)})_i^T\right)^{-1}$$

$\mathbf{Z}_i^{(m)}$ is the i^{th} row of the design matrix in the probit submodel given $D_i^{(m)}$ and $C_i^{(m)}$.

(5) The prediction of outcome for a new validation sample i, $i = 1, \cdots, \tilde{n}$ can be based on $\tilde{p}_i^{(m)} = \Phi(\boldsymbol{\eta}^{(m)T} \tilde{\mathbf{Z}}_i^{(m)})$, where $\tilde{\mathbf{Z}}_i^{(m)}$ contains $\tilde{D}_i^{(m)}$ and $\tilde{C}_i^{(m)}$ for LC model.

joint MSRE model:

After initialize the chain, we repeat the following steps (1) to (5) for $m = 1, \cdots, M$:

(1) update individual level latent variables $\tilde{\mathbf{z}}$ for validation sample (note: this is only conditional on the longitudinal trajectories $\tilde{\mathbf{y}} = \{\tilde{\mathbf{y}}_i\}_{i=1}^{\tilde{n}}$)

- **draw** $\tilde{D}_i^{(m)}$, $i = 1, ..., \tilde{n}$ from Multinomial$(\tilde{\pi}_{i1}^D, ..., \tilde{\pi}_{iK_D}^D)$, where

$$\tilde{\pi}_{i1}^D = \frac{\pi_d^{D(m)} |\Sigma_d^{(m)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})\right\}}{\sum_{d=1}^{K_D} \pi_d^{D(m)} |\Sigma_d^{(m)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})\right\}}$$

- **draw** $\tilde{\mathbf{b}}_i^{(m)}$, $i = 1, ..., \tilde{n}$ from $\text{MVN}\left(\tilde{\beta}_i, \tilde{\Sigma}_i\right)$, where

$$\tilde{\Sigma}_i = \left[(\Sigma^{(m)}_{\tilde{D}_i^{(m)}})^{-1} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{\tilde{n}_i} \mathbf{t}_{ij} \mathbf{t}_{ij}' \right]^{-1}$$

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\Sigma}_i \left[(\Sigma^{(m)}_{\tilde{D}_i^{(m)}})^{-1} \boldsymbol{\beta}^{(m)}_{\tilde{D}_i^{(m)}} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{n_i} \tilde{y}_{ij} \mathbf{t}_{ij} \right]$$

  $\mathbf{t}_{ij}$ is a vector of functional forms of the observation time t_{ij} for the longitudinal measurement $\tilde{y}_{ij}$ such that $\tilde{y}_{ij} \sim \text{N}\{f(\tilde{\mathbf{b}}_i; t_{ij}), \sigma_i^2\}$ with $f(\tilde{\mathbf{b}}_i; t_{ij}) = \tilde{\mathbf{b}}_i^T \mathbf{t}_{ij}$.

- **draw** the variance class memberships $\tilde{C}_i^{(m)}$, $i = 1, ..., \tilde{n}$ from Multinomial$(\tilde{\pi}^C_{i1}, ..., \tilde{\pi}^C_{iK_C})$, where

$$\tilde{\pi}^C_{i1} = \frac{\pi_c^{C(m)} \exp\left\{ -\frac{1}{2(\tau^{(m)})^2} \left(\log(\tilde{\sigma}_i^{(m)})^2 - \mu_c^{(m)} \right)^2 \right\}}{\sum_{c=1}^{K_C} \pi_c^{C(m)} \exp\left\{ -\frac{1}{2(\tau^{(m)})^2} \left(\log(\tilde{\sigma}_i^{(m)})^2 - \mu_c^{(m)} \right)^2 \right\}}$$

- **draw** the variances $(\tilde{\sigma}_i^{(m)})^2$, $i = 1, ..., \tilde{n}$ from

$$\pi((\tilde{\sigma}_i^{(m)})^2 | .) \propto \prod_{c=1}^{K_C} \text{LN}\left(\tilde{\sigma}_i^{(m)})^2; \tilde{\mu}_c^{(m)}, (\tau^{(m)})^2 \right)^{\text{I}(\tilde{C}_i^{(m)} = c)} \prod_{j=1}^{\tilde{n}_i} \text{N}\left\{ y_{ij}; f(\tilde{\mathbf{b}}_i^{(m)}; t_{ij}), (\tilde{\sigma}_i^{(m)})^2 \right\}$$

