Joint Longitudinal and Survival Models for Intensive Longitudinal Data from Mobile Health Studies

by

Madeline R. Abbott

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Walter Dempsey, Co-Chair
Professor Jeremy M. G. Taylor, Co-Chair
Research Associate Professor Inbal Nahum-Shani
Associate Professor Zhenke Wu

Madeline R. Abbott
mrabbott@umich.edu
ORCID iD: 0000-0002-5344-3732

ACKNOWLEDGEMENTS

I am sincerely grateful for all of the support that I have received from my co-advisors, Jeremy M. G. Taylor and Walter Dempsey, over these past 4-6 years. I am lucky to have had two advisors who have met with me week after week, who have consistently provided me with prompt and thoughtful feedback, and who have never become impatient when my progress has been slow. They have given me both flexibility and a lot of support as I figured out each of my dissertation projects. Through their excellent mentorship, I have learned how to develop and frame a research project, to be persistent when the first approach fails, and to ask questions of my results when things seem off. I appreciate their encouragement to attend conferences and workshops, to submit papers to student competitions, and to write an F31 grant that provided the framework for these dissertation projects. During our weekly meetings, I have always appreciated Walter sharing his enthusiasm for algebra and Jeremy sharing his wisdom and experience. Thank you for all of your suggestions, advice, questions, patience, and mentorship.

I am also thankful for the advice and support that I have received from Billie Nahum-Shani. I am grateful for her valuable feedback as I wrote, and rewrote, my F31. Through her mentorship, she helped me develop skills that have made me a better communicator and collaborative statistician. I would also like to acknowledge researchers at the University of Utah, who provided the data motiving much of this work. Dave Wetter, Cho Lam, and Lindsey Potter generously shared data with me from three different smoking cessation studies, enabling me to illustrate the statistical methods in this dissertation with interesting examples. Both Lindsey and Jamie Yap patiently answered my many questions as I worked to understand these datasets.

I would also like to thank Zhenke Wu, for serving on my committee, as well as the broader members of the Biostatistics Department and University, including Matt Schipper, members of TaBaBooHe, d3c, and the administrative and computing support staff, including Mike Kleinsasser and Dan Barker. I would like to acknowledge the funding support that I have received from the NIH Cancer Biostatistics Training Grant and National Research Service Award for Individual Predoctoral Fellows. Thank you to my current and former labmates, officemates, and friends—Elizabeth, Emily (also the department's #1 peer mentor), Nate, Lam, Fatema, Jess, Grant, and Rachel—for your moral support and advice.

I am grateful for all of my (current and former) Ann Arbor friends, my volleyball teammates, and fellow track clubbers. Thank you to Deesha, Kim, Irena, and Elyse for keeping me well-fed and well-entertained, with freshly baked late-night chocolate chip cookies, floats down the river, adventures to Detroit, and long bike rides. Thank you especially to Laura and Marge and Kaitlyn, for keeping things silly and making sure I stay a fun nerd.

Finally, thank you to my parents for their unwavering support throughout my entire education, from preschool through 22nd grade. Thank you to Tristan for "keeping me humble" [3] and to Tenley for paving the way to full-time employment.

This dissertation was made possible by the support from so many different people and I am sincerely grateful for everyone, including others not named above, who have helped me get to where I am today.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vii

LIST OF TABLES xi

LIST OF APPENDICES xii

LIST OF ACRONYMS xiii

ABSTRACT xiv

CHAPTER

1 Introduction 1

2 A Continuous-Time Dynamic Factor Model for Intensive Longitudinal Data Arising from Mobile Health Studies 6

2.1 Introduction 6
2.2 Motivating Data 9
2.3 Methods 9
2.3.1 Measurement Submodel 9
2.3.2 Structural Submodel 11
2.3.3 Likelihood Definition 12
2.3.4 Identification Issues 14
2.3.5 Estimation Algorithm 15
2.4 Simulation Study 17
2.4.1 Data Generation for Assessing Bias and Variance 18
2.4.2 Bias and Variance Results 19
2.4.3 Data Generation for Model Selection 19
2.4.4 Model Selection Results 22
2.5 Application to mHealth Emotion Data 22
2.6 Discussion 26

3 A Latent Variable Approach to Jointly Modeling Longitudinal and Cumulative Event Data Using a Weighted Two-Stage Method 29

3.1 Introduction 29
3.1.1 Related Work 30
3.1.2 Main Contributions and Outline 31
3.2 Motivating Data 32
3.3 Methods 34
3.3.1 Longitudinal Submodel 34
3.3.2 Cumulative Risk Submodel 36
3.3.3 Estimation 37
3.3.4 Inference 39
3.4 Simulation Study 40
3.5 Application to Smoking Cessation Data 43
3.6 Discussion 47

4 A Bayesian Joint Longitudinal-Survival Model with a Latent Stochastic Process for Intensive Longitudinal Data 50

4.1 Introduction 50
4.1.1 Related Work 50
4.1.2 Main Contributions and Outline 52
4.2 Motivating Data 53
4.3 Methods 55
4.3.1 Measurement Submodel 55
4.3.2 Structural Submodel 55
4.3.3 Survival Submodel 56
4.3.4 Likelihood 56
4.4 Simulation Study 59
4.4.1 Discrete Approximation of the Survival Function 60
4.4.2 Simulation Results 60
4.5 Analysis of Smoking Cessation Data 61
4.6 Discussion 66

5 Estimation of Time-Varying Treatment Effects in a Joint Model for Longitudinal and Recurrent Event Outcomes in Mobile Health Data 71

5.1 Introduction 71
5.2 Motivating Data 74
5.3 Methods 76
5.3.1 Modeling the Impact of Treatment on the Latent Process 78
5.3.2 Modeling the Impact of Treatment on the Hazard of Recurrent Events 81
5.3.3 Inference 82
5.4 Simulation Study 85
5.4.1 Data Generation 85
5.4.2 Results 87

5.5 Analysis of the MRT Data 87
5.6 Discussion 93

6 Conclusion **96**

APPENDICES 101

BIBLIOGRAPHY 195

LIST OF FIGURES

FIGURE

2.1 Responses to the EMA questions over time for one participant in the mHealth study, separated by positive and negative emotions. 10
2.2 Parameter estimates from the block coordinate descent algorithm for the three different settings in which the true OU process differs. 20
2.3 Comparison of estimated standard errors (from Fisher information) and standard deviation of point estimates. 21
2.4 Point estimates and corresponding 95% confidence intervals (CI) for each of the parameter matrices in our two-factor OUF model. 24
2.5 The top panel shows the decay in autocorrelation and cross-correlation between latent factors that represent positive affect ($\eta_1(t)$) and negative affect ($\eta_2(t)$) across increasing gap times, where time is measured in hours. 25

3.1 Plot of self-reported emotions from the random EMAs and self-reported cigarette use from all EMAs for three individuals in the smoking cessation study. 33
3.2 Box plots of point estimates of cumulative risk model parameters from simulation study. 42
3.3 Point estimates and 95% confidence intervals for cumulative risk model parameters applied to mHealth smoking cessation study data. 45
3.4 Expected smoking rate (in units of cigarettes per 12 hour interval) using estimated coefficients and assuming constant values of either (a) average positive affect (PA = 0) and average negative affect (NA = 0), (b) above-average positive affect (PA = 1) and below-average negative affect (NA = -1), or (c) below-average positive affect (PA = -1) and above-average negative affect (NA = 1). 46

4.1 Longitudinal responses to the 9 emotion-related questions for one individual in the smoking cessation study. 54
4.2 For data generated under settings 1 and 2 with each of the four measurement patterns, we use box plots to summarize the distribution of the posterior medians for all parameters across the 100 simulated datasets. 62
4.3 For data generated under settings 1 and 2 with each of the four measurement patterns, we summarize the coverage rate of 90% credible intervals across the 100 simulated datasets with the colored dots. 63
4.4 For data generated under settings 1 and 2 with each of the four measurement patterns, we use box plots to summarize the distribution of the difference between the posterior medians and true values by grid width for the survival submodel parameters across the 100 simulated datasets. 64

4.5 Plot of posterior medians and 95% credible intervals for parameters in the joint model with a piecewise constant baseline hazard fit to data from the mHealth smoking cessation study. 67
4.6 Posterior samples of the latent factors (interpreted as positive and negative affect) and cumulative hazards for four individuals in the mHealth smoking cessation study. 68

5.1 Simplified diagram of the micro-randomized trial design. 74
5.2 Responses to emotion-related questions, timing of recurrent poly-substance use events, and timing of treatments for eight different participants in the motivating MRT. 76
5.3 Treatment effect models for the longitudinal process. 82
5.4 For data generated under settings 1 and 2 with different hazards and treatment effect models, we use box plots to summarize the distribution of the posterior medians for all parameters across the 100 simulated datasets. 88
5.5 For data generated under settings 1 and 2 with different hazards and treatment effect models, we summarize the coverage rate of 95% credible intervals across the 100 simulated datasets with the colored dots. 89
5.6 Posterior means and 95% credible intervals for parameters in the four different joint models that make different assumptions about how the treatment impacts the latent process and hazard. 92

A.1 Point estimates for each of the parameter matrices in our one-factor OUF model. 119
A.2 Point estimates for each of the parameter matrices in our three-factor OUF model. 121
A.3 Computation time (in minutes) for our estimation algorithm and the Bayesian method proposed in Tran et al. (2021) [112] 125
A.4 Final parameter estimates from our block coordinate descent algorithm and the Bayesian method proposed in Tran et al. (2021) [112]. 126

B.1 Illustration of two event intervals as defined by three measurement occasions (i.e., random EMAs). 128
B.2 Illustration of data collected over a single event interval for a single individual. . 133
B.3 Stage 1 point estimates from the simulation study. 138
B.4 Point estimates from simulation study assessing sensitivity to R. True coefficient values are indicated with dashed orange horizontal lines. 139
B.5 Computation time (in minutes) for stage 1 and stage 2 in the simulation study, where times are summarized across the 100 replicates using box plots. 140
B.6 Point estimates and corresponding 95% confidence intervals (CIs) for parameters in the longitudinal submodel applied to the mHealth smoking cessation study data with $N = 218$ individuals. 141
B.7 Point estimates and 95% confidence intervals (CIs) for cumulative risk model parameters applied to the subset of mHealth smoking cessation study data with $N = 214$ individuals. 142

C.1 Each column displays the true values of the four longitudinal outcomes for a different individual in the simulated data under setting 1. 146

C.2 Each column displays the true values of the four longitudinal outcomes for a different individual in the simulated data under setting 2. 147
C.3 Kaplan-Meier curves for a single dataset with each measurement pattern (1-4) simulated using the true set of parameters in setting 1. 149
C.4 Kaplan-Meier curves for a single dataset with each measurement pattern (1-4) simulated using the true set of parameters in setting 2. 150
C.5 For data generated under settings 1 and 2 with each of the four measurement patterns, we summarize the time required to run Stan for 3000 iterations. . . . 151
C.6 For data generated under settings 1 and 2 with each of the four measurement patterns, we use box plots to summarize the distribution of the posterior medians for all parameters across the 100 simulated datasets. 153
C.7 For data generated under settings 1 and 2 with measurement pattern 3, we use box plots to summarize the distribution of the posterior medians for all parameters across the 100 simulated datasets when fitting the model without added grid points. 154
C.8 For data generated under settings 1 and 2 with each of the four measurement patterns, we summarize the coverage rate of 90% credible intervals across the 100 simulated datasets with the colored dots. 155
C.9 The Kaplan-Meier curve for the time-to-event outcome of time until first lapse. 156
C.10 Trace plot of posterior samples (after burn-in) for the joint model with the piecewise constant baseline hazard fit to data from the mHealth smoking cessation study. 159
C.11 Posterior densities of parameters for the joint model with the piecewise constant baseline hazard applied to data from the mHealth smoking cessation study. . . 160
C.12 Goodness of fit for the joint model with the piecewise constant baseline hazard in the survival submodel 161
C.13 Goodness of fit for the joint model with the Weibull baseline hazard in the survival submodel. 162

D.1 Timing of recurrent poly-substance use events. 166
D.2 Mean cumulative function for recurrent poly-substance use. 166
D.3 Mean cumulative function for recurrent poly-substance use by pre- and post-quit periods. 167
D.4 Trace plot for the joint model with an additive model for treatment effect in the longitudinal submodel and with a single treatment parameter in the hazard model. 169
D.5 Trace plot for the joint model with an additive model for treatment effect in the longitudinal submodel and with separate pre-quit and post-quit treatment parameters in the hazard model. 170
D.6 Trace plot for the joint model with a drift model for treatment effect in the longitudinal submodel and with a single treatment parameter in the hazard model. 171
D.7 Trace plot for the joint model with a drift model for treatment effect in the longitudinal submodel and with separate pre-quit and post-quit treatment parameters in the hazard model. 172
D.8 Decay in cross- and auto-correlation in bivariate OU process from fitted joint models. 173

D.9 Estimated treatment-related terms from fitted joint models. 175
D.10 Diagram illustrating potential mechanisms (a)–(c) for the effect of a single treatment at time $t_{j'}$, $a_i(t_{j'})$, on future values of the latent process $\eta_i(t_{j+1})$, the measured longitudinal outcomes $Y_i(t_{j+1})$, and the recurrent event outcome (T_{ir}, δ_{ir}). 177
D.11 Curves for the $g(\cdot)$ functions that capture the association between the hazard of an event as a function of time (in days) since the most recent prior event. . . . 180
D.12 Setting 1: Treatment effect is additive and the hazard takes the form of model 1. 181
D.13 Setting 1: Treatment effect is modeled as drift and the hazard takes the form of model 1. 182
D.14 Setting 2: Treatment effect is additive and the hazard takes the form of model 1. 183
D.15 Setting 2: Treatment effect is modeled as drift and the hazard takes the form of model 1. 184
D.16 For data generated under settings 1 and 2 with hazard model 2 and different treatment effect models in the longitudinal submodel, we use box plots to summarize the distribution of the posterior medians for all parameters across the 5 simulated datasets. 186
D.17 For data generated under settings 1 and 2 with hazard model 2 and different treatment effect models in the longitudinal submodel, we summarize the coverage rate of 95% credible intervals across the 5 simulated datasets with the colored dots. 187
D.18 Comparison of approximate log-likelihood values for each individual evaluated at the posterior mean to the exact log-likelihood for each individual using the true marginal distribution evaluated at the posterior mean. 192
D.19 Comparison of approximate log-likelihood values for each individual evaluated at each posterior sample to the exact log-likelihood for each individual using the true marginal distribution evaluated at each posterior sample s. 193

LIST OF TABLES

TABLE

2.1 For datasets generated under each true model, we summarize the percent of times that the model-selection metric chose the fitted model with the indicated number of factors. 22

3.1 Coverage rates (%) for 95% confidence intervals for the cumulative risk submodel parameters in the simulation study. 43

5.1 WAIC for the joint models fit to the motivating MRT data. 93

A.1 For datasets generated under each true model, we summarize the percent of times that the model-selection metric chose the fitted model with the indicated number of factors. 117

A.2 For datasets generated under each true model, we summarize the number of datasets (out of 100) on which the algorithm converged or reached the maximum number of block-wise iterations prior to convergence (when $\delta = 1 \times 10^{-6}$). . . . 117

A.3 For datasets generated under each true model, we summarize the number of datasets (out of 100) on which the algorithm converged or reached the maximum number of block-wise iterations prior to convergence (when $\delta \leq 1 \times 10^{-3}$). . . . 118

A.4 For datasets generated under each true model, we summarize the percent of times that the model-selection metric chose the fitted model with the indicated number of factors. 118

A.5 Behavioral science literature supporting the division of the positive emotions into two groups representing no-to-low arousal positive affect and high arousal positive affect. 120

B.1 Coverage rates (CRs; %) for 80% and 95% confidence intervals from the simulation study assessing sensitivity to R, where confidence intervals are calculated from von Hippel standard errors. 137

D.1 Number of times that each number of puffs was observed and conversion of puffs to events. 165

D.2 Empirical variability in measured longitudinal outcomes in the motivating MRT. 176

D.3 DIC for the joint models fit to the motivating MRT data. 194

LIST OF APPENDICES

A Supplementary Material for: A Continuous-Time Dynamic Factor Model for Intensive Longitudinal Data Arising from Mobile Health Studies . . . 101

B Supplementary Material for: A Latent Variable Approach to Jointly Modeling Longitudinal and Cumulative Event Data Using a Weighted Two-Stage Method . . . 127

C Supplementary Material for: A Bayesian Joint Longitudinal-Survival Model with a Latent Stochastic Process for Intensive Longitudinal Data 143

D Supplementary Material for: Estimation of Time-Varying Treatment Effects in a Joint Model for Longitudinal and Recurrent Event Outcomes in Mobile Health Data . . . 163

LIST OF ACRONYMS

AIC Akaike information criterion

AR autoregressive

BIC Bayesian information criterion

DIC deviance information criterion

EMA ecological momentary assessment

HIV human immunodeficiency virus

HMC Hamiltonian Monte Carlo

ILD intensive longitudinal data

JITAI just-in-time adaptive intervention

MCEM Monte Carlo Expectation Maximization

mHealth mobile health

MRT micro-randomized trial

PCA principal components analysis

OU Ornstein-Uhlenbeck

SDE stochastic differential equation

VAR vector autoregressive

WAIC Watanabe-Akaike information criterion

ABSTRACT

Mobile health (mHealth) technology enables the collection of intensive longitudinal data (ILD) and, as a result, serves as a rich source of information on both the short-term and long-term dynamics of multiple outcomes measured over time. When combined with time-to-event outcomes, ILD can provide insight into factors that elevate the risk of an event. Motivated by mHealth studies of smoking cessation in which participants report both longitudinal data on the intensity of many emotions multiple times per day and event-time information on cigarette use, this dissertation presents methods for jointly modeling multivariate ILD and time-to-event outcomes.

In the first project, we develop a dynamic factor model that summarizes ILD as a smaller number of time-varying latent factors. The evolution of these latent factors is modeled using a multivariate continuous-time Ornstein-Uhlenbeck stochastic process. We propose a block coordinate descent algorithm for maximum likelihood estimation and apply our method to mHealth data to summarize the dynamics of 18 emotions as two latent factors. These latent factors are interpreted by behavioral scientists as the psychological constructs of positive and negative affect.

In the second project, we extend this dynamic factor model to consider an event outcome. Specifically, we use the latent factors as time-varying predictors of a cumulative event outcome (e.g., the total number of cigarettes smoked across repeated intervals of time), which we model using Poisson regression. We take a two-stage approach to estimation; we use weights—based on importance sampling—to account for potential bias that could result from the two-stage approach.

In the third project, we extend this dynamic factor model to model the longitudinal process jointly with a traditional survival outcome (e.g., the time of first cigarette use after attempted quit). In this joint longitudinal-survival model, the hazard of a time-to-event outcome is a function of the low-dimensional latent process. Joint estimation of this model is challenging due to the combination of ILD and the presence of a stochastic process as a time-varying covariate in our hazard model. We fit our joint model with a Bayesian approach and use it to analyze data from another mHealth study of smoking cessation. We summarize the longitudinal self-reported intensity of nine emotions as the psychological states of positive and negative affect; these time-varying latent states capture the risk of the first smoking lapse

after attempted quit.

In the fourth project, we present a model-based approach for estimating the effect of repeatedly delivered treatments in a micro-randomized trial (MRT) via an extension of our joint model. We discuss different ways that these repeated treatment effects can be incorporated into the joint model; these different model specifications correspond to different mechanisms by which treatment is assumed to impact the longitudinal and event processes. Taking a Bayesian approach to inference, we model the association between repeated app-based notifications, longitudinally-measured emotions, and recurrent events of substance use in an mHealth MRT.

CHAPTER 1

Introduction

The ubiquity of smartphones presents researchers with abundant opportunities to collect rich data and develop new and effective interventions in mobile health (mHealth) studies. mHealth methods for data collection are particularly useful for measuring longitudinal changes in outcomes of interest, ranging from blood-based biomarkers to self-reported engagement in certain health behaviors. These methods make frequent measurement possible, allowing researchers to capture outcomes that vary rapidly. mHealth methods also allow for real-time data collection without requiring individuals to attend in-person clinic visits. As termed in Stone and Shiffman (1995) [105], ecological momentary assessment (EMA) is one such mHealth method that focuses on frequently assessing outcomes of interest as they are experienced by individuals in their natural environments. EMAs often consist of surveys sent to smartphones that prompt study participants to respond to a set of questions assessing numerous factors related to their current state (e.g., emotional intensity or stress levels) and context (e.g., geographical location or social setting).

Not only do mHealth methods allow for frequent measurement of multiple outcomes of interest, but they also facilitate delivery of interventions in real time. A just-in-time adaptive intervention (JITAI) is a type of intervention that can be delivered at specific moments in time; these moments may include instances when, for example, a participant is at increased risk of engaging in a certain type of behavior, is expected to be more receptive to the intervention, and/or is in a situation that presents an opportunity for positive change [70]. Often, JITAIs take the form of app-based notifications or text messages that prompt individuals to make certain decisions or encourage them to behave in certain ways. micro-randomized trial (MRT)s, a specific type of trial in which individuals are repeatedly randomized to potentially receive an intervention, were developed to help investigators design evidence-based JITAIs [45]. For example, investigators may use an MRT to better understand what type of intervention is most effective when individuals are in different states or contexts. An MRT can also be used to learn rules that define the probability of sending an intervention

in different contexts. Recent examples of MRTs used to develop JITAIs include Klasnja et al. (2019) [46], Nahum-Shani et al. (2021) [68], and Jeganathan et al. (2022) [42].