  $\text{LN}\left((\tilde{\sigma}_i^{(m)})^2; \mu_c^{(m)}, (\tau^{(m)})^2 \right)$ represents the density of log normal disribution with mean $\mu_c^{(m)}$ and variance $(\tau^{(m)})^2$ evaluated at $(\tilde{\sigma}_i^{(m)})^2$ and $\text{N}\left\{ y_{ij}; f(\tilde{\mathbf{b}}_i^{(m)}; t_{ij}), (\tilde{\sigma}_i^{(m)})^2 \right\}$ represents the density of normal distribution with mean $f(\tilde{\mathbf{b}}_i^{(m)}; t_{ij})$ and variance $(\tilde{\sigma}_i^{(m)})^2$ evaluated at $\tilde{y}_{ij}$.

(2) update individual level latent variables z for old sample (note: this is conditional on both the longitudinal trajectories $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^n$ and the outcome $\mathbf{o} = \{o_i\}_{i=1}^n$)

- **draw** the mean profile class memberships $D_i^{(m)}$ from Multinomial($\tilde{\pi}_{i1}^D, ..., \tilde{\pi}_{iK_D}^D$), where

$$\tilde{\pi}_{i1}^D = \frac{\pi_d^{D(m)} |\Sigma_d^{(m)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})\right\}}{\sum_{d=1}^{K_D} \pi_d^{D(m)} |\Sigma_d^{(m)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})'(\Sigma_d^{(m)})^{-1}(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)})\right\}}$$

- **draw** the random effects $\mathbf{b}_i$, $i = 1, ..., n$ from $\sim \text{MVN}\left(\tilde{\beta}_i, \tilde{\Sigma}_i\right)$, where

$$\tilde{\Sigma}_{id} = \left((\Sigma_{D_i^{(m)}}^{(m)})^{-1} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{n_i} \mathbf{t}_{ij} \mathbf{t}_{ij}^T + \mathbf{g}(\boldsymbol{\eta}^{(m)}, \sigma_i^{(m)}) \mathbf{g}(\boldsymbol{\eta}^{(m)}, \sigma_i^{(m)})^T \right)^{-1}$$

$$\tilde{\boldsymbol{\beta}}_i^{(m)} = \tilde{\Sigma}_{id} \left((\Sigma_{D_i}^{(m)})^{-1} \boldsymbol{\beta}_{D_i}^{(m)} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{n_i} y_{ij} \mathbf{t}_{ij} + (W_i^{(m)} - \tilde{g}(\boldsymbol{\eta}^{(m)}, \sigma_i^{(m)})) \mathbf{g}(\boldsymbol{\eta}^{(m)}, \sigma_i^{(m)}) \right)$$

where $\mathbf{t}_{ij}$ is a vector of functional forms of the observation time t_{ij} for the longitudinal measurement y_{ij} such that $y_{ij} \sim \text{N}\{f(\mathbf{b}_i; t_{ij}), \sigma_i^2\}$ with $f(\mathbf{b}_i; t_{ij}) = \mathbf{b}_i^T \mathbf{t}_{ij}$. $\mathbf{g}(\boldsymbol{\eta}, \sigma_i)$ is a vector such that $\mathbf{Z}_i^T \boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\eta}, \sigma_i)^T \mathbf{b}_i + \tilde{g}(\boldsymbol{\eta}, \sigma_i)$ in the MSRE primary outcome model. $\mathbf{Z}_i$ is the i^{th} row of the design matrix in the probit submodel given $\mathbf{b}_i$ and σ_i.