Overall, the use of mHealth technology in research has resulted in an abundance of complicated and rich data—on multivariate intensive longitudinal outcomes, time-to-event outcomes, and repeated treatments—that present numerous opportunities and challenges for statistical analysis. In this dissertation, we attempt to fill some gaps in the statistical methods available for analyzing this type of complex data. Methods appropriate for modeling intensive longitudinal data (ILD) will only become more important as the popularity of mHealth studies, and thus availability of this type of data, continuous to increase.

Chapter 2 focuses on the longitudinal data collected in mHealth studies; specifically, ILD. ILD generally consist of a moderate to large number of longitudinal outcomes measured frequently over time; this type of data tends to contain more measurement occasions than traditional longitudinal data. In behavioral science research, ILD is common; for example, multiple emotions are sometimes measured with the goal of understanding trends in a smaller number of underlying and unobservable psychological states. Factor models are commonly applied in psychology and behavioral science as a way of summarizing and understanding the relationship between multiple outcomes [125, 20]. Traditional factor models assume that the summarizing latent factors are cross-sectional (e.g., [8, 104, 34]), while dynamic factor models capture the longitudinal aspects of both the measured outcomes and the latent factors (e.g., [113, 52, 121]).

When analyzing ILD on rapidly varying outcomes, such as emotions, it is important to both use a model that is flexible enough to account for abrupt correlated change and a method that is computationally efficient enough to handle ILD. Although random effect models (e.g., models with a random slope and random intercept) are useful for capturing smooth trends in correlated longitudinal variables, they are less suitable for modeling outcomes that change abruptly (e.g., rapidly varying psychological states). Recent work by Tran et al. (2021) [112] proposed a flexible modeling approach that combines a dynamic factor with a latent stochastic process. This work was motivated by clinical data on a neurologic disease (amyotrophic lateral sclerosis). In this dissertation, we focus on higher-dimensional longitudinal data—specifically ILD—collected in mHealth studies; for example, 10-20 emotions measured up to 4 times per day over 10 days using EMA. In Chapter 2, we present a similar model to that in Tran et al. (2021) [112] but develop a computationally efficient approach to estimation that facilitates use of this type of model with ILD. Our work helps fulfill the need for methods that scale for use with large datasets consisting of many subjects and measurements (e.g., EMAs) per subject.

While understanding the dynamics of latent variables and how they related to sources of

variability in measured multivariate longitudinal outcomes is interesting in its own right, ILD may also contain information on vulnerability to certain event-time outcomes. A multitude of studies have established a link between psychological state—e.g., positive and negative affect or self-efficacy—and engagement in risky health behaviors, such as smoking and other drug use [39, 33, 48]. While many studies rely on events of interest to be self-reported in EMAs, some studies also use sensors to pick up additional information (e.g., outcomes such as stress [38] or smoking [99]). Depending on the phrasing of EMA questions, we may record information on events either in a traditional time-to-event format or as a total number of events experienced across some known interval of time (e.g., the time between the current and prior EMA). In Chapter 3, we present a model that allows us to assess the association between time-varying predictions of latent factors (e.g., positive and negative affect) and the risk of cumulative event outcomes (e.g., total numbers of cigarettes smoked across repeated intervals of time).

A straightforward way to connect EMA responses (e.g., emotions) with events of interest is to first model the longitudinal trajectory of the EMA responses, and then funnel the output of this model into another model to predict the risk of an event (e.g., numbers of cigarettes smoked). This type of approach, originally developed in Tsiatis et al. (1995) [115], is called a two-stage approach. Since publication of Tsiatis et al. (1995) [115], however, numerous papers have described a well-known issue with this strategy: it does not account for measurement error in the longitudinal outcome and can result in biased estimates of association between the longitudinal and the event-time outcomes. Much work has been done to develop corrections for this bias (e.g., [57, 5]), since this two-stage approach conveniently requires less intensive computation than joint estimation. In Chapter 3, we take a similar strategy and develop a two-stage estimation method that combines the estimation algorithm from Chapter 2 and model-fitting functions from standard `R` packages for a simple and effective approach to estimation. As in Mauff et al. (2020) [57], we use importance sampling-based weights to reduce potential bias that could result from this two-stage method.

While the two-stage approach has the advantage of being a fast and easy way to link longitudinal and event submodels, since the publication of Tsiatis et al. (1995) [115], joint estimation has gained popularity as an elegant and statistically efficient alternative. Joint models—models that are fit simultaneously to both the longitudinal and event-time outcomes—are powerful tools for enabling less-biased estimation of the association between temporal variations in longitudinal outcomes (e.g., emotions) and the risk of event-time outcomes (e.g., lapses in smoking cessation). Tsiatis and Davidian (2004) [114] present a nice overview of traditional joint models.

Much statistical work on joint models has been motivated by data arising from studies on

human immunodeficiency virus (HIV) and cancer. In HIV, joint models have been developed to model changes in longitudinal measurements of CD4 cell counts and the risk of disease progression or death [22, 109, 122]. Data on prostate cancer has motivated models that allow researchers to jointly model changes in prostate specific antigen levels and the risk of cancer recurrence [129, 110]. Predictors in the hazard models often include the current value of the longitudinal process, but can also include terms based on the integral or derivative of the longitudinal process.

A major challenge in the development of joint models is their computational cost; namely, the need to evaluate complex and often intractable integrals in the survival function and across unobserved random effects in the longitudinal submodel. Joint models can quickly become prohibitively computationally expensive if multiple longitudinal responses are considered simultaneously. Much existing literature has employed two general strategies for estimation and inference of joint models; these strategies include (i) an imputation-based approach in which values of the longitudinal outcome are estimated using either empirical Bayes estimators or best linear unbiased predictors (BLUPs) [128, 5] and (ii) a likelihood-based approach using an expectation maximization (EM) algorithm that involves integrating over the latent random effects in the longitudinal submodel [127, 36]. This first approach is computationally intensive due to its iterative nature; the second approach is computationally costly because it often involves evaluating intractable integrals. Numerical approaches to approximating integrals have been suggested and include Gauss Hermite quadrature or Monte Carlo approximations. To reduce the computational cost of integration, discrete-time approximations have previously been used (e.g., [122, 36]). In cases when the longitudinal submodel involves a complicated function of high-dimensional random effects and the number of repeated measurements per individual is high, Rizopoulos et al. (2009) [93] propose using the fully exponential Laplace approximation. Bayesian approaches are also common when fitting joint models, as they handle latent variables naturally. In Chapter 4, we combine the dynamic factor model from Chapter 2 with a time-to-event model and take a Bayesian approach to inference. In our joint model, the hazard depends on a multivariate stochastic process, so calculating the cumulative hazard, which requires integrating over this stochastic process as a function of time, is challenging. We present a method that can handle ILD, filling a gap in the literature for joint models suitable for use with ILD.

In the final project of this dissertation, described in Chapter 5, we consider an additional feature of some mHealth studies: mHealth interventions. We specifically consider JITAIs. These types of interventions often consist of text messages or app-based notifications delivered directly to study participants' smartphones. Standard approaches to estimating treatment effects from this type of intervention are based on generalized estimating equations.

Weighted and centered least-squares is the usual approach [9], with various adaptations of this method also having been proposed recently (e.g., [79] and [101]). Model-based approaches to treatment effect estimation are less common in mHealth settings, but are often used in other areas, like cancer, for example. In joint models, the risk of bias in the coefficient estimates linking the longitudinal and event processes decreases and statistical efficiency increases, as this joint approach considers dependencies between the longitudinal and survival processes. This decrease in bias and increase in efficiency also applies to treatment effect estimation [40]. In Chapter 5, we develop different treatment effect models and incorporate them into our joint longitudinal event-time model. In this chapter, we also consider recurrent events, rather than single time-to-event outcome, since recurrent events are commonly recorded in mHealth studies due to the high frequency of measurement.

The methods proposed in Chapters 2-5 of this dissertation are motivated by three specific mHealth studies of smoking cessation: two observational studies [74, 14, 119] and one MRT [68]. All studies enrolled current smokers who attempted to quit. The studies used EMAs to record the intensity of many emotions multiple times per day over the course of the study. The use of both cigarettes and other types of substances was also recorded. In the MRT, participants were randomized to be sent interventions multiple times per day [68]. We use data from the observational studies to motivate the development of the methods proposed in Chapters 2, 3, and 4, and use data from the MRT to drive the work proposed in Chapter 5. Although these motivating datasets focus on smoking cessation, the statistical methods proposed in this dissertation are applicable to similarly structured mHealth studies and ILD in other domains.

CHAPTER 2

A Continuous-Time Dynamic Factor Model for Intensive Longitudinal Data Arising from Mobile Health Studies

2.1 Introduction

Intensive longitudinal data (ILD) can capture rapid changes in outcomes over time. Often in mobile health (mHealth) studies, many longitudinal outcomes are measured with the aim of understanding the temporal dynamics of unobservable constructs related to mental or physical health. Our work is motivated by an observational mHealth study in which the intensity of emotions was collected over time. Participants self-reported the intensity of 18 different emotions up to four times per day over 10 days, resulting in a substantial quantity of rich data. For behavioral scientists, understanding the temporal dynamics of the latent psychological states that underlie these emotions—and how well these emotions measure the specific latent states—is of scientific interest.

The volume and complexity of ILD, however, make them challenging to analyze since longitudinal outcomes are often measured irregularly across many individuals. Thus statistical methods must be able to handle the irregular spacing of this high volume of data. At the same time, the frequent measurements in ILD create many opportunities to discover new information, particularly if the latent constructs of interest vary rapidly. We present a dynamic factor model that is motivated by the need to model multiple longitudinal outcomes measured frequently over time in a flexible yet interpretable manner. The model, which is similar to that described in Tran et al. (2021) [112], consists of two submodels: (i) a measurement submodel—a factor model—that summarizes the multiple observed longitudinal outcomes as lower-dimensional latent factors and (ii) a structural submodel—an Ornstein-Uhlenbeck (OU) stochastic process—that captures the evolution of the multiple correlated latent factors over time. Together, these components of our dynamic factor model

are flexible enough to capture abrupt changes in the longitudinal outcomes while avoiding use of a non-parametric or other many-parameter model that inhibits interpretability. The low-dimensional nature of the structural submodel also greatly reduces computational complexity, as opposed to fitting a high-dimensional stochastic process directly to the observed outcomes.

One standard approach to modeling changes in multiple correlated longitudinal variables is to use an autoregressive (AR) model. These models, which are called vector autoregressive (VAR) models when data are multivariate, have been widely used to model observed outcomes as well as latent variables. For example, Dunson (2003) [23], Cui and Dunson (2014) [21], and Tran (2019) [113] have proposed related methods in which observed longitudinal outcomes are summarized as time-varying lower-dimensional latent variables. The correlation of these latent variables is then modeled with AR or VAR processes. VAR models, however, are specified for balanced data. This situation is often not realistic in the case of ILD, which generally consists of irregularly-measured outcomes, and can lead to biased estimates in cases where the assumption is made but does not hold.

Mixed models have been proposed as alternatives to discrete-time processes for modeling the evolution of latent variables over time, and have been previously used in combination with factor models. Unlike the AR and VAR processes, mixed models do not require balanced data. Existing work has focused both on the development of mixed models for modeling the evolution of a single latent factor over time (e.g., [96, 76, 77]) or multiple latent factors (e.g., [53, 121]). Overall, these mixed model-based approaches are useful tools for capturing smooth trends in latent factors but may have trouble capturing changes that happen rapidly (e.g., abrupt jumps in psychological states).

The OU process, which can be thought of as a continuous-time analog of the AR or VAR process, is a stochastic process well-suited for capturing abrupt variations over time. Existing work has frequently focused on using the OU process or integrated OU process to model longitudinal outcomes that have been directly observed (or observed with measurement error); e.g., [109, 106, 71, 72].

Most closely related to our proposed approach is the work in Tran et al. (2021) [112]. Like us, the authors propose a longitudinal latent variable model that consists of two parts: a measurement submodel to summarize observed outcomes as lower dimensional latent factors and an OU process as the structural submodel for the latent factors. While we differ in the exact specification of the measurement submodel, our chosen models are related. The key distinction between this existing work and the work presented in this manuscript lies in the approach to estimation and inference. Tran et al. (2021) [112] take a Bayesian approach, which uses a form of the likelihood that requires sampling values of the latent factors at

each measurement occasion. In the ILD setting, we need approaches that can scale to large numbers of repeated measurements. Here, we choose to work in the frequentist framework and directly maximize the marginal log-likelihood of the observed longitudinal outcome by integrating out latent variables, resulting in a method more suitable for ILD. While Tran et al. (2021) [112] present an approach that relies on algebraic constraints to fit models with two or three latent factors, our maximum-likelihood approach enables us to easily extend our model to larger numbers of latent factors through the use of penalties, rather than algebraic constraints. Finally, although we—like Tran et al. (2021) [112]—assume that the number of latent factors is known, our work additionally investigates the use of information criteria to select the true model among models with misspecified numbers of latent factors in a simulation study. The marginal log-likelihood of the observed data that we use here better facilitates the use of information criterion to compare models with different numbers of latent factors, as opposed to a version of the likelihood that conditions on the latent variables (see Merkle et al. (2019) [60] for more discussion of marginal vs. conditional likelihoods for factor models).

In this work, we build on the model from Tran et al. (2021) [112] by developing and evaluating the performance of an efficient estimation algorithm that has the computational ability to handle ILD. Our work enables the analysis of high-dimensional ILD using low-dimensional stochastic latent variable models; as a result, these models can be used to understand how well observed longitudinal outcomes measure underlying states, how correlated these latent states are over time, and how much of the variation in the longitudinal outcome is related to short-term changes within an individual vs. longer-term differences across individuals. Designed specifically for the ILD setting, our novel methodological contributions include (i) a closed-form likelihood for the marginal distribution of the observed outcome, (ii) the derivation of the computationally-simpler sparse precision matrix for the multivariate OU process, (iii) identifiability constraints imposed via scaling constants, and (iv) a block coordinate descent algorithm for estimation and inference in a maximum likelihood framework.

The remainder of this paper is organized as such: In Section 2.2, we describe the motivating ILD from an mHealth study; in Section 2.3, we present the model and our novel methodological contributions; in Section 2.4, we demonstrate the performance of our method via simulation; in Section 2.5, we use our method to analyze intensive longitudinal emotion data collected in an mHealth study; and in Section 2.6, we provide a discussion.

2.2 Motivating Data

The ILD motivating this work consist of self-reported emotional states collected in an observational mHealth study [74]. Over a period of 10 days, ecological momentary assessments (EMAs), which enable repeated sampling of individuals' current states and contexts in real time, were used to frequently track participants' emotions as they were experienced. Specifically, participants were prompted to respond to a series of questions sent to their smartphones multiple times per day at random occasions; the original study design intended for individuals to receive up to four EMAs per day. The EMAs contained a set of questions that assessed the current intensity of multiple emotions measured on a 5-point Likert scale. We focus on a set of 18 emotions; these emotions are active, angry, ashamed, attentive, calm, determined, disgusted, enthusiastic, grateful, guilty, happy, irritable, joyful, lonely, nervous, proud, sad, and scared. The resulting data contain frequent measurements of a substantial number of longitudinal outcomes, where the number of measurement occasions per person ranges from 2 to 47 (mean = 17). The variability in total number of observations per person is due to a combination of intermittent non-response to the EMAs and dropout.

The high rate of measurement enables us to capture rapid changes in emotions—and thus different aspects of the latent psychological states—over time. Note that these data are the subset of the full study data that were available at the time of drafting this manuscript (N = 218 individuals). Additional details on the study can be found in [74].

We illustrate the variability in these longitudinal outcomes in Figure 2.1, which shows the responses to emotion-related EMA questions over time for one participant in the study. Understanding the dynamics of individuals' latent psychological states that underlie the measured responses, as well as investigating the appropriate number of latent states to summarize the observed responses, is of scientific interest among behavioral scientists.

2.3 Methods

In this section, we present the OUF model that jointly models multiple observed longitudinal outcomes (here, emotions) and the lower dimensional latent factors (representing, for example, psychological states) assumed to generate the observed longitudinal outcomes. The model consists of two submodels: a measurement submodel and a structural submodel.

2.3.1 Measurement Submodel

Let $\boldsymbol{Y}_i(t) = [Y_{i1}(t), Y_{i2}(t), ..., Y_{iK}(t)]^\top$ be a $K \times 1$ vector of measured longitudinal outcomes (e.g., emotions in the motivating data) for individual $i, i = 1, ..., N$, at time t. Assume that

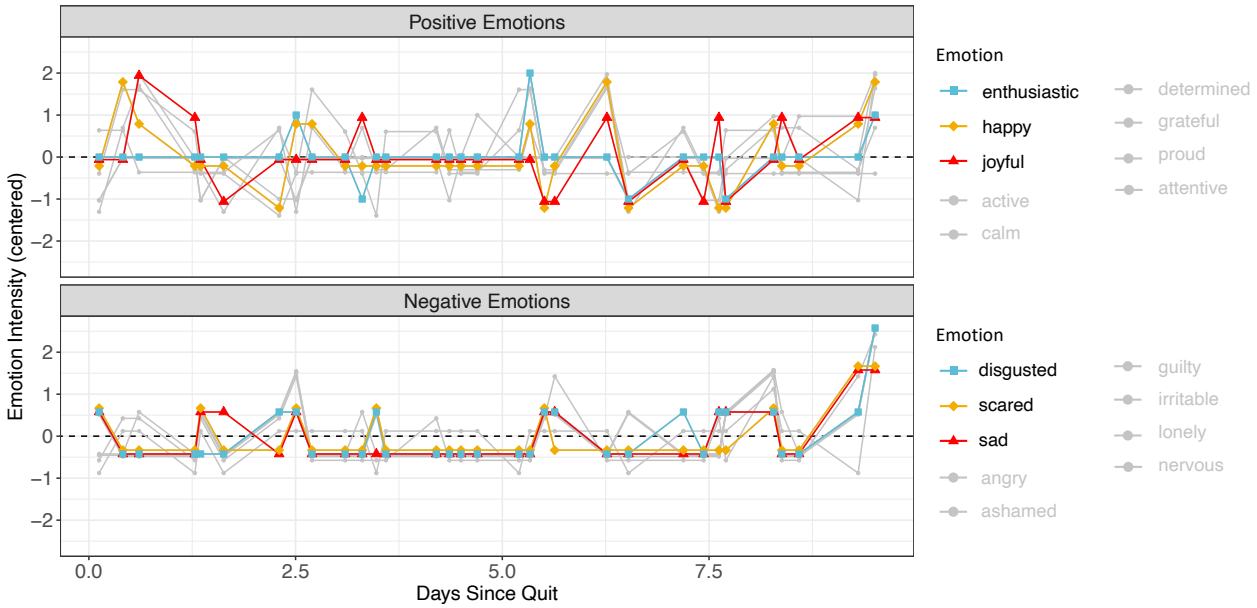


Figure 2.1: Responses to the EMA questions over time for one participant in the mHealth study, separated by positive and negative emotions. In this plot, a subset of three positive emotions and three negative emotions are highlighted solely for illustrative purposes; all 18 emotions are later included in the model described in Section 2.5. Note both the high correlation and abrupt fluctuations of these longitudinal outcomes over time.

individual i has longitudinal outcomes measured at n_i occasions (e.g., at n_i EMAs). Using the measurement submodel, we model the observed longitudinal outcome $\boldsymbol{Y}_i(t)$ as

$$\boldsymbol{Y}_i(t) = \boldsymbol{\Lambda}\boldsymbol{\eta}_i(t) + \boldsymbol{u}_i + \boldsymbol{\epsilon}_i(t) \tag{2.1}$$

where $\boldsymbol{\eta}_i(t)$ is a vector of p time-varying latent factors (where $p < K$); $\boldsymbol{\Lambda}$ is a $K \times p$-dimensional time-invariant loadings matrix with elements $\lambda_{k,j}$ that captures the degree of association between the latent factors and observed longitudinal outcomes; $\boldsymbol{u}_i \sim N(0, \boldsymbol{\Sigma}_u)$ is a vector of length K of random intercepts; and $\boldsymbol{\epsilon}_i(t) \sim N(0, \boldsymbol{\Sigma}_\epsilon)$ is a vector representing measurement error, where $\boldsymbol{\Sigma}_\epsilon$ is assumed to be a diagonal matrix.

This model builds upon a standard factor model but also includes (i) a random intercept and (ii) a multivariate model for the evolution of the correlated latent processes $\boldsymbol{\eta}_i(t)$ (described in the next section). This random intercept was previously introduced in [112]. We assume that $\boldsymbol{\Sigma}_u$ is diagonal, as we include this term to account for baseline differences across individuals, but then model the correlated change in outcomes through the structural submodel. Allowing a non-diagonal $\boldsymbol{\Sigma}_u$ is possible, but we opt not to do so to avoid the substantial increase in computational cost associated with estimation of these extra parameters. In the context of modeling emotions over time, we can interpret the random intercept as

accounting for differences in psychological traits (i.e., a construct that is more stable within a person) while the dynamic latent factors capture changes in psychological state (i.e., a construct that varies more quickly [100]).

We also assume that $\mathbf{\Lambda}$ contains many structural zeros such that each row of the loadings matrix contains only one non-zero element; this structure means that each observed outcome is a measurement of only a single latent factor. The decision to incorporate structural zeros in the loadings matrix is supported by behavioral science concepts (e.g., Positive and Negative Affect Schedule [124]), which can be used to classify a given emotion as a measurement of a specific category of emotional state.