- **draw** the variance class memberships $C_i^{(m)}$, $i = 1, ..., n$ from Multinomial($\tilde{\pi}_{i1}^C, ..., \tilde{\pi}_{iK_C}^C$) where

$$\tilde{\pi}_{i1}^C = \frac{\pi_c^{C(m)} \exp\left\{-\frac{1}{2(\tau^{(m)})^2}\left(\log(\sigma_i^{(m)})^2 - \mu_c^{(m)}\right)^2\right\}}{\sum_{c=1}^{K_C} \pi_c^{C(m)} \exp\left\{-\frac{1}{2(\tau^{(m)})^2}\left(\log(\sigma_i^{(m)})^2 - \mu_c^{(m)}\right)^2\right\}}$$

- **update** the variances $(\sigma_i^{(m)})^2$, $i = 1, ..., n$

$$\pi((\sigma_i^{(m)})^2 | .) \propto \prod_{c=1}^{K_C} \text{LN}\left((\sigma_i^{(m)})^2; \mu_c^{(m)}, (\tau^{(m)})^2\right)^{\text{I}(C_i^{(m)} = c)} \prod_{j=1}^{n_i} \text{N}\left\{y_{ij}; f(\mathbf{b}_i^{(m)}; t_{ij}), (\sigma_i^{(m)})^2\right\}$$
$$\times \text{N}(W_i^{(m)}; (\mathbf{Z}_i^{(m)})^T \boldsymbol{\eta}^{(m)}, 1)$$

$\text{LN}\left((\sigma_i^{(m)})^2;\mu_c^{(m)},(\tau^{(m)})^2\right)$ represents the density of log normal disribution with mean $\mu_c^{(m)}$ and variance $(\tau^{(m)})^2$ evaluated at $(\sigma_i^{(m)})^2$ and $\text{N}\left\{y_{ij};f(\mathbf{b}_i^{(m)};t_{ij}),(\sigma_i^{(m)})^2\right\}$ represents the density of normal distribution with mean $f(\mathbf{b}_i^{(m)};t_{ij})$ and variance $(\sigma_i^{(m)})^2$ evaluated at y_{ij} and similarly $\text{N}(W_i^{(m)};(\mathbf{Z}_i^{(m)})^T\boldsymbol{\eta}^{(m)},1)$ represents the density of normal distribution with mean $(\mathbf{Z}_i^{(m)})^T\boldsymbol{\eta}^{(m)}$ and variance 1 evaluated at $W_i^{(m)}$.

- **draw** W_i, $i = 1,...,n$ for old sample from:

$$[W_i^{(m)}|o_i = 1,\cdot] \sim \text{N}((\mathbf{Z}_i^{(m)})^T\boldsymbol{\eta}^{(m)},1)I_{(0,\infty)}(\cdot)$$
$$[W_i^{(m)}|o_i = 0,\cdot] \sim \text{N}((\mathbf{Z}_i^{(m)})^T\boldsymbol{\eta}^{(m)},1)I_{(-\infty,0)}(\cdot)$$

where, $\mathbf{Z}_i^{(m)}$ contains $\mathbf{b}_i^{(m)}$ and $\sigma_i^{(m)}$ for joint MSRE model.

(3) update population level parameter $\phi_{\textbf{long}}$ in the longitudinal submodel

- **draw** $\boldsymbol{\beta}_d^{(m)}$, $d = 1,\cdots,K_D$: Assuming the prior for $\boldsymbol{\beta}_d \overset{ind}{\sim} \text{MVN}(\boldsymbol{\nu},\mathbf{V})$, then the full conditional posterior density for $\boldsymbol{\beta}_d$ for $d = 1,...,K_D$ is $\left[\boldsymbol{\beta}_d|.\right] \sim \text{MVN}\left(\tilde{\boldsymbol{\nu}}_d,\tilde{\mathbf{V}}_d\right)$ where

$$\tilde{\nu}_d = \tilde{\mathbf{V}}_d\left\{\boldsymbol{\nu} + (\Sigma_d^{(m)})^{-1}\left(\sum_{i=1}^{n}\text{I}(D_i^{(m)} = d)\mathbf{b}_i^{(m)} + \sum_{i=1}^{\tilde{n}}\text{I}(\tilde{D}_i^{(m)} = d)\tilde{\mathbf{b}}_i^{(m)}\right)\right\}$$
$$\tilde{\mathbf{V}}_d = \left\{\mathbf{V}^{-1} + (\Sigma_d^{(m)})^{-1}\left(\sum_{i=1}^{n}\text{I}(D_i^{(m)} = d) + \sum_{i=1}^{\tilde{n}}\text{I}(\tilde{D}_i^{(m)} = d)\right)\right\}^{-1}$$

- **draw** $\Sigma_d^{(m)}$, $d = 1,\cdots,K_D$: Assuming the prior for $\Sigma_d \overset{ind}{\sim}$ Inverse-Wishart(m,Λ), then the full conditional posterior density is, $[\Sigma_d|\cdot] \sim$ Inverse-Wishart$\left(\tilde{m}_d,\tilde{\Lambda}_d\right)$

where

$$\tilde{m}_d = m + \sum_{i=1}^{n} \mathrm{I}(D_i^{(m)} = d) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{D}_i^{(m)} = d)$$

$$\tilde{\Lambda}_d = \left\{ \Lambda^{-1} + \sum_{i=1}^{n} \mathrm{I}(D_i^{(m)} = d) \left(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right) \left(\mathbf{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right)' + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{D}_i^{(m)} = d) \left(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right) \left(\tilde{\mathbf{b}}_i^{(m)} - \boldsymbol{\beta}_d^{(m)}\right)' \right\}^{-1}$$

- **update** the mixing proportion $\{\pi_d^{D(m)}\}_d$: assuming $\{\pi_d^D\}_d \sim \text{Dirichlet}(e_1^D, ..., e_{K_D}^D)$ then the full conditional posterior distribution is $\left[\{\pi_d^D\}_d | \cdot\right] \sim \text{Dirichlet}(\{e_d^D + \sum_{i=1}^{n} \mathrm{I}(D_i^{(m)} = d) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{D}_i^{(m)} = d)\}_d)$.

- **update** $\mu_c^{(m)}$: assuming the prior for $\mu_c \overset{ind}{\sim} \mathrm{N}(a, b)$, then the full conditional posterior distribution is, $\left[\mu_c | \cdot\right] \sim \mathrm{N}\left(\tilde{a}, \tilde{b}\right)$ where

$$\tilde{a} = \frac{(\tau^{(m)})^{-2} \left(\sum_{i=1}^{n} \mathrm{I}(C_i^{(m)} = c) \log \sigma_i^2 + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{C}_i^{(m)} = c) \log \tilde{\sigma}_i^2 \right) + a/b}{1/b + (\tau^{(m)})^{-2} \left(\sum_{i=1}^{n} \mathrm{I}(C_i^{(m)} = c) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{C}_i^{(m)} = c) \right)}$$

$$\tilde{b} = \left\{ 1/b + (\tau^{(m)})^{-2} \left(\sum_{i=1}^{n} \mathrm{I}(C_i^{(m)} = c) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{C}_i^{(m)} = c) \right) \right\}^{-1}$$

- **draw** $(\tau^{(m)})^2$: assuming $\tau^2 \sim \text{Inverse-Gamma}(v, e)$, then the full conditional posterior distribution is $\left[\tau^2 | \cdot\right] \sim \text{Inverse-Gamma}\left\{v + \frac{n+\tilde{n}}{2}, \tilde{e}\right\}$, where $\tilde{e} = e + \sum_{c=1}^{K_C} \left(\sum_{i=1}^{n} \frac{1}{2} \mathrm{I}(C_i^{(m)} = c) \left(\log(\sigma_i^{(m)})^2 - \mu_c^{(m)} \right)^2 + \sum_{i=1}^{\tilde{n}} \frac{1}{2} \mathrm{I}(\tilde{C}_i^{(m)} = c) \left(\log(\tilde{\sigma}^{(m)})_i^2 - \mu_c^{(m)} \right)^2 \right)$

- **update** the mixing proportion $\{\pi_c^{C(m)}\}_c$: assuming $\{\pi_c^C\}_c \sim \text{Dirichlet}(e_1^C, ..., e_{K_C}^C)$ then the full conditional posterior distribution is

$$\left[\{\pi_c^C\}_c | \cdot\right] \sim \text{Dirichlet}\left(\left\{ e_c^C + \sum_{i=1}^{n} \mathrm{I}(C_i^{(m)} = c) + \sum_{i=1}^{\tilde{n}} \mathrm{I}(\tilde{C}_i^{(m)} = c) \right\}_c \right).$$