2.3.2 Structural Submodel

The structural submodel captures the evolution of the latent factors, $\boldsymbol{\eta}_i(t)$, over time. In the motivating data, these latent factors are psychological states (e.g., positive/negative affect, valence, arousal, etc.) assumed to generate the measured emotions. We use a multivariate OU process, which can be understood as a continuous-time analog of a VAR process and has the ability to capture abrupt temporal variation. Here, we assume a bivariate OU process ($p = 2$) for illustrative purposes. The stochastic differential equation (SDE) definition of the bivariate OU process is

$$d\begin{bmatrix}\eta_{i1}(t)\\ \eta_{i2}(t)\end{bmatrix} = -\underbrace{\begin{bmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\end{bmatrix}}_{:=\boldsymbol{\theta}}\begin{bmatrix}\eta_{i1}(t)\\ \eta_{i2}(t)\end{bmatrix}dt + \underbrace{\begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}}_{:=\boldsymbol{\sigma}} d\begin{bmatrix}W_{i1}(t)\\ W_{i2}(t)\end{bmatrix} \tag{2.2}$$

where the diagonal elements of matrix $\boldsymbol{\theta}$ capture the mean-reverting tendency of the latent factors (where the mean is assumed to be 0) and the off-diagonal elements of $\boldsymbol{\theta}$ capture correlation between the latent factors. The diagonal elements of $\boldsymbol{\theta}$ are required to be positive. The matrix $\boldsymbol{\sigma}$, with elements σ_{11} and $\sigma_{22} > 0$, describes the volatility of the process, where $W_{i1}(t)$ and $W_{i2}(t)$ are both standard Brownian motion. In general, the standard definition of the OU process allows $\boldsymbol{\sigma}$ to take non-zero values in the off-diagonal. By restricting $\boldsymbol{\sigma}$ to be a simpler diagonal matrix here, we consider the Brownian motion terms as separate noise processes for each latent factor and thus capture all correlation between the latent factors through the $\boldsymbol{\theta}$ matrix. We also require that all eigenvalues of the $\boldsymbol{\theta}$ matrix have a positive real part; this constraint ensures a mean-reverting process [112].

2.3.3 Likelihood Definition

Rather than taking a Bayesian strategy or relying on the complete-data likelihood and taking an expectation-maximization (EM) approach to estimation, we directly maximize the likelihood of the observed data. Direct maximization of the marginal likelihood allows us to avoid repeatedly calculating values of the latent factors at each measurement occasion (via posterior sampling in a Bayesian framework or via complex integrals in the E-step of the EM algorithm). Thus, our approach is more scalable to the ILD setting.

In existing literature, the OU process is most often defined using its conditional distribution. If our p latent factors for individual i, denoted by vector $\boldsymbol{\eta}_i$, follow an OU process, then the conditional distribution of the latent factors at time t given the previous value at time s, where $s < t$, is

$$\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s) \sim N\left(e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s), \boldsymbol{V} - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t-s)}\right)$$

This distribution assumes that the initial value of the OU process is drawn from its stationary distribution, $\boldsymbol{\eta}_i(t_0) \sim N(0, \boldsymbol{V})$, where the stationary variance is $\boldsymbol{V} := vec^{-1}\big\{(\boldsymbol{\theta} \oplus \boldsymbol{\theta})^{-1} vec\{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\}\big\}$. Here, $\oplus$ denotes the Kronecker sum, defined for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of sizes a and b, respectively, as $\boldsymbol{A} \oplus \boldsymbol{B} = \boldsymbol{A} \otimes \boldsymbol{I}_b + \boldsymbol{I}_a \otimes \boldsymbol{B}$; and the $vec\{\boldsymbol{A}\}$ operation consists of stacking the columns of matrix $\boldsymbol{A}$ into a column vector.

The conditional distribution can be challenging to work with in the context of ILD, as it requires computing products sequentially across all measurement times within the likelihood. To simplify computation in our ILD setting, we integrate out the latent factors so that we can simply maximize the observed data log-likelihood. This marginal likelihood depends on the joint distribution of the latent factors. The joint distribution of $\boldsymbol{\eta}_i = \left[\boldsymbol{\eta}_i^\top(t_1), \boldsymbol{\eta}_i^\top(t_2), ..., \boldsymbol{\eta}_i^\top(t_{n_i})\right]^\top$ is

$$\boldsymbol{\eta}_i \sim N(\boldsymbol{0}, \boldsymbol{\Psi}_i)$$

where

$$\boldsymbol{\Psi}_i = \begin{bmatrix} \boldsymbol{V} & \boldsymbol{V}e^{-\boldsymbol{\theta}^\top|t_2-t_1|} & \dots & \boldsymbol{V}e^{-\boldsymbol{\theta}^\top|t_{n_i-1}-t_1|} & \boldsymbol{V}e^{-\boldsymbol{\theta}^\top|t_{n_i}-t_1|} \\ e^{-\boldsymbol{\theta}|t_2-t_1|}\boldsymbol{V} & \boldsymbol{V} & \dots & \boldsymbol{V}e^{-\boldsymbol{\theta}^\top|t_{n_i-1}-t_2|} & \boldsymbol{V}e^{-\boldsymbol{\theta}^\top|t_{n_i}-t_2|} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-\boldsymbol{\theta}|t_{n_i-1}-t_1|}\boldsymbol{V} & e^{-\boldsymbol{\theta}|t_{n_i-1}-t_2|}\boldsymbol{V} & \dots & \boldsymbol{V} & \boldsymbol{V}e^{-\boldsymbol{\theta}^\top|t_{n_i-1}-t_{n_i}|} \\ e^{-\boldsymbol{\theta}|t_{n_i}-t_1|}\boldsymbol{V} & e^{-\boldsymbol{\theta}|t_{n_i}-t_2|}\boldsymbol{V} & \dots & e^{-\boldsymbol{\theta}|t_{n_i}-t_{n_i-1}|}\boldsymbol{V} & \boldsymbol{V} \end{bmatrix}$$

The dimension of the marginal OU covariance matrix $\boldsymbol{\Psi}_i$ still scales with the number of

longitudinal measurements and so to make our approach computationally amenable to the ILD setting, we take advantage of the fact that the OU process has the Markov property. As a result of this property, the inverse of the marginal covariance matrix—the precision matrix—is block tri-diagonal; thus it is much simpler to evaluate the likelihood for the OU process when written in terms of the sparse precision matrix, compared to either the dense marginal covariance matrix or as a product of many conditional distributions. As one of the key contribution of this paper, we derive this sparse precision matrix: Let $\boldsymbol{\Omega}_i$ be the precision matrix of the OU process observed at n_i occasions. Then $\boldsymbol{\Omega}_i$ has the structure

$$\boldsymbol{\Omega}_i = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} & 0 & \cdots & 0 \\ \boldsymbol{\Omega}_{12}^\top & \boldsymbol{\Omega}_{22} & \boldsymbol{\Omega}_{23} & \cdots & 0 \\ 0 & \boldsymbol{\Omega}_{23}^\top & \boldsymbol{\Omega}_{33} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \boldsymbol{\Omega}_{n_i-1,n_i} \\ 0 & 0 & \cdots & \boldsymbol{\Omega}_{n_i-1,n_i}^\top & \boldsymbol{\Omega}_{n_i n_i} \end{bmatrix} \tag{2.3}$$

and each block indexed by j for $1 < j < n_i$ in the tri-diagonal matrix is

$$\begin{aligned}
\boldsymbol{\Omega}_{11} &= \left[\boldsymbol{V} - \boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_2-t_1)}\boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_2-t_1)}\boldsymbol{V}\right]^{-1} \\
\boldsymbol{\Omega}_{j,j+1} &= -\left[\boldsymbol{V} - \boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_{j+1}-t_j)}\boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_{j+1}-t_j)}\boldsymbol{V}\right]^{-1}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_{j+1}-t_j)}\boldsymbol{V}^{-1} \\
\boldsymbol{\Omega}_{jj} &= \boldsymbol{V}^{-1} + \boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_j-t_{j-1})}\boldsymbol{V}\left[\boldsymbol{V} - \boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_j-t_{j-1})}\boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_j-t_{j-1})}\boldsymbol{V}\right]^{-1}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_j-t_{j-1})}\boldsymbol{V}^{-1} \\
&\quad + \left[\boldsymbol{V} - \boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_{j+1}-t_j)}\boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_{j+1}-t_j)}\boldsymbol{V}\right]^{-1}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_{j+1}-t_j)}\boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_{j+1}-t_j)} \\
\boldsymbol{\Omega}_{n_i n_i} &= \boldsymbol{V}^{-1} + \boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_{n_i}-t_{n_i-1})}\boldsymbol{V}\left[\boldsymbol{V} - \boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_{n_i}-t_{n_i-1})}\boldsymbol{V}^{-1}e^{-\boldsymbol{\theta}(t_{n_i}-t_{n_i-1})}\boldsymbol{V}\right]^{-1} \\
&\quad \cdot \boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t_{n_i}-t_{n_i-1})}\boldsymbol{V}^{-1}
\end{aligned} \tag{2.4}$$

The derivation for each block is given in Section A.3. Later, during estimation, we leverage the sparse precision matrix to simplify computation. This sparsity becomes particularly advantageous as the number of individuals and observations per individual (e.g., EMAs per individual) in a dataset increases, and it is critical to the scalability of our model to the ILD setting.

Together, the measurement and structural submodels imply that the observed longitudinal outcomes are normally distributed with mean 0 and covariance $\boldsymbol{\Sigma}_i^* := Var(\boldsymbol{Y}_i) = (\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Lambda})Var(\boldsymbol{\eta}_i)(\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Lambda})^\top + \boldsymbol{J}_{n_i} \otimes \boldsymbol{\Sigma}_u + \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}_\epsilon$, where $\boldsymbol{I}_{n_i}$ is an $n_i \times n_i$ identity matrix and $\boldsymbol{J}_{n_i}$ is an $n_i \times n_i$ matrix of ones. We estimate the OUF model by minimizing the following function, which equal to twice the negative log-likelihood up to a constant: $-2logL(\boldsymbol{Y}) = \sum_{i=1}^N log|\boldsymbol{\Sigma}_i^*| + \sum_{i=1}^N \boldsymbol{Y}_i^\top \boldsymbol{\Sigma}_i^{*-1}\boldsymbol{Y}_i$.

2.3.4 Identification Issues

Before fitting our model, we must make additional assumptions to address identifiability issues common to factor models. Because both $\boldsymbol{\Lambda}$ and $\boldsymbol{\eta}_i(t)$ are unknown, multiplying $\boldsymbol{\Lambda}$ by some matrix, say $\boldsymbol{A}$, and multiplying $\boldsymbol{\eta}_i(t)$ by $\boldsymbol{A}^{-1}$ will result in the same model. To make a factor model identifiable, constraints must be placed on either the loadings matrix or the latent factors. Aguilar and West (2000) [4] and Carvalho et al. (2008) [19], for example, make the standard assumption of requiring the loadings matrix to be triangular while Tran et al. (2019) [113], for example, fix the variance of the latent factors at 1. The main disadvantage of assuming that $\boldsymbol{\Lambda}$ has a triangular structure is that the order of the longitudinal outcomes matters, and so the structure of this matrix is less intuitive to specify based on behavioral science literature. Assuming that the latent factors have a variance of 1 simply means that we model the latent psychological constructs on the correlation scale.

Thus, to make our model identifiable, we fix the scale of the latent factors but propose a novel approach for doing so. Letting $\boldsymbol{\eta}_i$ be the $(p \times n_i)$-length vector of latent variables values stacked over measurement occasions, we constrain $Var(\boldsymbol{\eta}_i)$ to have diagonal elements equal to 1. This constraint means that the OU process must have a stationary variance equal to 1. By fixing the scale of the latent factors, we can allow the elements of the loadings matrix $\boldsymbol{\Lambda}$ to vary almost freely during estimation. For a generic $\boldsymbol{\Lambda}$ (without structural zeros), the only constraint on the loadings matrix is that the sign of the first element must be positive. Together these constraints make our model identifiable; the constraint on the OU process identifies the scale and the constraint on the first element of the loadings matrix identifies the direction. Because we later make the simplifying assumption that $\boldsymbol{\Lambda}$ contains structural zeros with a single non-zero loading per row, flipping the signs on both the loadings and the latent factors results in the same model; we choose to keep the signs that correspond to the most relevant interpretation of the model given the application. Another constraint could be added to require that one loading per column of $\boldsymbol{\Lambda}$ is positive; this would avoid sign flipping.

To impose this identifiability constraint, we use a set of p constants to re-scale the OU process parameters. We summarize this identifiability constraint for the bivariate ($p = 2$) OU process as: Using a pair of positive scalar constants c_1 and c_2, we can re-scale an arbitrary OU process parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ to have stationary variance of 1, where this re-scaled OU process is parameterized by $\boldsymbol{\theta}^*$ and $\boldsymbol{\sigma}^*$ according to

$$\begin{bmatrix} \theta^*_{11} & \theta^*_{12} \\ \theta^*_{21} & \theta^*_{22} \end{bmatrix} = \begin{bmatrix} \theta_{11} & \frac{c_1}{c_2}\theta_{12} \\ \frac{c_2}{c_1}\theta_{21} & \theta_{22} \end{bmatrix} \text{ and } \begin{bmatrix} \sigma^*_{11} & 0 \\ 0 & \sigma^*_{22} \end{bmatrix} = \begin{bmatrix} c_1\sigma_{11} & 0 \\ 0 & c_2\sigma_{22} \end{bmatrix} \tag{2.5}$$

In Section A.4, we show why this re-scaling approach works for any mean-reverting OU

process. This constraint can also be extended to OU processes of higher dimensions.

Although this identifiability assumption allows us to identify the magnitude of the loadings in the factor model, it does so only up to a sign change. Consider again the case of a bivariate OU process. To make this example more concrete, suppose also that one of the latent factors, η_1, is measured by the positive emotions and the other latent factor, η_2, is measured by the negative emotions collected in the motivating mHealth study. The likelihood for our model is equivalent for pairs of scaling constants $(c_1 = 1, c_2 = 1)$ and $(c_1 = 1, c_2 = -1)$. In practice, the model would be the same under both pairs of scaling constants (and so we restrict c_1 and c_2 to be positive during estimation) but interpretation of model parameters would differ. After estimation, the signs of estimated model parameters can easily be flipped to match the most relevant interpretation of the data by multiplying estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\theta}$ by a $p \times p$ matrix with the constants along the diagonal. In this two-factor example, it would make sense to choose signs such that η_1 and η_2 are negatively correlated and higher values of the latent factors correspond to higher values of the measured emotions. As a result, η_1 could be interpreted as representing positive affect and η_2 as negative affect, both of which are two traditional psychological constructs often used in behavioral science [124].

2.3.5 Estimation Algorithm

To fit this model, we take an iterative approach to estimation in which we directly maximize the marginal likelihood of our observed longitudinal outcome using a block coordinate descent algorithm and rely on simpler existing models to inform the initial parameter estimates. To increase the computational efficiency of this estimation algorithm, we (i) take advantage of tractable analytic gradients for the measurement submodel, avoiding the need to calculate computationally expensive numerical gradients; (ii) leverage the Markov property of the OU process and use the computationally-simpler sparse precision matrix derived in Equation 2.3, rather than the dense covariance matrix; and (iii) implement the code used to repeatedly calculate these numerical gradients and the sparse precision matrix in C++, using R for the rest of our code.

In the block coordinate descent algorithm, we split parameters into two different blocks: one block for parameters in the measurement submodel ($\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}_u$, $\boldsymbol{\Sigma}_\epsilon$) and the other for parameters in the structural submodel ($\boldsymbol{\theta}$, $\boldsymbol{\sigma}$). Note that each element of these blocks is actually a matrix of parameters. Within each block-wise iteration, we minimize the log-likelihood with respect to one block of parameters, given the current estimates of the other block of parameters, using Newton algorithms as implemented in `R`'s `stats` package [81]. By updating parameters in blocks, we can leverage the availability of analytic gradients

for parameters in the measurement submodel. The Kronecker structure of the covariance matrix for each individual's longitudinal outcomes $\boldsymbol{Y}_i$ allows us to derive these analytic gradients. The gradient of the log-likelihood for a single individual with respect to one of the measurement submodel parameters, Θ_j, has the general form

$$\frac{\partial logL(\boldsymbol{Y}_i)}{\partial \Theta_j} = -\frac{1}{2}\left[tr\left\{\left(I - \boldsymbol{\Sigma}_i^{*-1}\boldsymbol{Y}_i\boldsymbol{Y}_i^{\top}\right)\Sigma_i^{*-1}\frac{\partial \boldsymbol{\Sigma}_i^*}{\partial \Theta_j}\right\}\right] \tag{2.6}$$

where the exact form of $\frac{\partial \boldsymbol{\Sigma}_i^*}{\partial \Theta_j}$ depends on the specific parameter; either λ_k, σ_{u_k}, or σ_{ϵ_k}. The complete set of analytic gradients is given in Section A.5. The computational advantage of using the analytic gradient, as opposed to a numerical approach to differentiation, is particularly notable as the number of longitudinal outcomes—and thus parameters in the measurement submodel—increases.

Prior to maximizing the marginal likelihood, we use a cross-sectional factor model to initialize $\boldsymbol{\Lambda}$, $\boldsymbol{\theta}$, and $\boldsymbol{\sigma}$, and use linear mixed models to initialize $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\epsilon$. Then, we iteratively update parameter estimates using the following block coordinate descent algorithm:

1. *Initialize estimates of* $\boldsymbol{\Lambda}^{(0)}, \boldsymbol{\Sigma}_u^{(0)}, \boldsymbol{\Sigma}_\epsilon^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\sigma}^{(0)}$. *Measurement submodel parameters are always initialized empirically; for structural submodel parameters, two sets of initial estimates are considered—an empirical set of values estimated from cross-sectional factor scores and a default set of values. The set of values that corresponds to the higher log-likelihood given the current data is used.*

2. *Set iteration index* $r = 1$ *and convergence indicator* $\delta = 0$. *While* $\delta = 0$,

   (a) *Update block 1 (measurement submodel):*

   $$\boldsymbol{\Lambda}^{(r)}, \boldsymbol{\Sigma}_u^{(r)}, \boldsymbol{\Sigma}_\epsilon^{(r)} = \underset{\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon}{argmax}\left\{logL(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon | Y; \boldsymbol{\theta}^{(r-1)}, \boldsymbol{\sigma}^{(r-1))})\right\}.$$

   Maximization is done via a Newton-type algorithm (`nlm`; [81]) using analytic gradients (Equation 2.6).

   (b) *Update block 2 (structural submodel):*

   $$\boldsymbol{\theta}^{(r)}, \boldsymbol{\sigma}^{(r)} = \underset{\boldsymbol{\theta}, \boldsymbol{\sigma}}{argmax}\left\{logL(\boldsymbol{\theta}, \boldsymbol{\sigma} | Y; \boldsymbol{\Lambda}^{(r)}, \boldsymbol{\Sigma}_u^{(r)}, \boldsymbol{\Sigma}_\epsilon^{(r)})\right\}.$$

   Maximization is done via a quasi-Newton algorithm (`nlminb`; [81]) using numerical gradients and takes advantage of the sparsity of the OU precision matrix to increase the speed of this step. A large positive penalty is added to the nega-

tive log-likelihood within the optimization algorithm if a proposed $\boldsymbol{\theta}$ does not have eigenvalues with positive real parts.

(c) *Using Equation 2.5, re-scale OU parameters to satisfy the identifiability constraint.*

(d) *Check for block-wise convergence: Let $\boldsymbol{\Theta}$ be a vector containing all elements of $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}_u$, $\boldsymbol{\Sigma}_\epsilon$, $\boldsymbol{\theta}$, and $\boldsymbol{\sigma}$. Then, calculate*

$$\delta = \max\Big\{ I\big\{|\boldsymbol{\Theta}^{(r)} - \boldsymbol{\Theta}^{(r-1)}|/\boldsymbol{\Theta}^{(r)} < 10^{-6}\big\}, \\ I\big\{logL(\boldsymbol{\Theta}^{(r)}|\boldsymbol{Y}) - logL(\boldsymbol{\Theta}^{(r-1)}|\boldsymbol{Y}) < 10^{-6}\big\}\Big\}$$

where all operations on $\boldsymbol{\Theta}$ are element-wise.

(e) *Update r: $r = r + 1$*

3. *Estimate Fisher Information-based standard errors from numerical approximations to the Hessian of the log-likelihood, $\frac{\partial^2}{\partial\Theta\partial\Theta^\top} logL(\boldsymbol{\Lambda}^{(r)}, \boldsymbol{\Sigma}_u^{(r)}, \boldsymbol{\Sigma}_\epsilon^{(r)}, \boldsymbol{\theta}^{(r)}|Y)$.*

Note that when estimating standard errors, the parameterization of the likelihood differs slightly: the likelihood now depends on only one of the parameter matrices in the structural submodel, $\boldsymbol{\theta}$, and not the other, $\boldsymbol{\sigma}$. This change in parameterization is a result of the identifiability constraint that is placed on the stationary variance of the OU process. Since we are no longer conditioning on fixed measurement submodel parameters in step (3), we restrict $\boldsymbol{\sigma}$ to be a function of $\boldsymbol{\theta}$, where this function is derived from the identifiability constraint; thus, the likelihood is not over-parameterized. Standard error estimates for $\boldsymbol{\sigma}$ can be calculated quickly and easily using a parametric bootstrap. By sampling values of $\boldsymbol{\theta}$ from a Normal distribution defined by its point estimate and estimated covariance matrix, bootstrapped values of $\boldsymbol{\sigma}$ are calculated as a function of $\boldsymbol{\theta}$ and a confidence interval can be estimated based on the empirical distribution. More details on the parameterization of the log-likelihood for standard error estimation are in Section A.6.