(4) update the parameter η in the primary outcome model

- **draw $\boldsymbol{\eta}^{(m)}$**: Assuming the prior for $\boldsymbol{\eta} \sim \text{MVN}(\boldsymbol{\nu}_{\boldsymbol{\eta}}, \mathbf{V}_{\boldsymbol{\eta}})$, then the full conditional posterior density for $\boldsymbol{\eta}^{(m)}$ is $[\boldsymbol{\eta}|.] \sim \text{MVN}\left(\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}}, \tilde{\mathbf{V}}_{\boldsymbol{\eta}}\right)$ where

$$\tilde{\mathbf{V}}_{\boldsymbol{\eta}} = \left[\mathbf{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n} \mathbf{Z}_i^{(m)} (\mathbf{Z}^{(m)})_i^T\right]^{-1}$$

$$\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}} = \tilde{\mathbf{V}}_{\boldsymbol{\eta}} \left[\mathbf{V}_{\boldsymbol{\eta}}^{-1} \boldsymbol{\nu}_{\boldsymbol{\eta}} + \sum_{i=1}^{n} W_i^{(m)} \mathbf{Z}_i^{(m)}\right]$$

$\mathbf{Z}_i^{(m)}$ is the i^{th} row of the design matrix in the probit submodel and contains $\mathbf{b}_i^{(m)}$ and $\sigma_i^{(m)}$ for MSRE model.

(5) The prediction of outcome for a new validation sample i, $i = 1, \cdots, \tilde{n}$ can be based on $\tilde{p}_i^{(m)} = \Phi(\boldsymbol{\eta}^{(m)T} \tilde{\mathbf{Z}}_i^{(m)})$, where $\tilde{\mathbf{Z}}_i^{(m)}$ contains $\tilde{\mathbf{b}}_i^{(m)}$ and $\tilde{\sigma}_i^{(m)}$ for MSRE model.

BIBLIOGRAPHY

BIBLIOGRAPHY

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in *Proceedings of the Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, pp. 267–281, Budapest: Akademiai Kiado.

Albert, J. H., and S. Chib (1993), Bayesian analysis of binary and polychotomous response data, *Journal of the American statistical Association*, *88*(422), 669–679.

Biernacki, C., G. Celeux, and G. Govaert (2000), Assessing a mixture model for clustering with the integrated completed likelihood, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(7), 719–725.

Brown, E. R., and J. G. Ibrahim (2003a), A bayesian semiparametric joint hierarchical model for longitudinal and survival data, *Biometrics*, *59*(2), 221–228.

Brown, E. R., and J. G. Ibrahim (2003b), Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials, *Biometrics*, *59*(3), 686–693.

Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterington (2006), Deviance information criteria for missing data models, *Bayesian Analysis*, *1*(4), 651–673.

Chen, H., and Y. Wang (2011), A penalized spline approach to functional mixed effects model analysis, *Biometrics*, *67*(3), 861–870.

Chen, M.-H., and P. Gustafson (2011), editors, *Lifetime Data Analysis*, *17*, special Issue.

Day, N. E. (1969), Estimating the components of a mixture of normal distributions, *Biometrika*, *56*(3), 463–474.

Durbán, M., J. Harezlak, M. P. Wand, and R. J. Carroll (2005), Simple fitting of subject-specific curves for longitudinal data, *Statistics in medicine*, *24*(8), 1153–1167.

Eilers, P. H. C., and B. D. Marx (1996), Flexible smoothing with B-splines and penalties, *Statistical Science*, *11*(2), 89–121.

Elliott, M. R. (2007), Identifying latent clusters of variability in longitudinal data, *Biostatistics*, *8*(4), 756–771.

Elliott, M. R., M. D. Sammel, and J. Faul (2012), Associations between variability of risk factors and health outcomes in longitudinal studies, *Statistics in Medicine*, *31*(23), 2745–2756.

Fawcett, T. (2006), An introduction to ROC analysis, *Pattern recognition letters*, *27*(8), 861–874.

Fonseca, T. C., M. A. Ferreira, and H. S. Migon (2008), Objective Bayesian analysis for the student-t regression model, *Biometrika*, *95*(2), 325–333.