2.4 Simulation Study

We conduct a simulation study to assess (i) the bias and variance of estimates when the OUF model is specified with the correct number of latent factors and (ii) the ability of Akaike information criterion (AIC) and Bayesian information criterion (BIC) to select the model with the correct number of latent factors among models with mis-specified numbers of latent factors and loadings matrices.

2.4.1 Data Generation for Assessing Bias and Variance

We assume that $K = 4$ longitudinal outcomes (e.g., emotions) are recorded over time for $N = 200$ individuals. For individual i, these longitudinal outcomes are measured at n_i different occasions (e.g., EMAs) where n_i takes a random integer value between 10 and 20. The gap time between each measurement occasion is drawn from a $Uniform(0.1, 2)$ distribution. Although our choice of 4 longitudinal outcomes is smaller than the number of outcomes often in seen in ILD, we chose this number to balance between the complexity of our data and model, and the computational demands of a simulation study.

We consider simulated data in three different settings in which the true bivariate OU process has varying degrees of autocorrelation (see Section A.7 for details). Using each true OU process, we generate the observed longitudinal outcomes by drawing from $\boldsymbol{Y}_i \sim N(0, \boldsymbol{\Sigma}_i^*)$ where $\boldsymbol{\Sigma}_i^*$ is defined using

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1.2 & 0 \\ 1.8 & 0 \\ 0 & -0.4 \\ 0 & 2 \end{bmatrix}, \boldsymbol{\Sigma}_u = \begin{bmatrix} 1.1 & 0 & 0 & 0 \\ 0 & 1.3 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & 0.9 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 0.6 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.7 \end{bmatrix}. \quad (2.7)$$

When fitting this model, we assume that the zeros within the loadings matrix, random intercept covariance matrix, and measurement error covariance matrix are known.

Importantly, some of the parameter values used to generate the data are different from the parameters that will be estimated by the model; this difference is a side effect of the identifiability assumption. While unbiased estimates of $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\epsilon$ will match the values used in data generation, the values of $\boldsymbol{\Lambda}$ and the OU process parameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ will differ. As a result of the re-scaling approach for identification, the estimated OU process has a stationary variance of 1. The additional variation present in the OU process during data generation must be absorbed by the loadings matrix $\boldsymbol{\Lambda}$. Specifically, the data-generating loadings matrix will be re-scaled according to $\boldsymbol{\Lambda D}$ where $\boldsymbol{D} := \sqrt{diag\{V(\boldsymbol{\theta}, \boldsymbol{\sigma})\}}$ and $\boldsymbol{V}$ is the stationary variance of the OU process. $\boldsymbol{\Lambda D}$ will be estimated by our algorithm. The data-generating OU parameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ will be re-scaled according to scalar constants chosen such that the stationary variance of the re-scaled OU process is equal to 1 via Equation 2.5. True parameter values indicated in the simulation results have all been re-scaled to match the values targeted by our estimation algorithm. In setting 2, the true OU process used to generate data does have a stationary variance equal to 1 and thus the target parameter values do match the data-generating parameter values.

2.4.2 Bias and Variance Results

In each of the three settings, we generate 1,000 datasets and carry out the estimation algorithm. Final point estimates are shown in Figure 2.2 and information-based standard errors are summarized in Figure 2.3. In all settings, we consistently recover unbiased estimates of the true values and find that the averages of the standard errors are similar to the empirical standard deviations of the point estimates, indicating that confidence intervals will have close to nominal coverage. For one dataset, numerical issues result in a negative variance estimate; this specific case is discussed in Section A.8.

2.4.3 Data Generation for Model Selection

Because ILD consist of many different outcomes, determining the appropriate number of latent factors for summarizing these multiple outcomes may frequently be of interest. As such, we carry out a simulation study in which we evaluate the ability of AIC and BIC to correctly select the true model among the misspecified models. The formulas for AIC and BIC take into account our identifiability constraints. Letting $\hat{L}$ denote the maximized value of the marginal (observed data) likelihood of the OUF model; q be the total number of non-zero parameters in $\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\theta}$ and $\boldsymbol{\sigma}$; p be the number of latent factors (which corresponds to the number of scaling constants needed to impose the identifiability constraint); and N be the total number of independent individuals in the data, then AIC is calculated as $2 \times (q - p) - 2log\hat{L}$ and BIC is calculated similarly as $2 \times log(N) \times (q - p) - 2log\hat{L}$.

Assuming the same true measurement submodel parameters as before, we now generate data from five different factor models: a one-factor model, a two-factor model with low signal (i.e., high correlation between latent factors), a two-factor model with high signal (i.e., low correlation between latent factors), a three-factor model with low signal, and a three-factor model with high signal. The various structures of these data-generating models can be interpreted as representing different beliefs about underlying psychological states. Data-generating parameter values are given in the Section A.7. For 100 datasets generated from each of these true models, we fit a one-, two-, and three-factor model and compare information criteria. For fitted models with misspecified numbers of latent factors, the loadings matrix is also misspecified; for fitted models with the true number of latent factors, the structure of the loadings matrix is correctly specified. We do not consider a four-factor model in this simulation study because our data only contain four longitudinal outcomes and so fitting a four-factor model would no longer fall into the dimension-reduction setting that motivates this work.

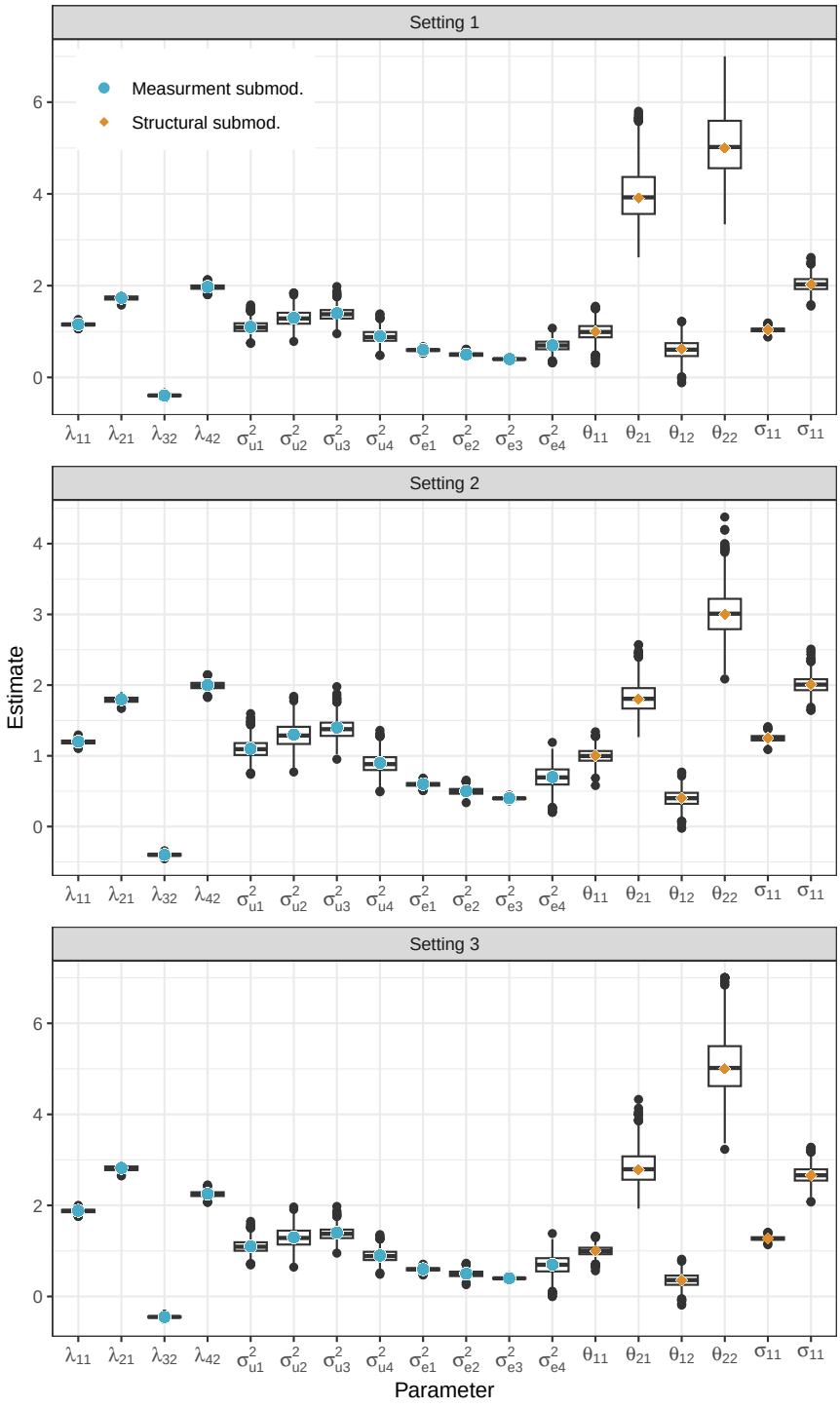


Figure 2.2: Parameter estimates from the block coordinate descent algorithm for the three different settings in which the true OU process differs. Point estimates are summarized across the 1000 simulated datasets with box plots and the dots indicate the true (target) parameter values.

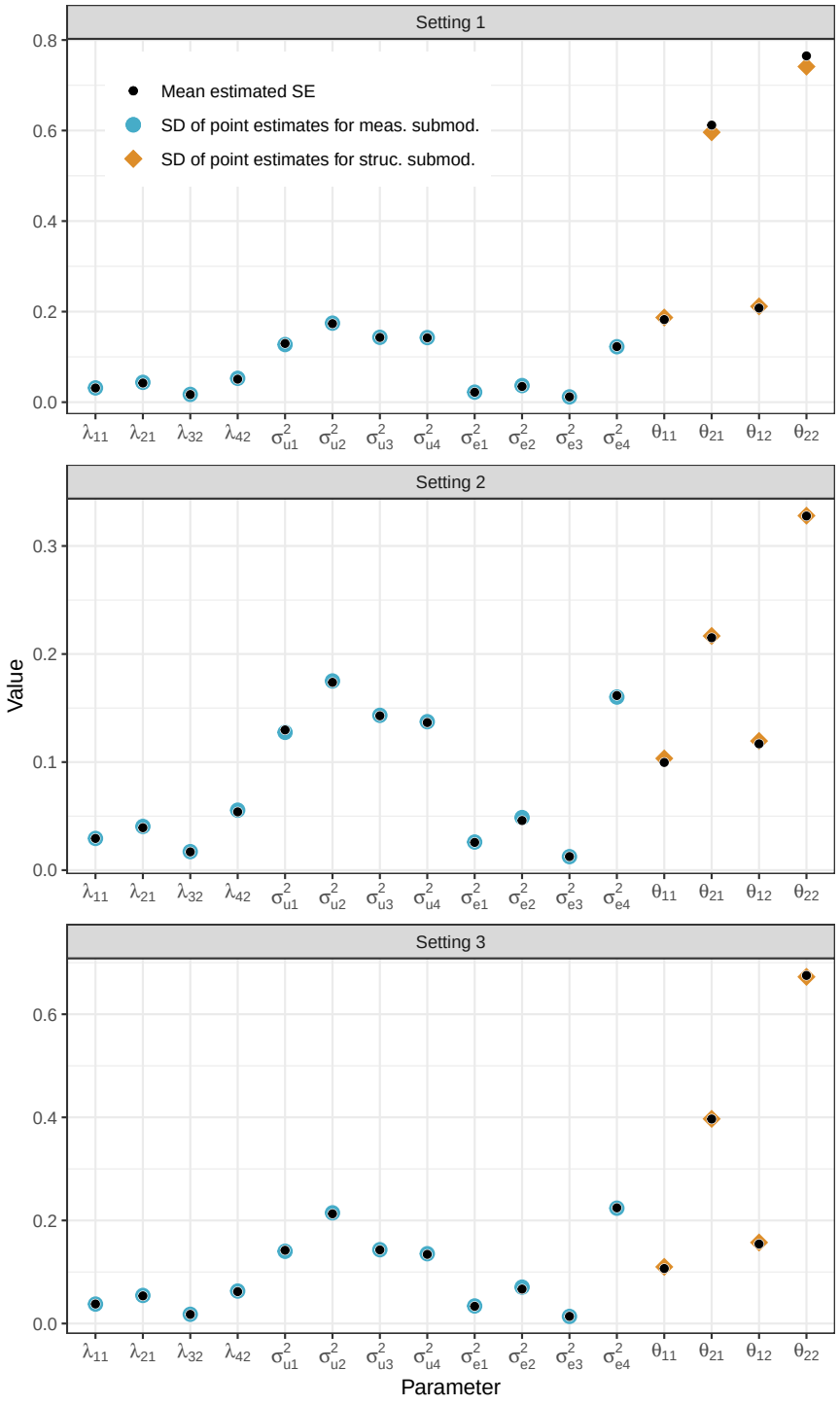


Figure 2.3: Comparison of estimated standard errors (from Fisher information) and standard deviation of point estimates. The similarity of the standard error estimates and empirical standard deviation suggests that the standard errors are of appropriate size. Note that the standard error estimate for $\sigma^2_{\epsilon_4}$ is missing for one datasets in setting 3 (see Section A.8 for more details).

True Model		# Factors in Fitted Model with Best AIC			# Factors in Fitted Model with Best BIC		
# Factors	Signal	1	2	3	1	2	3
1	-	**99**	0	1	**100**	0	0
2	Low	0	**93**	7	4	**96**	0
2	High	0	**100**	0	0	**100**	0
3	Low	0	0	**100**	0	8	**92**
3	High	0	0	**100**	0	0	**100**

Table 2.1: For datasets generated under each true model, we summarize the percent of times that the model-selection metric chose the fitted model with the indicated number of factors. When generating data from models with 2 and 3 factors, we considered two different settings: a high signal setting in which latent factors have lower correlation and a low signal setting in which latent factors have high correlation. The settings in which the fitted model has the same number of factors as the true data-generating model are emphasized with bold orange text. These results are presented for datasets on which the algorithm either converged or reached the maximum number of iterations (200) for all three models. See Section A.8 for more details.

2.4.4 Model Selection Results

We present model selection results in Table 2.1. In both the high and low signal settings, the model with the lowest AIC and BIC most often has the same number of factors as the true model used to generate the data. For models fit to data generated from a true model with three factors, BIC incorrectly selects a model with two factors more often than AIC. This difference make sense given the increased penalty that BIC places on model complexity. For datasets of this size ($N = 200$), estimation becomes more difficult as the number of factors increases and so, for a few simulated datasets, our algorithm did not converge within the allotted maximum number of block-wise iterations (see Section A.8 for details). While AIC and BIC perform similarly, we recommend use of BIC in practice, as the increased penalty placed on model complexity aligns well with the dimension-reduction goal of factor models.

2.5 Application to mHealth Emotion Data

We use our method to analyze the data on momentary emotions collected in the mHealth study. We fit three different OUF models in which we summarize the longitudinal responses to 18 emotion-related questions as either one, two, or three latent factors. The measured emotions that we model are: happy, joyful, enthusiastic, active, calm, determined, grateful, proud, attentive, sad, scared, disgusted, angry, ashamed, guilty, irritable, lonely, and nervous.

Behavioral scientist have a variety of theories that describe how these measured emotions relate to underlying psychological states (e.g., [95, 84, 85, 86, 32, 59]), and so we aim to compare the fit of models with different numbers of latent factors using this mHealth data.

The one-factor OUF model assumes that positive and negative emotions are generated from a single common underlying factor (i.e., a single spectrum that ranges from positive to negative affect) [95]. The two-factor OUF model assumes that the emotions are measurements of two distinct-but-correlated emotional states, which we interpret as positive affect and negative affect ([84]). In this model, happy, joyful, enthusiastic, active, calm, determined, grateful, proud, and attentive measure positive affect; and sad, scared, disgusted, angry, ashamed, guilty, irritable, lonely, and nervous measure negative affect. Finally, in the three-factor OUF model, we further divide the positive emotions into two latent factors that differ by the level of activation or arousal; we call these factors high arousal positive affect—measured by feeling grateful, proud, enthusiastic, active, determined, attentive—and no-to-low arousal positive affect—measured by feeling calm, happy, and joyful [85, 86, 32, 59]. The negative emotions are still assumed to be generated from one latent factor. Specifying these three models and comparing their fits allows us to investigate what level of dimension-reduction is appropriate for capturing the dynamics of the emotions measured in this mHealth study.

Both AIC and BIC indicate that, of the three models considered, the two-factor model fits best: $AIC_{1 \text{ factor}} = 123,309$ vs. $AIC_{2 \text{ factors}} = 121,069$ vs. $AIC_{3 \text{ factors}} = 124,957$ and $BIC_{1 \text{ factor}} = 123,791$ vs. $BIC_{2 \text{ factor}} = 121,577$ vs. $BIC_{3 \text{ factor}} = 125,509$. Some psychological theories support our conclusion that two factors represent our data better than one as it suggests that positive and negative affect are not opposites, rather they capture distinct-but-correlated components of psychological state [84]. The lower AIC and BIC of the two-factor model compared to the three-factor model suggest that the emotions corresponding to high arousal positive affect and no-to-low arousal positive affect are not different enough to justify the additional complexity of the three-factor model given the current data. The strong estimated correlation (0.995) between the latent factors for high arousal positive affect and no-to-low arousal positive affect further supports this conclusion.

For the bivariate OUF model, point estimates and 95% confidence intervals are in Figure 2.4. Coefficient estimates from the fitted one-factor and three-factor OUF models are given in Section A.9. For the two-factor model, measures of happiness, joy, and enthusiasm are most strongly correlated with positive affect and measures of sadness and irritability are most strongly correlated with negative affect. We use the estimated parameters of the OU process to understand the latent dynamics of positive and negative affect by plotting the degree of correlation for these two latent variables across varying time intervals between consecutive observations (see Figure 2.5). We see that positive and negative affect are negatively correlated as expected, and that the correlation between the latent states decays slowly.

We can also examine the variance estimates for all components of our model—the latent

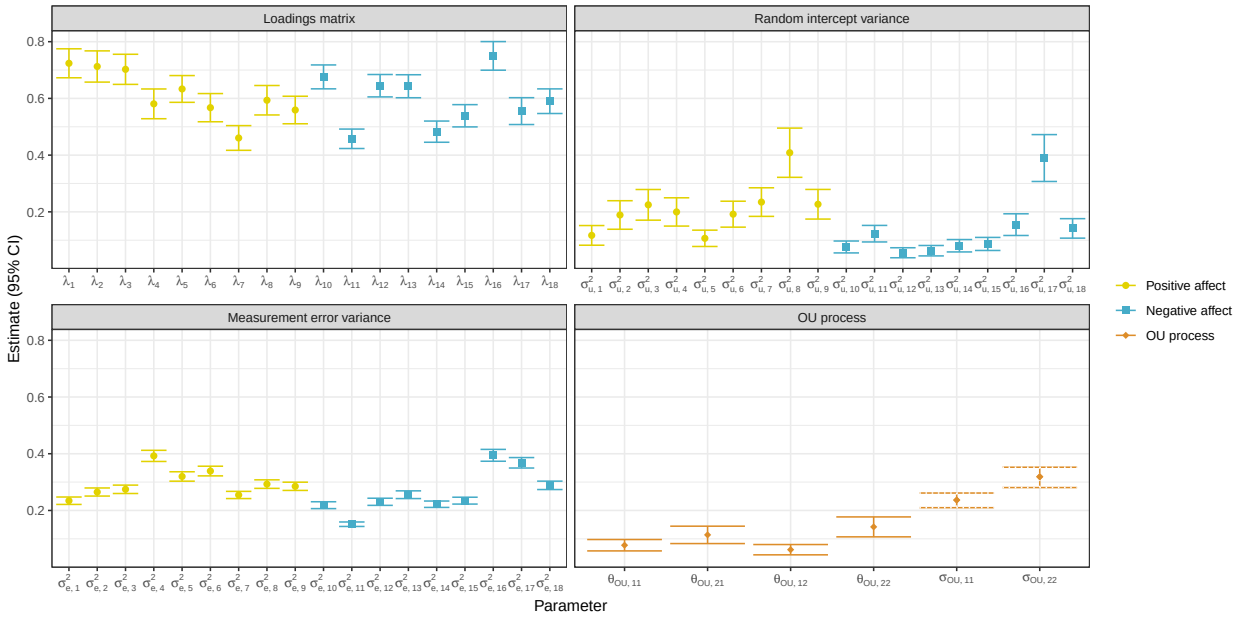


Figure 2.4: Point estimates and corresponding 95% confidence intervals (CI) for each of the parameter matrices in our two-factor OUF model. Intervals for OU parameters σ_{11} and σ_{22} are based on a parametric bootstrap. Because we assume structural zeros in the loadings matrix are known, each emotion has only a single loading. Parameter subscripts 1-18 correspond to the emotions as follows: 1 = happy, 2 = joyful, 3 = enthusiastic, 4 = active, 5 = calm, 6 = determined, 7 = grateful, 8 = proud, 9 = attentive, 10 = sad, 11 = scared, 12 = disgusted, 13 = angry, 14 = ashamed, 15 = guilty, 16 = irritable, 17 = lonely, 18 = nervous.

factors, random intercepts, and error terms—in order to help understand potential sources of variability. The relatively high variance estimates for the random intercepts for pride and loneliness suggest that these two emotions have higher variability across participants and vary less within participants; this pattern is consistent across all three OUF models. To gain further insight into the role of state vs. trait within this set of emotions, we can calculate the proportion of total variance explained by the latent process vs. the random intercepts for the set of 18 emotions at a fixed time point. We find that the dynamic latent factors explain more of the variability in happy (69% from the latent factors vs. 15% from the random intercept) and disgusted (79% vs. 11%), for example. On the other hand, the random intercepts explain more variability in proud (30% from the latent factors vs. 35% from the random intercept) and lonely (28% from the latent factors vs. 36% from the random intercept). The remaining proportion of variance is attributed to measurement error.

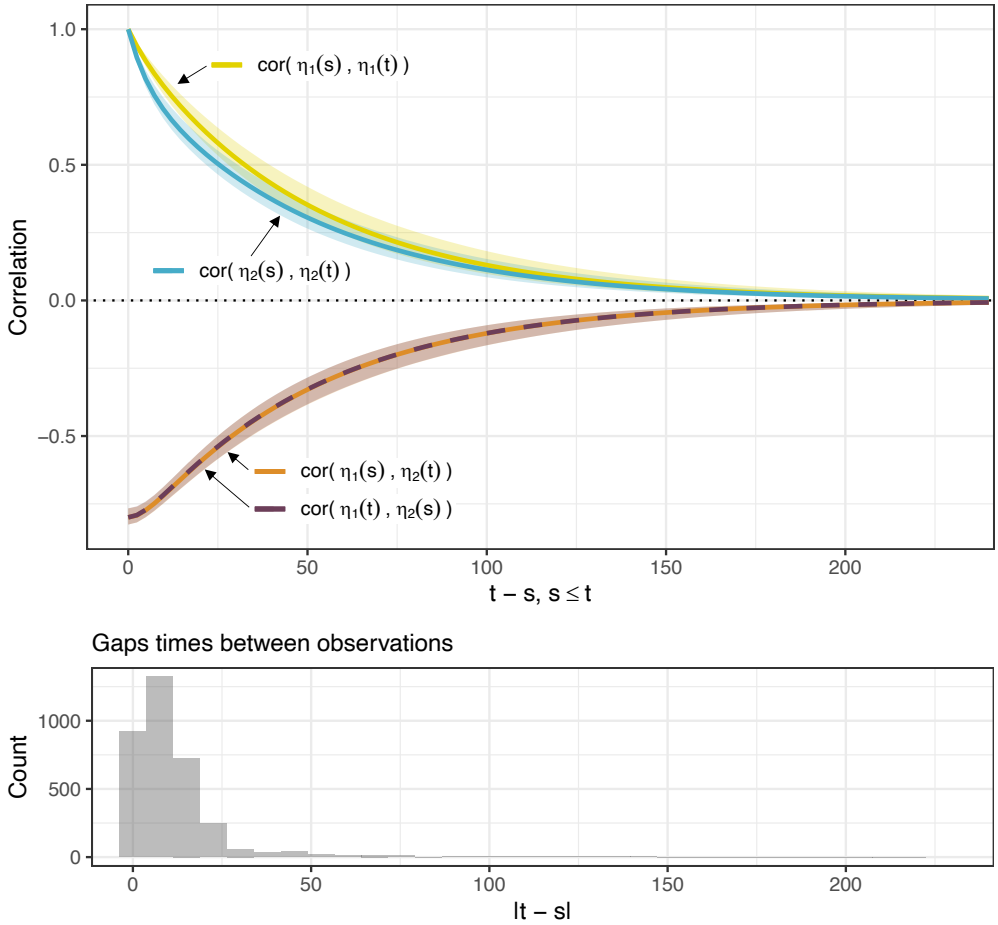


Figure 2.5: The top panel shows the decay in autocorrelation and cross-correlation between latent factors that represent positive affect ($\eta_1(t)$) and negative affect ($\eta_2(t)$) across increasing gap times, where time is measured in hours. Curves are calculated using OU parameters estimated from emotions measured in the mHealth study. The shaded bands indicate the 2.5th and 97.5th percentiles of a parametric bootstrap. The bottom plot summarizes the distribution of the observed gap times (in hours) between measurements for all individuals in the mHealth study.

2.6 Discussion

We developed an estimation method for a dynamic OUF model that combines a factor model to summarize multivariate observed longitudinal outcomes (e.g., emotions) as lower dimensional latent factors (e.g., psychological states) and an OU process to describe the temporal evolution of the latent factors in continuous time. By using the OU process, instead of a discrete time approach such as a VAR process, the model can be applied to irregularly-measured ILD commonly produced by mHealth studies. Importantly, to make the model suitable for the ILD setting, we (i) derive a close-form likelihood for the marginal distribution of the observed longitudinal outcome that integrates over latent variables, (ii) derive the sparse precision matrix for the multivariate OU process, and (iii) leverage a mix of analytic and numeric gradients in our block coordinate descent algorithm. Together, these methodological contributions enable us to use our model to study the short- and long-term dynamics of the intensity of momentary emotions using ILD from an mHealth study. Our derivation of the sparse precision matrix for the multivariate OU process, in particular, enables us to simplify computation and avoid the conditional distribution of the OU process, which is often used in practice but becomes computational costly in the ILD setting. Furthermore, the marginal log-likelihood used in our method makes it more amenable to comparing models using information criteria, such as AIC or BIC. The Bayesian approach for fitting a similar model developed in Tran et al. (2021) [112] uses the conditional likelihood, which has not been marginalized over the latent factors or the random intercepts. This conditional likelihood is quite convenient for Bayesian inference, but is less convenient for information criteria-based comparisons of models. Generally, the marginal likelihood is preferred when calculating information criteria; the use of conditional vs. marginal likelihoods for comparisons of factor models is discussed further in Merkle et al. (2019) [60].

Through the marginal distribution of the multivariate OU process, we parameterize our likelihood in terms of the standard OU drift ($\boldsymbol{\theta}$) and volatility ($\boldsymbol{\sigma}$) parameters. Having estimates for these parameters enables us to understand the dynamics of the latent factors, including generating new trajectories using the estimated values and examining the decay in the trajectories' correlation over time. Through examination of decay in correlation over time, our method could help inform the design of future studies that aim to collect ILD by providing insight into how frequently the longitudinal outcomes must be measured in order to capture the correlation between them. In our simulation study for assessing bias and variance, we generated data under true OU processes that showed reasonably slow decay in correlation over time given the intervals between measurements. We found that estimation of the OU parameters is difficult if correlation decays quickly relative to gaps

between measurements. When longitudinal outcomes are measured frequently enough that correlation between consecutive measurements is captured, our method consistently returns unbiased estimates of the OU process parameters.

In addition to understanding decay in correlation, we can use the OUF model to partition the variance of our observed outcome into contributions from different sources; specifically, the latent factors, random intercepts, and measurement error. Comparing the relative magnitude of these contributions allows us to gain insight into the importance of short-term variations within individuals and long-term differences across individuals. In the motivating mHealth data, these short-term variations are interpreted as emotional states and the long-term differences are interpreted as traits. The results of our analysis of these data suggest that it may be more important to measure certain emotions (e.g., happy and disgusted) frequently if understanding their dynamics is of interest, while other emotions (e.g., proud and lonely) may require less frequent measurements as they are more stable within an individual. EMA questionnaires often ask study participants to respond multiple times per day to numerous questions that assess their current state and context, and so understanding the optimal frequency at which to measure certain outcomes of interest could help reduce the burden on participants.

Because we focus on the analysis of ILD on momentary emotions, behavioral science theories can be used to inform the placement of the structural zeros in the loadings matrix. In a different setting, the relationship between the longitudinal outcomes and the latent factors may be more difficult to specify based on existing domain-specific literature; in this case, extending this work to enable learning of the location of the structural zeros could be a useful. An easy way to use the current method to gain insight into the structure of the loadings matrix would be to use AIC or BIC to compare models that are constant in the number of latent factors but differ in the locations of the structural zeros.

Our model currently assumes that data are missing at random; however, given that our emotions are self-reported, this assumption may not hold in practice. Considering additional methods for modeling informative missingness is another important extension of this work. In many mHealth studies, individuals tend to respond to EMAs less frequently over time due to the gradually accumulating burden. As a result, the missingness mechanism may be different for these individuals, compared to individuals who have intermittent missingness. Thus, modeling various missingness mechanisms could be an important and useful direction for future work.

Although we use the sparse OU precision matrix, leverage the availability of analytic gradients for the measurement submodel parameters, and implement a portion of our algorithm in C++, the computation time of our estimation algorithm increases rapidly as the number

of longitudinal outcomes increases. We successfully fit our model to a dataset containing 18 longitudinal outcomes but this does require approximately 27 hours. In order to make application of our model to datasets with larger numbers of longitudinal outcomes feasible, computational efficiency must be increased. However, our proposed marginal likelihood-based method has substantial computational benefits when compared to alternative methods. In comparison to the Bayesian approach proposed for fitting a similar model in Tran et al. (2021) [112], our approach requires less computation time. In a simulation study with $K = 4$ longitudinal outcomes measured at 10-20 occasions on $N = 200$ individuals, we found that estimation via our block coordinate descent algorithm required approximately 5% of the time required by the Bayesian approach proposed in Tran et al. (2021) [112] given the same computing resources. More details on this comparison are given in Section A.11.

In the simulation study and real data analysis presented in this work, we fit OUF models with one, two, or three latent factors, but the methods presented here extend to models with larger numbers of latent factors. Tran et al. (2021) [112] also focus on models with two or three latent factors. To fit their model, they derive an algebraic constraint on the $\boldsymbol{\theta}$ matrix requiring that $\boldsymbol{\theta}$ has eigenvalues with positive real parts; this constraint results in a latent process that is mean-reverting but possibly oscillating. The authors acknowledge that a limitation of this approach is that this constraint may not be easy to derive for a larger number of latent factors. In our work, we follow the eigenvalue constraint recommended in Tran et al. (2021) [112] but implement this constraint by adding a penalty to the likelihood. This penalty-based approach only requires us to calculate the eigenvalues of the $\boldsymbol{\theta}$ matrix, rather than derive an algebraic solution, and thus it is straightforward to increase the number of factors in our model.

Finally, the mHealth dataset to which we applied our method comes from a smoking cessation study and also contains information on demographic characteristics and on the timing of cigarette use. Including baseline covariates in either the measurement or structural submodel would be a useful extension. In behavioral science, specific emotional states, such as negative affect or craving, are expected to be correlated with cigarette use and so future work could involve combining our OUF model with a submodel for event-time outcomes. Our model could also be modified to account for treatment or for drift in the OU process to better capture the dynamics of the latent processes after a key event such as smoking cessation or relapse.

CHAPTER 3

A Latent Variable Approach to Jointly Modeling Longitudinal and Cumulative Event Data Using a Weighted Two-Stage Method

3.1 Introduction

Widespread use of mobile health (mHealth) technology, ranging from smartphones to wearable devices, has increased the availability of intensive longitudinal data (ILD). Ecological momentary assessment (EMA), which is a data collection method often used in mHealth studies, consists of repeatedly sampling individuals' current states and contexts. This method allows for the collection of data in real time and in individuals' natural environments. When outcomes of interest are measured multiple times per day, researchers can record rich data that capture temporal variations in individuals' current states and contexts. These EMA data can be particularly useful in helping researchers understand what factors (e.g., psychological states or environmental stimuli) represent risk for adverse health events (e.g., smoking, sedentarism, unhealthy eating). We consider the setting in which researchers are interested in examining the association between multiple dynamic latent factors—which are measured through a larger number of observed longitudinal outcomes—and the risk of repeated adverse health events, where these events are indirectly observed as the total number of events across some interval of time.

In many research areas, longitudinal outcomes are measured with error and not necessarily at the same time as the adverse health events. Thus, simply including the longitudinal outcomes as time-varying predictors in a time-to-event model, such as a Cox proportional hazards model, is not appropriate [114]. A more suitable standard alternative is to jointly fit a longitudinal and time-to-event model. However, existing joint longitudinal-survival models are limited in their ability to flexibly model multiple longitudinal outcomes in a computationally efficient manner. This computational challenge is of particular concern in

the setting of ILD when many longitudinal outcomes are measured simultaneously.

Additionally, rather than recording the timing of each individual adverse health event, the mHealth study motivating this work was designed such that we only observe cumulative numbers of events over windows of time, further complicating our setting. Here, we take an important first step to address these challenges—stemming from both the partial observation of the event outcome and the computational complexity due to ILD—by developing a practically useful approach for modeling the association between multiple longitudinal factors and the risk of repeated adverse health events. Our approach aims to take advantage of existing software while at the same time balance flexibility, computational complexity, and scientific utility.

3.1.1 Related Work

Much of the existing work within the field of joint longitudinal-survival models has focused on combining univariate longitudinal processes with a single time-to-event outcome. Modeling multivariate longitudinal outcomes can rapidly increase the computational cost of joint estimation. Aiming to address the computational challenges of joint modeling with multivariate longitudinal data, Rustand et al. (2023) [98] propose an approximate Bayesian method based on the integrated nested Laplace approximation. Adaptation of this technique to fit our specific model is less than straightforward, particularly due to our use of a multivariate continuous time stochastic process. An alternative sampling-based approach to modeling time-varying covariates and survival outcomes was proposed by Rathbun et al. (2013) [83]. This work was motivated by a desire to improve computational efficiency and avoid specification of a model for the longitudinal process; however, it requires many simplifying assumptions—such as lack of measurement error—that do not generally hold with use of EMAs in practice.

An alternative strategy to joint estimation of a longitudinal-survival model for multiple longitudinal outcomes, a two-stage approach is commonly used. Two stage estimation, however, has the well-documented risk of introducing bias into estimates of model coefficients [114, 88, 57]. Incorporating weights has been shown to help reduce this bias. For example, Mauff et al. (2020) [57] recently proposed a two-stage method that leverages weights based on importance sampling within a Bayesian framework.

To lower the high computational cost associated with directly modeling the observed multivariate longitudinal outcomes, we reduce complexity by modeling a smaller number of latent longitudinal factors via a dynamic factor model. Factor models have long been used in the fields of psychology and behavioral science to represent highly correlated outcomes

as simpler underlying states, where these states can be interpreted as affect, motivation, vulnerability, etc. [104, 34, 8]. Numerous variations of latent variable models have been proposed in the statistical literature, many of which involve modeling the evolution of the latent factors over time [77, 64, 96, 23, 76, 121, 54, 107, 78, 65]. Extending this dimension-reduction strategy to the joint model setting, McCurdy et al. (2019) [58], Liu et al. (2019) [53], and Larsen (2000) [47] have proposed combining factor models and hazard models to jointly model longitudinal and time-to-event data. Liu et al. (2019) [53] modeled the evolution of the latent factors using mixed models while McCurdy et al. (2019) [58] and Larsen (2000) [47] did not model correlation over time. Muthén and Muthén (2017) [66] and Asparouhov and Muthén (2018) [6] have implemented methods to fit models like that proposed by Larsen (2000) [47] in their software Mplus but suggest an approach to estimation that relies on numerical integration. We avoid the computational complexity of numerical integration by using a two-stage approach. Furthermore, our use of a lower-dimensional stochastic process to model the evolution of the latent factors falls beyond the scope of the longitudinal models considered in these prior works.

Importantly, our proposed approach differs from the standard joint model setting due to the partially unobservable nature of our event outcome. Rather than observing the time of each event (or censoring), we observe total numbers of events over defined windows of time. As a result, our joint model consists of a longitudinal submodel and an event submodel suitable for this count data—specifically, a Poisson regression model.

3.1.2 Main Contributions and Outline

Overall, there exists a need for further development of statistical methods well-suited for ILD that can connect multiple longitudinal outcomes with the risk of adverse events in a flexible and interpretable manner. Rather than attempting to jointly estimate the longitudinal and risk submodels in this work, we take a two-stage estimation approach, using weights to alleviate potential bias that can result from non-joint estimation. Our main contribution is the development of a scientifically useful approach to modeling multiple highly variable longitudinal outcomes and their associations with recurrent events; we do so by combining a dynamic factor model and a Poisson regression model for cumulative risk via an approximate method that leverages existing software in a practically useful way.

The remainder of this paper is organized as follows: in Section 3.2, we describe the motivating data collected as part of a smoking cessation study that uses EMAs; in Section 3.3, we outline our proposed method; in Section 3.4, we demonstrate its performance via simulation; in Section 3.5, we use our method to analyze a subset of data recorded in the

smoking cessation study; and in Section 3.6, we discuss our results and provide concluding remarks.

3.2 Motivating Data

Data motivating this work come from a longitudinal study examining the influence of contextual, biobehavioral/psychosocial, and demographic factors; social history; and acute momentary precipitants on smoking cessation among 302 African Americans attempting to quit (Break Free II; R01MD010362). Participants answered EMAs on a pre-programmed smartphone from 4 days prior to their quit date through 10 days post-quit. Smartphones were programmed to deliver up to four random EMAs per day. Participants also answered event-contingent EMAs based on whether wearable chest and wrist sensors detected possible smoking or stress [99, 38]. Each EMA assessed smoking behaviors, mood, and other cognitive, interpersonal, and contextual factors. All participants received nicotine patch therapy, self-help materials, and brief quitting advice [28] and were compensated for their time. Detailed information about study procedures have been published in Potter et al. (2023) [74]. Our work uses data that were available at the time of drafting this manuscript from 218 participants and focuses on their psychological state (measured via emotion items) and smoking (measured via the number of cigarettes recently smoked) self-reported using EMAs over the 10 day post-quit period. The total number of random EMAs to which individuals responded over the post-quit period varied from 2 to 47 (average = 17). Due to the non-random delivery of event-contingent EMAs, we only use event outcome (smoking) data from these EMAs and do not use information about the longitudinal process (emotions). Data collected from three individuals are illustrated in Figure 3.1.

An important characteristic of these data is the way that smoking information is collected. Each EMA—whether random or event-contingent—prompts the individual to respond to a series of questions that capture their cigarette use in the period since the previous EMA. We use the data to partition the days into non-overlapping time intervals. In each interval, we either observe the count outcome (taking a 0 or positive value) or we do not observe the count outcome and mark it as unobserved. The unobserved windows include sleep times when the individual is assumed not "at-risk". The partitions during which counts are observed are called "event intervals." A general illustration of this data structure is provided in the Appendix (see Figure B.1).

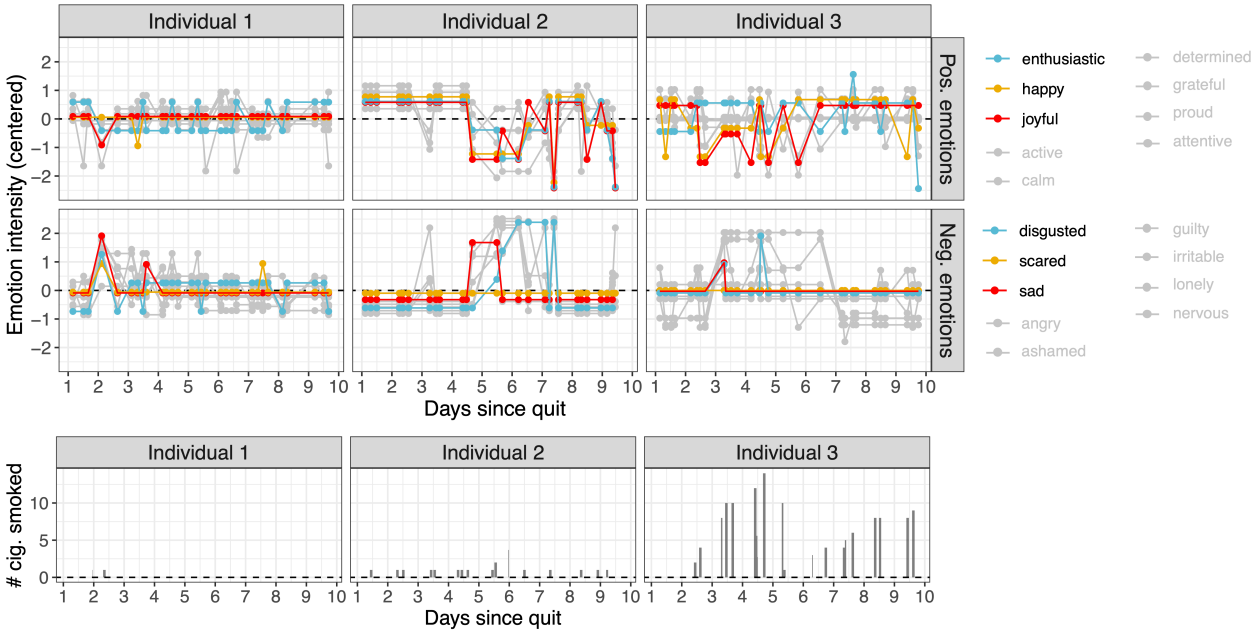


Figure 3.1: Plot of self-reported emotions from the random ecological momentary assessments (EMAs) and self-reported cigarette use from all EMAs for three individuals in the smoking cessation study. A subset of the 18 positive and negative emotions are highlighted for illustrative purposes only. In the plot summarizing reported cigarette use, the width of each bar corresponds to the interval over which an individual reported smoking at least one cigarette and the height of the bar corresponds to the total number of cigarettes reported smoked. Note that any smoking reported at a rate greater than 8 cigarettes per hour has been truncated at 8 cigarettes per hour.

3.3 Methods

We propose a cumulative risk model for the summarized version of the event outcome (e.g., cigarette counts), which we model as a function of the partially latent predictors (e.g., psychological state). Although we could treat each individual EMA item (e.g., emotion) as a measurement of its own latent factor, we instead take a dimension-reduction approach to modeling these observed longitudinal outcomes (e.g., we assume that the emotions are measurements of psychological states capturing positive and negative affect). Advantages of this dimension-reduction approach are two-fold: (i) it improves interpretability of the cumulative risk model while minimizing information loss due to the high correlation between many observed longitudinal measurements and (ii) it increases computational efficiency by using a low-dimensional latent process to summarize the many measured longitudinal outcomes, rather than directly modeling them via a higher-dimensional process.