Freeman, E. W., M. D. Sammel, J. A. Grisso, M. Battistini, B. Garcia-Espagna, and L. Hollander (2001), Hot flashes in the late reproductive years: risk factors for African American and Caucasian women, *Journal of Women's Health and Gender-Based Medicine*, *10*(1), 67–76.

Freeman, E. W., M. D. Sammel, H. Lin, and D. B. Nelson (2006), Associations of hormones and menopausal status with depressed mood in women with no history of depression, *Archives of General Psychiatry*, *63*(4), 375.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, New York: Springer.

Garrett, E. S., and S. L. Zeger (2000), Latent class model diagnosis, *Biometrics*, *56*(4), 1055–1067.

Gelman, A., and J. Hill (2007), *Data Analysis using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press.

Gelman, A., X.-L. Meng, and H. Stern (1996), Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica*, *6*(4), 733–760.

Gelman, A., Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen (2000), Diagnostic checks for discrete data regression models using posterior predictive simulations, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *49*(2), 247–268.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003), *Bayesian data analysis*, CRC press.

Gelman, A., J. Hwang, and A. Vehtari (2013), Understanding predictive information criteria for Bayesian models, *Statistics and Computing*, dOI: 10.1007/s11222-013-9416-2.

Geweke, J. (1993), Bayesian treatment of the independent student-t linear model, *Journal of Applied Econometrics*, *8*(S1), S19–S40.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2001), *Bayesian survival analysis*, New York: Springer-Verlag.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2004), Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials, *Statistica Sinica*, *14*(3), 863–884.

Ibrahim, J. G., H. Chu, and L. M. Chen (2010), Basic concepts and methods for joint models of longitudinal and survival data, *Journal of Clinical Oncology*, *28*(16), 2796–2801.

James, G. M. (2002), Generalized linear models with functional predictors, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 411–432.

Jeffreys, H. (1973), *Scientific Inference*, third ed., New York: Cambridge University Press.

Johnson, V. E. (2004), A Bayesian $\chi 2$ test for goodness-of-fit, *The Annals of Statistics*, *32*(6), 2361–2384.

Johnson, V. E. (2007), Bayesian model assessment using pivotal quantities, *Bayesian Analysis*, *2*(4), 719–734.

Kass, R. E., and R. Natarajan (2006), A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper), *Bayesian Analysis*, *1*(3), 535–542.

Lang, S., and A. Brezger (2004), Bayesian P-splines, *Journal of Computational and Graphical statistics*, *13*(1), 183–212.

Lange, K. L., R. J. Little, and J. M. Taylor (1989), Robust statistical modeling using the t distribution, *Journal of the American Statistical Association*, *84*(408), 881–896.

Law, N. J., J. M. Taylor, and H. Sandler (2002), The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure, *Biostatistics*, *3*(4), 547–563.

Little, R. J., and D. B. Rubin (2002), Statistical analysis with missing data, *New York: Wiley*.

Manson, J. M., M. D. Sammel, E. W. Freeman, and J. A. Grisso (2001), Racial differences in sex hormone levels in women approaching the transition to menopause, *Fertility and sterility*, *75*(2), 297–304.

McLachlan, G., and D. Peel (2000), *Finite Mixture Models*, New York: Wiley.

Muthén, B., and K. Shedden (1999), Finite mixture modeling with mixture outcomes using the EM algorithm, *Biometrics*, *55*(2), 463–469.

Neelon, B., A. J. O'Malley, and S.-L. T. Normand (2011), A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity, *Biometrics*, *67*(1), 280–289.

Pemstein, D., K. M. Quinn, and A. D. Martin (2007), The scythe statistical library: An open source C++ library for statistical computation., *Journal of Statistical Software*, *1*, 29.

Plummer, M. (2002), Discussion of the paper by Spiegelhalter et al., *Journal of the Royal Statistical Society: Statistical Methodology*, *4*(64), 620.

Plummer, M. (2006), Comment on article by Celeux et al., *Bayesian Analysis*, *4*(1), 681–686.

Plummer, M. (2008), Penalized loss functions for Bayesian model comparison, *Biostatistics*, *9*(3), 523–539.

Proust-Lima, C., and J. M. Taylor (2009), Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment psa: a joint modeling approach, *Biostatistics*, *10*(3), 535–549.