Overall, our proposed approach consists of two parts: (i) a longitudinal submodel that summarizes the observed longitudinal outcome (e.g., emotions) as lower dimensional dynamic latent factors (e.g., psychological states) and (ii) a cumulative risk submodel that links time-varying latent factors with risk of recurrent events (e.g., cigarette use).

In the description of these submodels, we use the following notation: Suppose that we have N independent individuals, indexed by $i = 1, ..., N$. For each individual, a set of K longitudinal outcomes, $\boldsymbol{X}_i(t_{ij}) = (X_{i1}(t_{ij}), ..., X_{ijK}(t_{ij}))^\top$, is collected at each measurement occasion t_{ij}, where $j = 1, ..., n_i$. This stacked vector of longitudinal outcomes is denoted as $\boldsymbol{X}_i = (\boldsymbol{X}_i^\top(t_{i1}), ..., \boldsymbol{X}_i^\top(t_{ij}), ..., \boldsymbol{X}_i^\top(t_{in_i}))^\top$. At each measurement occasion, individuals also report event intensity data, denoted by $Y_i(t_{ij'})$ for individual i at time j'. Note that we do not require that measurement occasions are evenly spaced or that the set measurement occasions for the longitudinal outcomes $\boldsymbol{X}$ and event intensity outcomes Y be the same within the same individual. Our notation currently assumes that no longitudinal outcomes are missing within a given measurement occasion, but our model could easily be adapted to allow for such a scenario.

3.3.1 Longitudinal Submodel

To summarize the observed emotions as a smaller number of latent psychological states, we take the approach described in Chapter 2. We first use a dynamic factor model to summarize the K observed EMA items (emotions) as a low-dimensional multivariate latent process (representing psychological states), given by

$$\boldsymbol{X}_i(t_{ij}) = \boldsymbol{\Lambda}\boldsymbol{\eta}_i(t_{ij}) + \boldsymbol{u}_i + \boldsymbol{\epsilon}_i(t_{ij})$$

where $\boldsymbol{\Lambda}$ is the $K \times p$ time-invariant loadings matrix; $\boldsymbol{\eta}_i(t_{ij})$ is the value of the p-dimensional latent process at time t_{ij}, where $p << K$; $\boldsymbol{u}_i \sim N(0, \boldsymbol{\Sigma}_u)$ is a random intercept with diagonal covariance matrix $\boldsymbol{\Sigma}_u$; and $\epsilon_i(t_{ij}) \sim N(0, \boldsymbol{\Sigma}_\epsilon)$ accounts for error in the measurement of the longitudinal outcome, with covariance matrix $\boldsymbol{\Sigma}_\epsilon$ also assumed to be diagonal. We make the simplifying assumption that each observed longitudinal outcome loads onto only a single latent factor; this means that each row of $\boldsymbol{\Lambda}$ only contains one non-zero element (and we furthermore assume that the locations of the non-zero elements are known). We make this assumption to clarify the interpretation of the model, but this assumption could be relaxed. Domain knowledge (e.g., behavioral science theories) can be incorporated into the structure of this submodel by using it to specify the location of the non-zero elements within the loadings matrix. We include the outcome-specific random intercepts to account for differences in underlying levels of reported responses across individuals. By using a diagonal $\boldsymbol{\Sigma}_u$, we assume that these levels are independent across outcomes and instead we capture the correlation between the observed responses using a multivariate stochastic process for the latent factors.

The temporal evolution of the p-dimensional latent process is then modeled with a continuous-time multivariate OU stochastic process, which can be thought of as a continuous-time analog of a multivariate autoregressive process. Use of this process allows us to capture correlation within and between multiple latent factors over time. The OU process implies that, for individual i, the marginal distribution of the latent factors across all times $t_{i1}, ..., t_{in_i}$ is

$$\boldsymbol{\eta}_i \sim N(\boldsymbol{0}, \boldsymbol{\Psi}_i)$$

where $\boldsymbol{\eta}_i = (\boldsymbol{\eta}_i^\top(t_{i1}), ..., \boldsymbol{\eta}_i^\top(t_{in_i}))^\top$ and $\boldsymbol{\Psi}_i$ is a $(pn_i) \times (pn_i)$ covariance matrix parameterized by two $p \times p$ matrices of parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$. The exact form of the covariance matrix is given in Section B.2. To make this dynamic factor model identifiable, we model the OU process on the correlation scale; for additional details, see Chapter 2.

If, for example, $p = 2$, then the OU process parameters are

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix} \text{ and } \boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$$

We assume that the off-diagonal elements of $\boldsymbol{\sigma}$ are 0, which effectively forces all correlation between the latent factors to be captured via $\boldsymbol{\theta}$; diagonal elements of $\boldsymbol{\sigma}$ are assumed to be positive. The diagonal elements of $\boldsymbol{\theta}$ must also be positive but the off-diagonals can take positive or negative values; $\boldsymbol{\theta}$ is also constrained to have eigenvalues with positive real parts to ensure that the latent process is mean-reverting [112]. Lastly, we assume that this process has a marginal mean of 0 and that it is stationary.

Together, these two submodels imply that the marginal distribution of the observed longitudinal outcome $\boldsymbol{X}_i$ for individual i is

$$\boldsymbol{X}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i^*) \text{ where } \boldsymbol{\Sigma}_i^* = (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda})\boldsymbol{\Psi}_i(\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda})^\top + (\mathbf{J}_{n_i} \otimes \boldsymbol{\Sigma}_u) + (\mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}_\epsilon)$$

where $\boldsymbol{\Psi}_i$ is the marginal correlation matrix of the OU process parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$, $\mathbf{I}_{n_i}$ is an identity matrix of dimension n_i, $\mathbf{J}_{n_i}$ is an $n_i \times n_i$ square matrix of ones, and $\otimes$ denotes the Kronecker product.

This approach to jointly modeling the observed longitudinal outcomes as dynamic latent factors and modeling the evolution of the latent factors as a multivariate OU stochastic process has the advantage of being both a flexible model and returning unbiased estimates of the parameters that characterized the latent process.

3.3.2 Cumulative Risk Submodel

The coarseness of the adverse health event measurements makes it challenging to estimate the association between latent longitudinal factors and the instantaneous risk of an event. Instead of modeling instantaneous risk, we propose a cumulative risk model for the association between the average value of the latent factors and total number of events across a small window of time, or event interval. Specifically, we propose using a Poisson regression model to connect the latent factors with risk of events.

Let $Y_i(t_{ij})$ denote the cumulative number of events recorded between times $t_{i,j-1}$ and t_{ij}. To model the intensity of events over the interval from $t_{i,j-1}$ to t_{ij} as a function of the latent process, we use

$$\boldsymbol{Y}_i(t_{ij}) \sim Poi\left((t_{ij} - t_{i,j-1})\lambda_i(t_{ij})\right) \text{ where } log(\lambda_i(t_{ij})) = \beta_0 + \beta_\eta^\top \left[\frac{1}{(t_{ij} - t_{i,j-1})} \int_{t_{i,j-1}}^{t_{ij}} \boldsymbol{\eta}_i(s) ds\right]$$

Following the bivariate latent process given as an example in Section 3.3.1, $\boldsymbol{\eta}_i(s)$ is a length-2 vector that contains the values of $\boldsymbol{\eta}_{i1}(s)$ and $\boldsymbol{\eta}_{i2}(s)$ for individual i at time s over the interval from $t_{i,j-1}$ to t_{ij}. We later denote the average value of the latent process over this interval as $\bar{\boldsymbol{\eta}}_i(t_{ij})$. This allows us to re-write our model as

$$Y_i(t_{ij}) \sim Poi\left((t_{ij} - t_{i,j-1})\lambda_i(t_{ij})\right) \text{ where } log(\lambda_i(t_{ij})) = \beta_0 + \boldsymbol{\beta}_\eta^\top \bar{\boldsymbol{\eta}}_i(t_{ij})$$

Because we assume a stationary OU process for the latent factors, we can derive the limiting distribution of this integral, which is also Gaussian. Details on this limiting distribution are given in Section B.3. The association between the cumulative event outcome and $\bar{\boldsymbol{\eta}}_i(t_{ij})$ is captured through coefficient vector $\boldsymbol{\beta}_\eta$.

3.3.3 Estimation

In our two-stage approach to estimation, we first fit the longitudinal submodel using an iterative block coordinate descent algorithm. This block-wise approach was developed in Chapter 2 with the goal of improving the computational efficiency of estimation by leveraging the availability of analytic gradients for a subset of parameters. Using results from the first stage of estimation, we then fit a weighted version of the cumulative risk model via R's `svyglm` package [55] where the weights account for uncertainty in the average values of the latent trajectories. This approach to the second stage of estimation, which is based on a Monte Carlo Expectation Maximization (MCEM) algorithm with importance sampling-based weights, involves iterating between updating the cumulative risk model coefficients and then updating the weights.

We briefly describe the motivation for using these weights: in an ideal setting, we would like to sample values of the latent process from the conditional distribution given the event outcome, $\boldsymbol{\eta}_i|\boldsymbol{Y}_i$. However, sampling from this distribution is difficult. We use weights that are proportional to this conditional distribution to adjust for the fact that, in practice, the values of the latent process are sampled from the marginal distribution, rather than the conditional distribution. By applying these estimated weights to up-weight or down-weight certain values of the latent process sampled in stage 1, we avoid the need to re-sample these values in stage 2. Additional discussion of this approximate approach is provided in Section

B.4.

The two stages that make up the full estimation approach are outlined below, along with a definition of the weights.

Stage 1:

1. Let $\boldsymbol{\Theta}_L = (\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\theta}, \boldsymbol{\sigma})$ be a vector containing all parameter matrices for the longitudinal submodel. Then, initialize parameters in the longitudinal submodel: consider both empirical initial values (based on fitting simpler models) and default initial values and select the set of initial values that correspond to the higher log-likelihood for the longitudinal submodel.
2. Estimate the longitudinal submodel via an iterative block coordinate descent algorithm.
3. Predict the factor scores $\hat{\boldsymbol{\eta}}_i$ using the parameter estimates, denoted $\hat{\boldsymbol{\Theta}}_L$, from the fitted longitudinal submodel where $\hat{\boldsymbol{\eta}}_i = \mathbb{E}\left(\boldsymbol{\eta}_i | \boldsymbol{X}_i; \hat{\boldsymbol{\Theta}}_L\right) = \hat{\boldsymbol{\Psi}}_i \left(\mathbf{I}_i \otimes \hat{\boldsymbol{\Lambda}}^\top\right) \hat{\boldsymbol{\Sigma}}_i^{*^{-1}} \boldsymbol{X}_i$. These factor scores are observed at longitudinal measurement occasions (e.g., times of the random EMAs).
4. Augment observed data by drawing a value for the event interval-specific average of the latent factors given known values of the factors scores, $\hat{\boldsymbol{\eta}}_i$, at each endpoint of the interval; that is, sample $\bar{\boldsymbol{\eta}}_i(t_{ij}) | \hat{\boldsymbol{\eta}}_i(t_{i,j-1}), \hat{\boldsymbol{\eta}}_i(t_{ij}) \sim N\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}\right)$. For the form of $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$, see Section B.3. We refer to these generated average values that augment the observed data as "synthetic values". For each of the N individuals, we consider $r = 1, ..., R$ possible synthetic values per event interval.

At this point, our data consists of two parts: $Y_i(t_{ij})$, the cumulative number of cigarettes smoked in the interval $(t_{i,j-1}, t_{ij}]$; and $\bar{\boldsymbol{\eta}}^{(r)}(t_{ij}), r = 1, \ldots, R$, the set of R possible average values of the latent process across each interval.

Stage 2:

1. Fit the cumulative risk model.
   (a) Initialize coefficients $\boldsymbol{\beta}$ in the cumulative risk model by fitting a model with uniform weights. That is, fit

$$Y_i^{(r)}(t_{ij}) \sim Poi\left((t_{ij} - t_{i,j-1})\lambda_i^{(r)}(t_{ij})\right) \text{ and } log(\lambda_i^{(r)}(t_{ij})) = \beta_0 + \boldsymbol{\beta}_\eta^\top \bar{\boldsymbol{\eta}}_i^{(r)}(t_{ij})$$

   and $Y_i^{(r)}(t_{ij}) = Y_i(t_{ij}) \forall r = 1, ..., R$.

(b) Given estimates $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_\eta^\top\right)$, update the weights, $w_{ij}^{(r)}$. Calculate weights for each set of synthetic latent factor averages (r) where the weight for each event interval j is calculated as

$$\begin{aligned}\widetilde{w}_{ij}^{(r)} &= \Pr\left(Y_i\left(t_{ij}\right) = y \mid \bar{\eta}_i^{(r)}\left(t_{ij}\right), \hat{\beta}\right) \\ &\propto e^{-(t_{ij}-t_{ijj-1})\lambda_i^{(r)}(t_{ij})}\left(t_{ij}-t_{i,j-1}\right)\left(\lambda_i^{(r)}\left(t_{ij}\right)\right)^y\end{aligned}$$

Weights are normalized such that $w_{ij}^{(r)} = \widetilde{w}_{ij}^{(r)} / \sum_r \widetilde{w}_{ij}^{(r)}$ and then again such that $\sum_i \sum_j \sum_r w_{ij}^{(r)} = N$.

(c) Update the coefficients in the cumulative risk model by refitting

$$\begin{aligned}Y_i\left(t_{ij}\right) &\sim \text{Poi}\left(\left(t_{ij}-t_{i,j-1}\right)\lambda_i^{(r)}\left(t_{ij}\right)\right) \\ \text{where } \log\left(\lambda_i^{(r)}\left(t_{ij}\right)\right) &= \beta_0 + \beta_\eta^\top \bar{\eta}_i^{(r)}\left(t_{ij}\right)\end{aligned}$$

using weights $w_{ij}^{(r)}$, which upweight trajectories (indexed by r) that are more likely given the observed event counts and the current cumulative risk model parameter estimates $\hat{\boldsymbol{\beta}}$.

(d) Let k index iterations. Then iterate between steps b. and c. until convergence, where convergence is defined as a small change in the parameter estimates,

$$\left|\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\right| < 1 \times 10^{-6}$$

2. Estimate standard errors using the final parameter estimates $\hat{\boldsymbol{\beta}}$ and weights via the approach described in Section 3.3.3.

When implementing stage 2 in practice, we can assume a quasi-Poisson distribution to account for potential overdispersion in the data.

3.3.4 Inference

Estimation of standard errors for the longitudinal submodel parameters in stage 1 is straightforward using a Fisher information-based approach (see Chapter 2). Our approach to stage 2 estimation, however, presents a challenging setting for inference due to the need to accurately capture additional uncertainty from the synthetic average values of the latent factors and the estimated weights in the cumulative risk model. Here, we focus on inference for stage 2 and present two approaches to standard error estimation often applied in multiple imputation settings that account for uncertainty in parameter estimates in different manners:

standard Rubin's rule for combining variance estimates [97] and a bootstrap-based approach for estimating variance in multiple imputation settings with uncongeniality or model misspecification [120].

When applying Rubin's rule, we generate M imputed datasets, each of which consists of R sets of synthetic average values for the latent factors. We then apply Rubin's rule in a standard manner to combine point estimates and variances across the M imputed datasets. Variance estimates from each imputed dataset are calculated using robust sandwich estimators to account for uncertainty due to weighting of the R synthetic values. Because we require generating multiple average values per event interval to estimate the weights that are used to help alleviate potential bias in our two-stage method, our data do not have the standard structure of multiply imputed datasets. We found that this implementation of Rubin's rule resulted in anti-conservative estimates of standard errors and so attempted to improve our estimates of uncertainty through use of the bootstrap-based approach.

Bartlett and Hughes (2020) [7] summarize various approaches for combining bootstrapping and multiple imputation that aim to improve coverage of confidence intervals in settings of uncongeniality or model misspecification. Among these approaches is a computationally efficient adaptation of bootstrapping followed by imputation that was originally proposed in von Hippel and Bartlett (2021) [120]. In this approach, called the von Hippel approach, data are bootstrapped, multiple imputation is applied to each bootstrapped dataset, and then the variance is estimated as a weighted sum of the mean sum of squares within and between bootstraps. In this adaptation, the authors suggest the multiple for imputation can be low ($M = 2$) for a reasonable number of bootstrapped samples ($B = 200$) and nominal coverage can still be achieved. We apply this method in our setting using the suggested number of replicates for imputation and bootstrapping. As in our implementation of Rubin's rule, we consider a single imputed dataset to consist of R synthetic average values per event interval. To pool point estimates, we simply average across results from all $M \times B$ imputed bootstrapped datasets.

3.4 Simulation Study

We conduct a simulation study to assess the performance of our two-stage method. Motivated by our mHealth data, here we assume $K = 18$ observed longitudinal outcomes (emotions) are measurements of $p = 2$ underlying latent factors. These longitudinal outcomes are observed for $N = 200$ individuals at n_i measurement occasions, where $n_i \sim Uniform(10, 20)$. The gaps between consecutive measurements range from 1 to 2 units of time and the timing of the measurement occasions define the endpoints of the event intervals. To generate complete

data for each individual, we take the following approach:

1. Set true values of $\boldsymbol{\theta}^*$ and $\boldsymbol{\sigma}^*$, which characterize the underlying bivariate OU stochastic process, and generate values of the latent factors at each measurement occasion.

   (a) We consider two sets of true values for $\boldsymbol{\theta}^*$ and $\boldsymbol{\sigma}^*$; one set corresponds to an OU process with higher autocorrelation and the other set corresponds to an OU process with lower autocorrelation.

2. Using the measurements of the true latent factors, generate true values for the observed longitudinal outcome, X_i, also at the measurement occasions.

   (a) In this simulation study, we consider only a single set of true values for the parameters $\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}_u^*, \boldsymbol{\Sigma}_\epsilon^*$ in the longitudinal submodel. We assume that each row of $\boldsymbol{\Lambda}^*$ only contains one non-zero element and that $\boldsymbol{\Sigma}_u^*$ and $\boldsymbol{\Sigma}_\epsilon^*$ are diagonal.

3. Using true values of $\boldsymbol{\theta}^*$ and $\boldsymbol{\sigma}^*$, generate a single set of true average values of the latent process $\bar{\eta}_i^*$, given the values of the factors at the measurement occasions, across each event interval.

4. Then, generate cumulative numbers of events that occurred across each event interval using the average value of each trajectory within the bivariate latent process by drawing from a Poisson distribution with true mean,

$$\lambda_i^*(t_{ij}) = (t_{ij} - t_{i,j-1}) \exp\{\beta_0^* + \beta_1^* \tilde{\eta}_{i1}^*(t_{ij}) + \beta_2^* \tilde{\eta}_{i2}^*(t_{ij})\}$$

   where $\beta_0^*, \beta_1^*, \beta_2^*$ are chosen such that, given $\boldsymbol{\eta}_1^*$ and $\boldsymbol{\eta}_2^*$, the average event rate is close to observed smoking rate in the mHealth data described in Section 3.2.

From these *complete data*, we then assume the *observed data* consist of: (a) the timing of the measurement occasions that define the event intervals, (b) the longitudinal outcomes measured at those occasions, and (c) the cumulative number of events that occurred within each event interval. For the true values of the parameters used to generate these data, see Section B.5. We then carry out the estimation and inference approaches described in Sections 3.3.3 and 3.3.4.

In our simulations, we generate $R = 50$ possible values for the average of the latent factors over each event interval (i.e., $\bar{\boldsymbol{\eta}}_i^{(r)}(t_{ij}), r = 1, ..., R = 50$). When calculating standard errors using the scaled von Hippel approach, we assume $M = 2$ imputed datasets (each of which consist of $R = 50$ synthetic average values per event interval) and $B = 200$ bootstrapped samples, as recommended in Bartlett and Hughes (2020) [7].

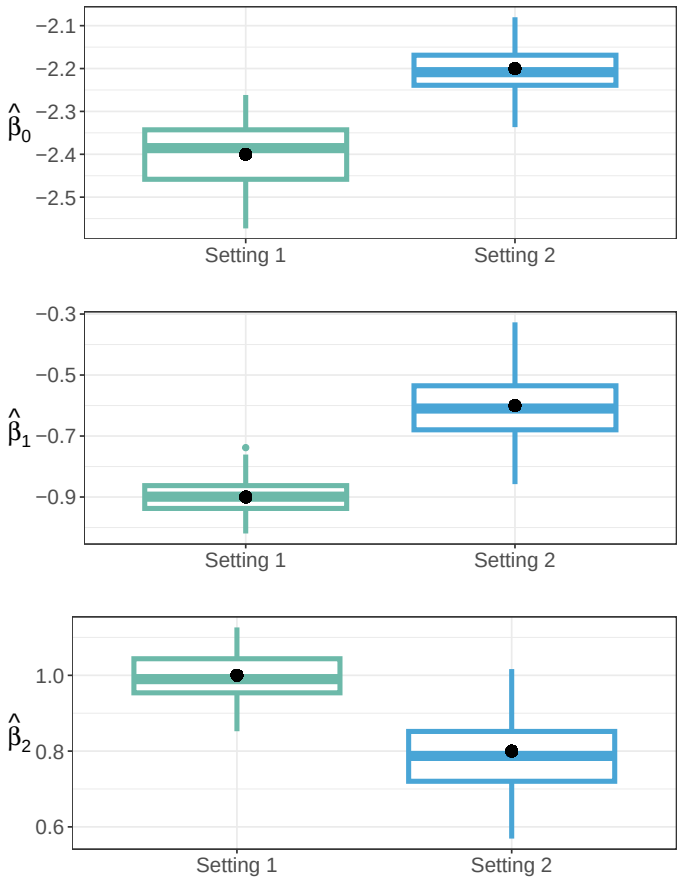


Figure 3.2: Box plots of point estimates of cumulative risk model parameters from simulation study. True values are indicated with black dots. Setting 1 corresponds to data generated from a true stochastic process with low noise and high correlation (i.e., estimation should be easier) and setting 2 corresponds to data generated with high noise and low correlation (i.e., estimation should be more difficult in this setting).

We simulate 100 datasets and summarize results across these replicates. We only consider 100 replicates because bootstrapping is computationally expensive. Point estimates from stage 1 are summarized in Figure B.3. For stage 2, we report point estimates in Figure 3.2 and summarize coverage rates for confidence intervals calculated using standard error estimates from Rubin's rule and the von Hippel approach in Table 3.1. We find that we consistently recover unbiased estimates of the coefficients in all submodels. For the cumulative risk submodel, coverage rates of the confidence interval show that, in our setting, Rubin's rule returns confidence intervals that are occasionally anti-conservative while the von Hippel approach results in confidence intervals with closer-to-nominal coverage. Based on additional empirical results, we find that a large value of R (e.g., $R = 50$) helps reduce bias in our point estimates. Additional discussion of the choice of R is provided in Section B.5.2.

CR	Coef.	Setting 1: high correlation		Setting 2: low correlation	
		RR	VH	RR	VH
80%	β_0	**69**	**71**	80	82
	β_1	82	85	73	81
	β_2	80	82	**67**	73
95%	β_0	**89**	**89**	94	94
	β_1	96	96	91	94
	β_2	95	99	**85**	92

Table 3.1: Coverage rates (%) for 95% confidence intervals for the cumulative risk submodel parameters in the simulation study. Coverage rates (CRs) are averaged across 100 datasets. CRs that fall outside of the range of values expected based on a 95% binomial proportion confidence interval are in bold font. Given that we have calculated the CR for 12 coefficients in each setting (using 80% and 95% confidence intervals), we would expect 0-1 of the CRs to fall outside of the expected range of values from the 95% binomial proportion confidence interval. Abbreviations: Rubin's Rule (RR), von Hippel (VH).

3.5 Application to Smoking Cessation Data

To illustrate this method, we apply it to a subset of data collected from the Break Free II study described earlier (Section 3.2). Our event outcome for the cumulative risk model is the self-reported total number of cigarettes smoked over repeated event intervals, where these intervals of time are defined by individuals' responses to the smoking-related EMA questions. Due to uncertainty surrounding the exact time at which each individual attempts to quit smoking, we restrict the cumulative risk model to use only data that were collected 24 hours after the designated quit date. This approach is intended to minimize the accidental inclusion of cigarette use that preceded quit. Furthermore, if a participant failed to respond to any EMA for a long period of time (i.e., more than 24 hours), we designated them only "at-risk" within the 24 hours prior to the most recent EMA. As a result, we assumed that data were missing-at-random when fitting the cumulative risk model. We also truncated self-reported cigarette use at a maximum rate of 8 cigarettes per hour; this value of 8 was selected based on domain knowledge of how many cigarettes a heavy smoker could reasonably smoke in an hour.

The observed outcomes used for the longitudinal submodel included the intensity of 18 emotions (listed in Figure 3.1). Although each random EMA assessed the intensity of a total of 23 emotions, our illustrative analysis included only 18 emotions due to computational constraints. These 18 emotions were selected by fitting the longitudinal submodel to the 6 emotions with the highest loadings based on a cross-sectional factor model and then gradually

adding emotions from the remaining 17 until computational cost became prohibitive. We assume that the 18 emotions are observed measurements of two latent factors that represent the psychological constructs of positive affect and negative affect [86]. We also assume that structural zeros within the loadings matrix of the longitudinal submodel are known, meaning that positive emotions are measurements of only positive affect and negative emotions are measurements of only negative affect. Thus, correlation between positive and negative affect is captured entirely through the multivariate latent process representing these unobservable affective states. The cumulative risk model is,

$$Y_i(t_{ij}) \sim \text{Poi}\left((t_{ij} - t_{i,j-1})\,\lambda_i(t_{ij})\right)$$

$$\log\left(\lambda_i(t_{ij})\right) = \beta_0 + \beta_1 \text{PA}_i(t_{ij}) + \beta_2 \text{NA}_i(t_{ij}) + \beta_3 q_{ij} + \beta_4 \text{PA}_i(t_{ij}) \times q_{ij} + \beta_5 \text{NA}_i(t_{ij}) \times q_{ij}$$

where PA indicates average positive affect across event interval t_{ij}, NA represents average negative affect, time q_{ij} corresponds to the midpoint time of event interval t_{ij} and is in units of weeks since quit, and the outcome $Y_i(t_{ij})$ has units of cigarettes smoked per hour. As a result of the identifiability constraint on the longitudinal submodel, a one-unit change in PA or NA corresponds to one standard deviation.

We previously applied the stage 1 longitudinal submodel to all available post-quit data in Chapter 2 and use these existing results here. Because we previously considered all post-quit emotion data when fitting the longitudinal submodel, our analyses included $N = 218$ individuals. However, since in the current illustrative analysis we have restricted the data to 24 hours post-quit, our sample size reduces to the subset of $N = 214$ individuals because responses for four individuals were only available in the first 24 hours post-quit. As these individuals contributed such limited information, we do not believe that the impact of including/excluding their emotion data when estimating the population-level parameters of the longitudinal submodel would substantially impact the results. Point estimates and confidence intervals for the longitudinal submodel are given in the Appendix (Figure B.6). Thus, we use results from previous application of the longitudinal submodel as our stage 1 results here.

In our estimation algorithm, we generate $R = 50$ synthetic average values of the latent factors for each event interval. For inference, we assume $M = 2$ and $B = 200$ as in the simulation study.

In Figure 3.3, we report point estimates and confidence intervals (on the log scale) from the fitted cumulative risk model. These confidence intervals are based on standard errors from the von Hippel approach. The point estimates indicate that above average levels of negative affect (where a value of 0 corresponds to the average) are associated with increased

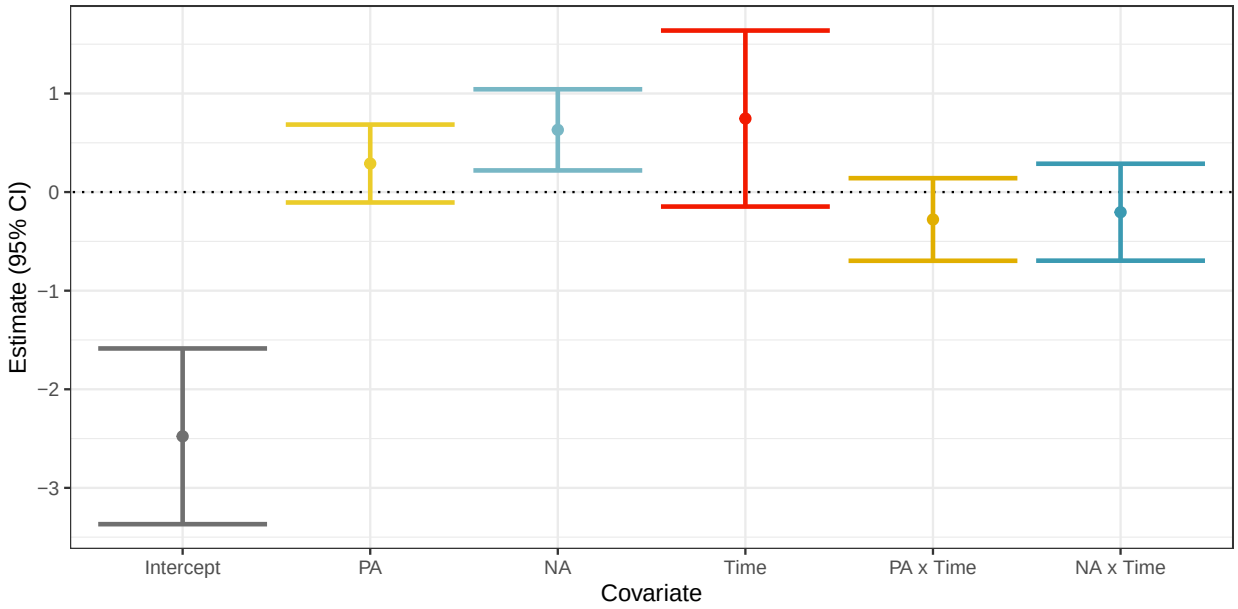


Figure 3.3: Point estimates and 95% confidence intervals for cumulative risk model parameters applied to mHealth smoking cessation study data. Error bars correspond to 95% confidence intervals calculated from von Hippel standard error estimates using $R = 50$, $M = 2$, and $B = 200$. Abbreviations: positive affect (PA), negative affect (NA).

smoking but the association between negative affect and smoking decays over time. The positive association between higher negative affect and increased smoking could be explained by smoking being used as a tool for coping with high stress or negative feelings. But given the observational nature of these data, and since the relative ordering of affective state and cigarettes smoked is unknown, we can only draw conclusions about associations and not causal effects. As such, other explanations may be plausible as this application does not distinguish between affect as a driver of cigarette use vs. changes in affect resulting from cigarette use.

Figure 3.4 illustrates trends in the association between smoking and positive and negative affect over time. This figure allows us to evaluate changes over time in expected smoking rates across hypothetical individuals with (a) average positive and negative affect (PA = 0, NA = 0), (b) above-average positive affect and below-average negative affect (PA = 1, NA = -1), or (c) below-average positive affect and above-average negative affect (PA = -1, NA = 1). Based on these results, the difference in expected smoking rates by affective state increases over time; however, all 95% confidence intervals overlap. Note that confidence intervals presented in this figure are point-wise confidence intervals calculated using standard errors from the von Hippel approach.

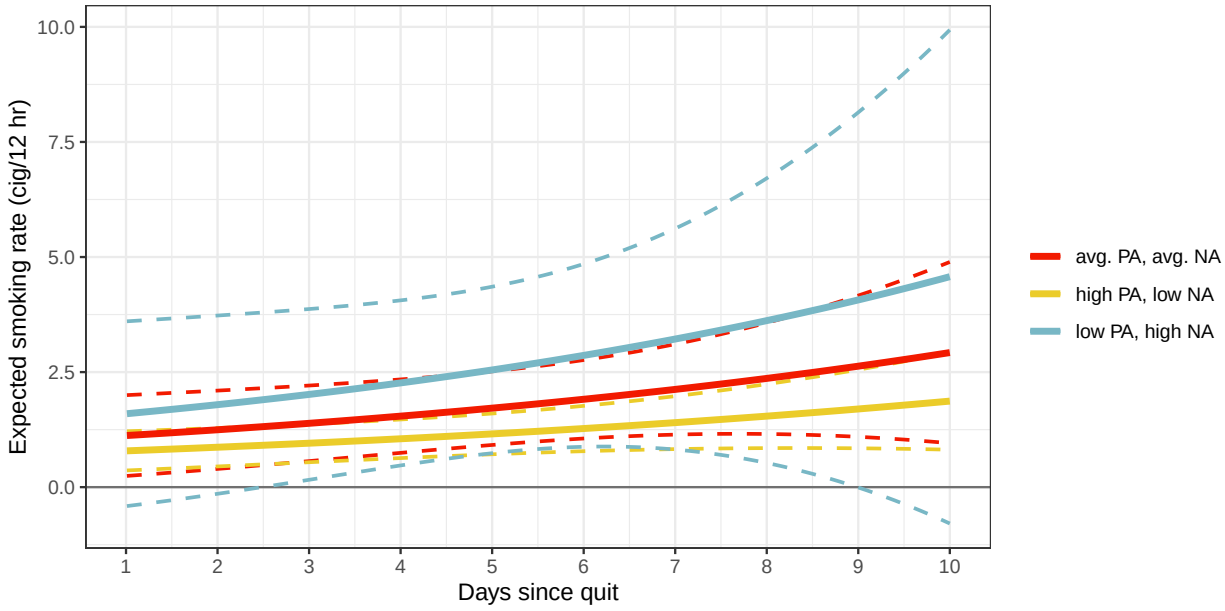


Figure 3.4: Expected smoking rate (in units of cigarettes per 12 hour interval) using estimated coefficients and assuming constant values of either (a) average positive affect (PA = 0) and average negative affect (NA = 0), (b) above-average positive affect (PA = 1) and below-average negative affect (NA = -1), or (c) below-average positive affect (PA = -1) and above-average negative affect (NA = 1). This figure illustrates the difference in expected smoking rates at a fixed time point under different hypothetical levels of positive and negative affect. Dotted lines indicate 95% point-wise confidence intervals estimated using von Hippel standard errors.

3.6 Discussion

In this paper, we proposed a two-stage approach for analyzing longitudinal EMA data to determine how longitudinal outcomes (in our case, emotions) are associated with increased risk of repeated events (here, intensity of cigarette use). We present a two-stage approach to fitting a longitudinal submodel—a dynamic factor model—that summarizes the time-varying dynamics of multiple measured longitudinal outcomes as a lower dimensional stochastic process and a cumulative risk model—a Poisson regression model—that captures the association between latent factors represented by the stochastic process and a count outcome. We use weights to address the partially unobservable nature of both the outcome and predictors and to reduce potential bias in our two-stage approach to estimation.

Readers may be concerned that we take a two-stage approach to fitting our model, which is commonly known to result in biased estimates of coefficients. We argue, however, that the focus of this paper is on the development of a practical method and this two-stage approach allows us to leverage existing methods and software. We include weights in our cumulative risk submodel, which seems to help alleviate bias. Readers might also express concern about our use of MCEM, which relies on approximations rather than direct maximization of the likelihood from our model. Because we assume that our count outcome follows a Poisson distribution, the complexity of our likelihood increases. We take advantage of the approximations with MCEM to greatly simplify the computations involved in the second stage of estimation. Our empirical investigations indicate that these approximations are close enough to yield unbiased estimates in our setting; however, future work could include the development of a computationally efficient approach for directly maximizing the likelihood in the two-stage approach or jointly estimating the longitudinal and cumulative risk submodels.

One might argue that if computational cost is a concern that inhibits use of joint models in practice, then why not fit a three-part model consisting of a cross-sectional factor model, a stochastic process fit to factor scores, and then the risk model? Through simulation, we found that treating the predicted factor scores as observed values of the latent process and then trying to estimate the OU process using those predicted factor scores resulted in poor estimates of the stochastic process parameters and thus biased coefficient estimates in our risk model. Stage 1 estimation does contribute a large portion of our method's computation time (see Figure B.5 in the Appendix for computation by stage). However, the method that we use for the first stage of estimation was shown in Chapter 2 (Section A.11) to require substantially less computation time than an alternative approach to fitting a similar model proposed in [112]. Although both the first stage of estimation and the bootstrapping within the second stage of estimation requires a significant amount of time,

stage 2 only takes approximately 6 minutes to return point estimates in the simulation study. Our ability to use off-the-shelf software in stage 2 means that the computational cost of this stage is low enough to easily allow for estimation of bootstrap-based standard errors. That is, while the bootstrap-based method for standard error estimation is computationally demanding, accurate estimation of standard errors is only possible because each subroutine (i.e., fitting each weighted Poisson model) is fast enough and easy to implement. Thus, exploration of various combinations, interactions, and parameterizations of the latent states in the cumulative risk model is feasible from a practical point of view.

Through simulation, we found that taking a standard likelihood-based approach to inference of the cumulative risk submodel parameters via Fisher information resulted in underestimates of standard errors and thus confidence intervals with low coverage. To improve our ability to fully capture the uncertainty of these point estimates, we then tried adjusting our standard error estimates via an application of Rubin's rule and the von Hippel approach. Our estimation approach does not fit into standard scenarios addressed by the Rubin's rule and so this approach resulted in occasional under-coverage. We found that the bootstrap-based von Hippel approach was slightly better at accurately capturing the uncertainty in our point estimates, although future work could include developing an approach that further improves our ability to quantify uncertainty. Through simulation studies, we found that coverage of confidence intervals is not sensitive to the choice of R. Development of an alternative method for standard error estimation could be a useful direction for future work, as it would avoid the added computation time of bootstrapping.

In our formulation of the longitudinal dynamic factor model, we assume that the locations of structural zeros within the loadings matrix are known. Behavioral science theories regarding how emotions relate to each other can be used to justify the locations of these zeros in our setting (e.g., [95, 84, 85, 86, 32, 59]). While our work could be extended to allow for estimation of an entirely non-null loadings matrix, this extension would present additional challenges related to model identifiability and interpretation; however, it would likely be useful in other setting in which scientific evidence for informing the structure of the loadings matrix is limited. In addition to specifying the locations of the structural zeros, we also require the number of latent factors (p) to be pre-specified. In our motivating application, use of two latent factors is a natural choice. However, if selecting the optimal number of latent factors for summarizing the observed longitudinal data is of interest, AIC and BIC could be used; we suggest using BIC as it tends to error on the side of selecting a slightly simpler model. For more discussion on using AIC and BIC to compare longitudinal submodels with different numbers of latent factors, please see Chapter 2. A further extension could integrate a data-driven approach for selecting the number of latent factors into the

estimation algorithm.

Importantly, our proposed method does not allow us to make any inference about the order of the values of latent affect states and occurrence of adverse health events; the method only provides information about associations. This limitation results, in part, from the modeling decision that we made to account for the partially unobservable nature of both the events of interest and longitudinal predictors recorded in the data. The observational nature of the data limits the ability to make causal conclusions about how affect impacts smoking, regardless of how the models are fitted (i.e., jointly or separately). Nonetheless, this method enables investigators to model multiple dynamic states of risk and their correlation with adverse health events in a way that can advance behavioral theory and generate hypotheses that can be tested in future research. For example, the results can guide the construction of JITAIs [67, 70] designed to target the newly identified states that are correlated with increased risk, and inform the development of MRTs that investigate the utility of delivering (vs. not delivering) these interventions.

CHAPTER 4

A Bayesian Joint Longitudinal-Survival Model with a Latent Stochastic Process for Intensive Longitudinal Data

4.1 Introduction

mHealth technology enables researchers to record longitudinal changes in a variety of biomedical indicators that capture temporal variations in harder-to-measure underlying states. The potentially high frequency of these measurements allows researchers to gain insight into short- and long-term patterns of change in underlying health-related states, such as mood, cognitive function, or disease severity. Here, we use the existing term "intensive longitudinal data" (ILD) to refer to data consisting of multiple outcomes recorded frequently over time. When these ILD are combined with information on the occurrence of time-to-event outcomes, the ILD can provide insight into factors that elevate the risk of an event. Motivated by ILD and event-time data collected in an mHealth study of smoking cessation, we propose a novel approach for jointly modeling a time-to-event outcome and multiple frequently measured—and possibly rapidly varying—longitudinal outcomes. The key contribution of this work is the development of a joint model suitable for analyzing multivariate ILD. Specifically, we use a multivariate continuous-time stochastic process to (a) flexibly model a smaller number of highly variable latent factors measured through a larger number of longitudinal outcomes and (b) represent risk of a time-to-event outcome by incorporating the latent factors as a time-varying covariates in a hazard model.

4.1.1 Related Work

Joint longitudinal-survival models are powerful tools for enabling estimation of the association between temporal variations in longitudinal outcomes and the risk of time-to-event

outcomes. A classic joint longitudinal-survival model consists of two parts: a longitudinal submodel and a survival submodel. Joint models attempt to account for the intermittent measurement of the longitudinal outcomes, measurement error, and informative drop-out. For a comprehensive review of joint models, see Tsiatis and Davidian (2004) [114]. A major challenge to the use of joint models in practice is their computational cost, which rises rapidly as the number of longitudinal outcomes increases. This increasing cost is due to the need to evaluate complex and often intractable integrals across the unobserved random effects in the longitudinal submodel, as well as in the survival function.

Existing literature for jointly modeling multivariate longitudinal outcomes and time-to-event outcomes contains a variety of strategies for dealing with long computation times. These strategies work within both frequentist and Bayesian frameworks. Variations of the two-stage approach—which involves first fitting the longitudinal submodel and then incorporated predicted values (e.g., BLUPS) from the longitudinal submodel as time-varying covariates in the survival submodel—have often been used in settings with multivariate longitudinal outcomes due to the computational speed (e.g., [49, 102, 44]). A well-known drawback of the two-stage approach is the risk of bias in coefficient estimates and so adaptations with bias corrections have also been proposed (e.g., [5, 24, 57]). Despite the lower computation time required by the two-stage approach, most existing work has not focused on the ILD setting and so approaches have not been developed specifically for large numbers of longitudinal outcomes.

As an alternative to the two-stage approach, strategies for joint estimation have also been developed for modeling multiple longitudinal outcomes and a time-to-event outcome. Many of these existing approaches have leveraged dimension-reduction tools to help lower computation time. Li and Luo (2019) [49] and Li et al. (2021) [50], for example, proposed joint models in which multiple longitudinal outcomes are summarized using variations of functional principal components analysis (PCA). Factor models, and related approaches such as item response models, have also been used to reduce the dimension of the longitudinal outcomes in the joint model setting; e.g., [35, 52, 43]. In these instances, the latent factors that summarize the multiple longitudinal outcomes are then used as time-varying covariates in the survival submodel.