Proust-Lima, C., M. Séne, J. M. Taylor, and H. Jacqmin-Gadda (2012), Joint latent class models for longitudinal and time-to-event data: a review, *Statistical Methods in Medical Research*.

Ramsay, J. O., and C. Dalzell (1991), Some tools for functional data analysis, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 539–572.

Redner, R. A., and H. F. Walker (1984), Mixture densities, maximum likelihood and the EM algorithm, *SIAM review*, *26*(2), 195–239.

Richardson, S., and P. J. Green (1997), On *b*ayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society: series B (statistical methodology)*, *59*(4), 731–792.

Rizopoulos, D. (2011), Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data, *Biometrics*, *67*(3), 819–829.

Rizopoulos, D. (2012), *Joint Models for Longitudinal and Time-to-Event Data: with Applications in R*, CRC Press.

Rizopoulos, D., and P. Ghosh (2011), A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event, *Statistics in medicine*, *30*(12), 1366–1380.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011), proc: an open-source package for R and S+ to analyze and compare ROC curves, *BMC bioinformatics*, *12*(1), 77.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003), *Semiparametric Regression*, New York: Cambridge University Press.

Schwarz, G. (1978), Estimating the dimension of a model, *The annals of statistics*, *6*(2), 461–464.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005), ROCR: visualizing classifier performance in R, *Bioinformatics*, *21*(20), 3940–3941.

Song, X., M. Davidian, and A. A. Tsiatis (2002), A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data, *Biometrics*, *58*(4), 742–753.

Sowers, M. R., H. Zheng, D. McConnell, B. Nan, S. Harlow, and J. F. Randolph (2008), Follicle stimulating hormone and its rate of change in defining menopause transition stages, *Journal of Clinical Endocrinology & Metabolism*, *93*(10), 3958–3964.

Speigelhalter, D., N. Best, B. Carlin, and A. Van der Linde (2003), Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society, Series B*, *64*(4), 583–616.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002), Measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Steel, R. J., and A. E. Raftery (2010), Performance of Bayesian model selection criteria for Gaussian mixture models, in *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, edited by M.-H. Chen, P. Müller, D. Sun, K. Ye, and D. K. Dey, pp. 113–130, New York: Springer.

Stephens, M. (2000), Dealing with label switching in mixture models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Steyerberg, E. W., A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan (2010), Assessing the performance of prediction models: a framework for some traditional and novel measures, *Epidemiology (Cambridge, Mass.)*, *21*(1), 128.

Taylor, J. M., M. Yu, and H. M. Sandler (2005), Individualized predictions of disease progression following radiation therapy for prostate cancer, *Journal of clinical oncology*, *23*(4), 816–825.

Taylor, J. M., Y. Park, D. P. Ankerst, C. Proust-Lima, S. Williams, L. Kestin, K. Bae, T. Pickles, and H. Sandler (2013), Real-time individual predictions of prostate cancer recurrence using joint models, *Biometrics*.

Tsiatis, A., V. Degruttola, and M. Wulfsohn (1995), Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids, *Journal of the American Statistical Association*, *90*(429), 27–37.

van der Linde, A. (2012), A Bayesian view of model complexity, *Statistica Neerlandica*, *66*(3), 253–271.

Vehtari, A., and J. Ojanen (2012), A survey of Bayesian predictive methods for model assessment, selection and comparison, *Statistics Surveys*, *6*, 142–228.

Wang, Y., and J. M. G. Taylor (2001), Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome, *Journal of the American Statistical Association*, *96*(455), 895–905.

Watanabe, S. (2009), *Algebraic geometry and statistical learning theory*, Cambridge University Press.

Watanabe, S. (2010), Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *The Journal of Machine Learning Research*, *9999*, 3571–3594.

Xu, J., and S. L. Zeger (2001), The evaluation of multiple surrogate endpoints, *Biometrics*, *57*(1), 81–87.

Yu, M., J. M. G. Taylor, and H. M. Sandler (2008), Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model, *Journal of the American Statistical Association*, *103*(481), 178–187.

Yuan, Y., and V. E. Johnson (2012), Goodness-of-fit diagnostics for Bayesian hierarchical models, *Biometrics*, *68*(1), 156–164.