Regardless of whether a dimension-reduction strategy is used to help handle the multiple longitudinal outcomes, a longitudinal submodel must also be specified (either for the latent factors representing summary states or for the observed longitudinal outcomes directly). Simple longitudinal submodels allow for easier integration within the joint estimation framework but with ILD, the larger number of longitudinal measurements allows for specification of a more flexible—and potentially complicated—longitudinal submodel. Numerous spline-

based approaches have been developed as a flexible way to model the longitudinal process; e.g., [13, 63, 90, 43, 50, 103, 108, 44, 126]. Gaussian processes have also been incorporated into the longitudinal submodel to capture important patterns, such as serial correlation; e.g., [78, 37].

Although substantial developments have been made in methods for jointly modeling multivariate longitudinal outcomes and survival data, little of this work has focused specifically on the setting of ILD. Rathbun et al. (2013) [83] propose an alternative sampling-based approach for jointly modeling multiple longitudinal outcomes and a time-to-event outcome. Their work is motivated by data collected in an mHealth study and is suitable for analyzing ILD. To deal with the high computational cost of ILD, they avoid specifying a longitudinal submodel altogether through a sampling-based approach. While this approach is computationally fast, the lack of a longitudinal submodel inhibits modeling of measurement error, which we believe is important to account for in our—and many other—settings. More recently, Wong et al. (2022) [126] developed an EM-based approach for non-parametric maximum likelihood estimation of a flexible spline-based joint model. Although this approach was not motivated by ILD, the authors suggest that their method would work with many longitudinal outcomes. This approach, however, does not involve any dimension-reduction of the observed longitudinal outcomes, which are then modeled non-parametrically with splines. In the ILD setting, summarizing the many—and possibly highly correlated—longitudinal outcomes as scientifically meaningful dynamic latent factors has the potential to improve the interpretability of states associated with changes in the risk of an event.

4.1.2 Main Contributions and Outline

We propose a joint model for ILD that is novel in its specific combination of three submodels. As in existing literature, we take a dimension reduction strategy: rather than using PCA, we use a factor model as it allows more incorporation of scientific understanding into the structure of the model and into the interpretation of the latent factors themselves. Rather than using splines to model the change in multiple latent factors over time, we use a continuous-time multivariate stochastic process; this approach allows us the flexibility to capture abrupt changes in multiple correlated latent factors over time but avoids the complexity of specifying the number and location of knots as in a spline-based approach. We then incorporate the latent factors as time-varying covariates in a hazard regression model. To fit our model, we take a Bayesian approach, which allows us to avoid the need to evaluate complex integrals over a multivariate continuous-time stochastic process. Altogether, this approach enables joint modeling of multivariate ILD and a time-to-event outcome via the novel combination

of a dynamic factor model, multivariate stochastic process, and hazard regression model. To the best of our knowledge, the combination of these three submodels with an estimation approach suitable for ILD does not exist in the current literature.

The remainder of this paper is organized as follows: in Section 4.2, we briefly introduce the mHealth smoking cessation study motivating this work; in Section 4.3, we describe our joint model and a corresponding strategy for estimation and inference; in Section 4.4, we demonstrate the performance of our method via simulation; in Section 4.5, we use our method to analyze data from the smoking cessation study; and in Section 4.6, we provide a discussion.

4.2 Motivating Data

Data motivating this work come from a Houston-based mHealth study. This longitudinal observational cohort study, which ran between 2005 and 2007, followed established smokers for four weeks after they attempted to quit smoking. This study, called CARE, has been previously described in other publications (e.g., [14, 119]). During the study, the current state and context of individuals were assessed in real time using EMAs. These EMAs were carried out via surveys sent to mobile Palmtop Personal Computers that prompted individuals to respond to a series of questions capturing their current emotional state and recent cigarette use, among a variety of other social and contextual factors. These EMAs were intended to be sent randomly at four occasions each day. In addition to random EMAs, individuals were instructed to self-initiate EMAs in certain situations (e.g, when feeling a strong urge to smoke, or immediately before or after smoking). We only use information on the longitudinal emotional states reported in the random EMAs. During the four-week post-quit period, individuals responded to an average of 34.3 random EMAs (median = 21.5, range = 1–122). We analyze the 5-point Likert scale responses to the set of nine questions that assessed the current intensity of six negative emotions and three positive emotions over time. The association between smoking and both positive and negative emotions is well-documented in the behavioral science and smoking cessation literature; for example, Vinci et al. (2017) [119] show that positive emotions are associated with a lower likelihood of smoking lapse and Potter et al. (2023) [73] demonstrate that negative emotions are associated with a higher risk of lapse. As such, we aim to use our model to investigate the association between positive and negative affect (a psychological concept related to mood), as captured by the nine emotions measured longitudinally, and the time-to-event outcome of first smoking lapse after attempted quit. Longitudinal responses for one individual are plotted in Figure 4.1.

We define time-to-lapse as the time until the first episode of cigarette use after attempted

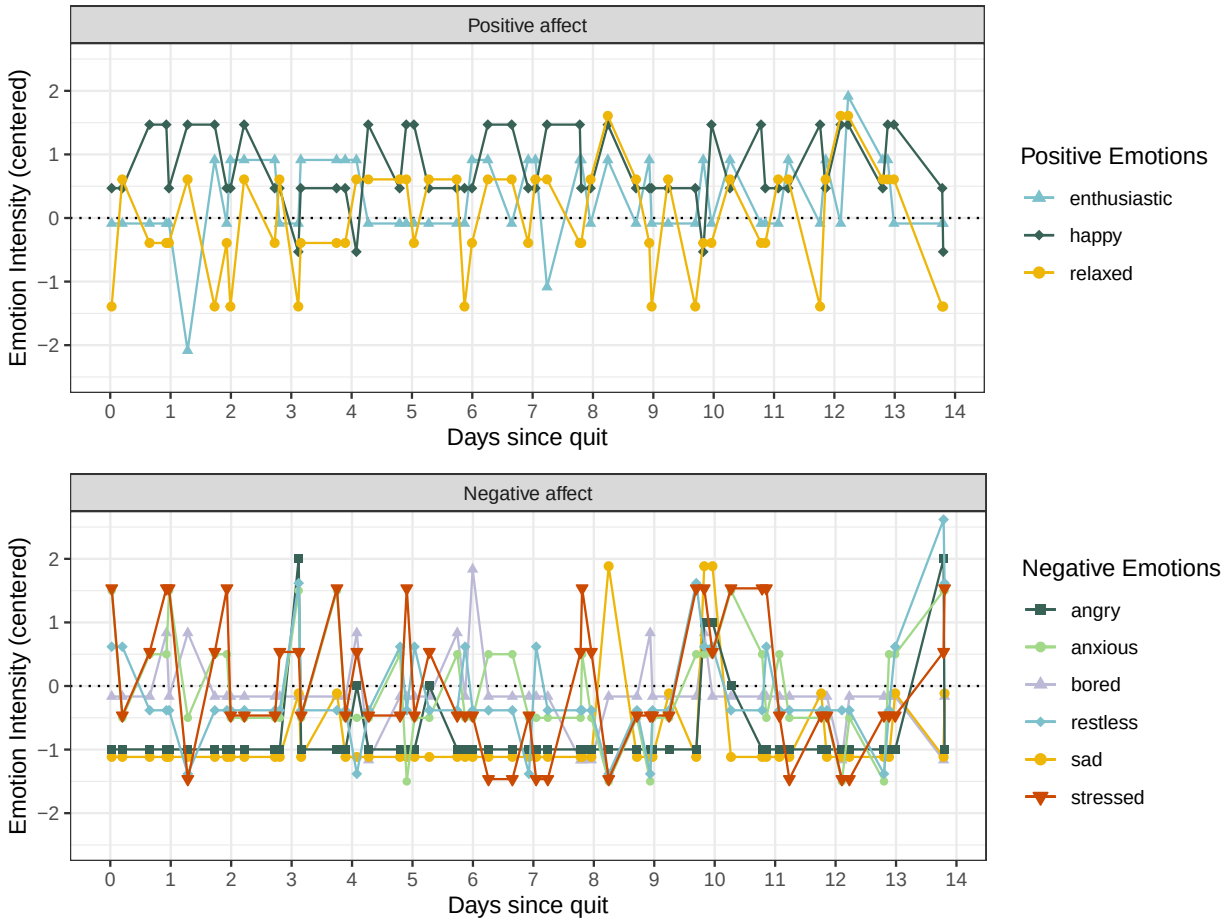


Figure 4.1: Longitudinal responses to the 9 emotion-related questions for one individual in the smoking cessation study. This individual experienced their first post-quit lapse on day 13.9.

quit. To determine the timing of this event, we use information about cigarette use collected from both the random and self-initiated EMAs. Due to uncertainty in the exact time of quit, we restrict our analysis to the subset of individuals who do not report any cigarette use in the first 12 hours after quit (i.e., within 12 hours of the pre-specified 4am quit time on their recorded quit day). Our analytic sample consists of 238 individuals who also responded to the emotion-related questions in at least one random EMA after the first 12 hours of the study. The time point of 4pm on the recorded quit date serves as time zero in our analysis. In the four weeks of follow-up, 71% of individuals are observed to have lapsed; the remaining 29% are censored either at the time of the final EMA to which they responded or at the end of the study. A Kaplan-Meier plot of time-to-lapse is presented in Figure C.9.

4.3 Methods

Our proposed joint model consists of three submodels: (i) a measurement submodel, (ii) a structural submodel, and (iii) an event-time submodel. Before describing these models in detail, we first define some notation. Suppose that our data contain information on $i = 1, ..., N$ individuals. For individual i, longitudinal outcomes $\boldsymbol{Y}_i(t_{ij})$ are measured at occasions $j = 1, ..., n_i$, where $\boldsymbol{Y}_i(t_{ij})$ is a vector of length K containing all measurements of the K longitudinal outcomes at time t_{ij}. Let $\boldsymbol{Y}_i$ be a $(K \times n_i)$-length vector that contains all measurements of the longitudinal outcomes over the n_i occasions. We assume that the longitudinal outcomes are (possibly noisy) observations of a smaller number of p underlying states, represented by p-length vector $\boldsymbol{\eta}_i(t_{ij})$. We let T_i and δ_i denote the observed event time and censoring indicator for individual i, where $T_i = min(\tilde{T}_i, C_i), \delta_i = I(\tilde{T}_i \leq C_i)$ using $\tilde{T}_i$ as the true event time and C_i as the censoring time.

4.3.1 Measurement Submodel

To model the set of K longitudinal outcomes observed for individual i at time t_{ij}, we use a dynamic factor model. This model is closely related to that developed in Tran et al. (2021) [112] and was also previously presented in Chapter 2. The dynamic factor model is written as $\boldsymbol{Y}_i(t_{ij}) = \boldsymbol{\Lambda}\boldsymbol{\eta}_i(t_{ij}) + \boldsymbol{u}_i + \boldsymbol{\epsilon}_i(t_{ij})$, where $\boldsymbol{\Lambda}$ is a $K \times p$-dimensional loading matrix and $\boldsymbol{\eta}_i(t_{ij})$ is p-length vector containing the current values of the p latent factors at time t, with $p << K$. We make the simplifying assumption that $\boldsymbol{\Lambda}$ contains structural zeros and that the location of these structural zeros are known; that is, we assume that we know which of the longitudinal outcomes is a measurement of which of the p latent factors and that each longitudinal outcome measures only a single latent factor. In the motivating study, this assumed structure of $\boldsymbol{\Lambda}$ is supported by behavioral science theories that relate certain emotions with certain underlying psychological states. To account for the correlation in repeated measurements, we include $\boldsymbol{u}_i \sim N_K(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ as an item-specified random intercept. $\boldsymbol{\epsilon}_i(t_{ij}) \sim N_K(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$ accounts for measurement error. We assume that $\boldsymbol{u}_i$ and $\boldsymbol{\epsilon}_i(t_{ij})$ are independent and that $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\epsilon$ are diagonal matrices. We include the random intercept $\boldsymbol{u}_i$ to account for differences in underlying levels of the measured longitudinal outcomes across individuals but then use the structural submodel to model correlated change over time.

4.3.2 Structural Submodel

The longitudinal evolution of the p latent factors is assumed to follow a p-dimensional multivariate OU stochastic process. The OU process can be thought of as a continuous-time

version of a multivariate autoregressive process and thus is suitable for modeling data with unevenly spaced measurement occasions. We assume that the OU process is stationary with a marginal mean of 0 and is parameterized by two $p \times p$-dimensional matrices, $\boldsymbol{\theta}_{OU}$ and $\boldsymbol{\sigma}_{OU}$. To ensure a mean-reverting OU process, $\boldsymbol{\theta}_{OU}$ is required to have eigenvalues with positive real parts, as discussed in [112]. $\boldsymbol{\sigma}_{OU}$ must have all positive elements. Assuming that the initial value of the latent process, $\boldsymbol{\eta}_i(t_{i1})$, is drawn from $N_p(\mathbf{0}, \boldsymbol{V})$, where $\boldsymbol{V} = vec^{-1}\left\{(\boldsymbol{\theta}_{OU} \oplus \boldsymbol{\theta}_{OU})^{-1} vec(\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top})\right\}$, then for $j = 2, ..., n_i$,

$$\boldsymbol{\eta}_i(t_{ij})|\boldsymbol{\eta}_i(t_{i,j-1}) \sim N_p\left(e^{-\boldsymbol{\theta}_{OU}(t_{i,j}-t_{i,j-1})}\boldsymbol{\eta}_i(t_{i,j-1}), \boldsymbol{V} - e^{-\boldsymbol{\theta}_{OU}(t_{i,j}-t_{i,j-1})}\boldsymbol{V}e^{-\boldsymbol{\theta}_{OU}^{\top}(t_{i,j}-t_{i,j-1})}\right)$$

Together, the measurement and structural model imply that

$$\boldsymbol{Y}_i(t_{ij})|\boldsymbol{\eta}_i(t_{ij}), \boldsymbol{u}_i \sim N_K\left(\boldsymbol{\Lambda}\boldsymbol{\eta}_i(t_{ij}) + \boldsymbol{u}_i, \boldsymbol{\Sigma}_{\epsilon}\right)$$

When describing the joint longitudinal-survival model that will capture the risk of event-time outcomes as a function of the time-varying latent factors, we will refer to the combined structural and measurement submodels as our longitudinal submodel.

4.3.3 Survival Submodel

We take a parametric approach to modeling the risk of an event and use the time-varying values of the latent factors to capture vulnerability to the outcome of interest. We define our hazard model as: $h_i\left(t|\mathcal{H}_i(t)\right) = h_0(t;\boldsymbol{\gamma}) \exp\left\{f(\mathcal{H}_i(t);\boldsymbol{\beta}) + \boldsymbol{\alpha}\boldsymbol{X}_i\right\}$ where $\mathcal{H}_i(t) = \{\boldsymbol{\eta}_i(s), 0 \leq s \leq t\}$ is the history of the latent process up until time t, $h_0(t)$ is a parametric baseline hazard function with parameter vector $\boldsymbol{\gamma}$, and $\boldsymbol{X}_i$ is a vector of baseline covariates with coefficients contained in $\boldsymbol{\alpha}$. We write the hazard as a general function of the history of the latent process, $f(\mathcal{H}_i(t);\boldsymbol{\beta})$, to allow for flexibility in how the association between the instantaneous risk of an event and the latent process is modeled. For example, we could simply use $f(\mathcal{H}_i(t);\boldsymbol{\beta}) = \boldsymbol{\beta}^{\top}\boldsymbol{\eta}_i(t)$ or we could choose a more complicated function such as $f(\mathcal{H}_i(t);\boldsymbol{\beta}) = \boldsymbol{\beta}^{\top}\int_s^t \boldsymbol{\eta}_i(u)du$.

4.3.4 Likelihood

We make the following assumptions, which are variations of assumptions standard in the joint longitudinal-survival literature: (i) the timing of the longitudinal measurements and censoring is non-informative; (ii) given the random effects and latent factors, the observed longitudinal outcomes and time-to-event outcomes are independent; (iii) conditional on the random effects and latent factors, the observed longitudinal outcomes within an individual

are also independent across time; (iv) the latent factors, random effects, and measurement error are independent. Using these assumptions, the joint log-likelihood of our observed data can be written as

$$logp(\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{Y}; \boldsymbol{\Theta}) = \sum_{i=1}^{N} log\Bigg\{ \int p(T_i, \delta_i|\boldsymbol{\eta}_i; \boldsymbol{\Theta}_T) \left[\int p(\boldsymbol{Y}_i|\boldsymbol{\eta}_i, \boldsymbol{u}_i; \boldsymbol{\Theta}_M) p(\boldsymbol{u}_i; \boldsymbol{\Theta}_M) d\boldsymbol{u}_i \right] \cdot p(\boldsymbol{\eta}_i; \boldsymbol{\Theta}_S) d\boldsymbol{\eta}_i \Bigg\} \tag{4.1}$$

where $p(T_i, \delta_i|\boldsymbol{\eta}_i, \boldsymbol{\Theta}_T) = h_i(T_i|\mathcal{H}_i(t))^{\delta_i} exp\left\{ -\int_0^{T_i} h_i(s)ds \right\}$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_T, \boldsymbol{\Theta}_M, \boldsymbol{\Theta}_S)$ contains all unknown parameters, with $\boldsymbol{\Theta}_T = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$, $\boldsymbol{\Theta}_M = (\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon)$, and $\boldsymbol{\Theta}_S = (\boldsymbol{\theta}_{OU}, \boldsymbol{\sigma}_{OU})$.

The main challenges to fitting our model stem from two integrals in the likelihood: one over the multivariate OU stochastic process and another within the survival function. For the integral over the multivariate OU process, we could use Monte Carlo integration or the fully exponential Laplace approximation [93]; instead, we opt for a fully Bayesian approach. We use Hamiltonian Monte Carlo (HMC) sampling as implemented in the software Stan [18]. In the following paragraphs, we address additional challenges stemming from the large number of latent variables in our model, identification of the longitudinal submodel parameters, and calculation of the survival function.

Reducing the number of latent variables: We could write the distribution of the observed longitudinal outcome conditional on all latent variables (i.e., the latent factors and the random intercept) as in Equation 4.1. However, attempting to estimate so many latent parameters within Stan is challenging and we found that using the likelihood in Equation 4.1 resulted in Monte Carlo samples with very poor mixing and high computational cost. With this parameterization of the likelihood, the number of parameters also increases by $N+2$ for each additional longitudinal outcome, which is potentially problematic in the ILD setting. To reduce the number of parameters that we need to sample, we can instead integrate the distribution of the observed longitudinal outcome over the random intercept so that the likelihood is conditional only on the latent factors. That is, the longitudinal component of the likelihood in our joint model is $\boldsymbol{Y}_i|\boldsymbol{\eta}_i$, rather than $\boldsymbol{Y}_i|\boldsymbol{\eta}_i, \boldsymbol{u}_i$. As a result, our posterior distribution becomes

$$p(\boldsymbol{\eta}_i, \boldsymbol{\Theta}) \propto \prod_{i=1}^{N} p(T_i, \delta_i|\boldsymbol{\eta}_i, \boldsymbol{\Theta}) p(\boldsymbol{Y}_i|\boldsymbol{\eta}_i, \boldsymbol{\Theta}) p(\boldsymbol{\eta}_i|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \tag{4.2}$$

where $\boldsymbol{Y}_i|\boldsymbol{\eta}_i \sim N_{K \times n_i}\left((\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Lambda})\boldsymbol{\eta}_i, (\boldsymbol{J}_{n_i} \otimes \boldsymbol{\Sigma}_u) + (\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}_\epsilon)\right)$, $\otimes$ is a Kronecker product, $\boldsymbol{J}_{n_i}$

is a n_i-dimensional matrix of ones, and $\boldsymbol{I}_{n_i}$ is a n_i-dimensional identity matrix. This posterior distribution now only involves the covariance matrix of the random intercept, $\boldsymbol{\Sigma}_u$, rather than the random intercepts themselves.

Identifying the longitudinal submodel parameters: Because we use a dynamic factor model as our longitudinal submodel, we require additional assumptions to identify both the loadings matrix $\boldsymbol{\Lambda}$ and the structural submodel parameters $\boldsymbol{\theta}_{OU}$ and $\boldsymbol{\sigma}_{OU}$. Common approaches to identifiability of factor models include either fixing the scale of the loadings matrix or fixing the scale of the latent factors; here, we fix the scale of the latent factors by modeling them on the correlation scale. In Tran et al. (2021) [112], the authors incorporated an OU process into a dynamic factor model and, rather than directly estimating $\boldsymbol{\theta}_{OU}$ and $\boldsymbol{\sigma}_{OU}$, they estimated $\boldsymbol{\theta}_{OU}$ and the stationary correlation matrix of the OU process, $\boldsymbol{V}$. We take the same approach here and parameterize our OU process in terms of $\boldsymbol{\theta}_{OU}$ and $\boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is vector of unknown parameters corresponding to the off-diagonals of $\boldsymbol{V}$. Converting between the $(\boldsymbol{\theta}_{OU}, \boldsymbol{\sigma}_{OU})$ and $(\boldsymbol{\theta}_{OU}, \boldsymbol{\rho})$ parameterizations of the OU process is straightforward (see Section C.1.1).

Calculating the survival function: The final challenge to fitting our model involves calculating the survival function. In our likelihood, evaluating the term corresponding to the survival submodel, $p(T_i, \delta_i|\boldsymbol{\eta}_i; \boldsymbol{\Theta})$, requires integrating over the hazard function, which depends on values of the latent factors at all times from 0 to T_i. We approximate this integral using a sum across small but discrete time intervals via a midpoint rule. For example, consider a simple hazard model that depends on a constant baseline hazard and the current values of two latent factors: $h_i(t|\mathcal{H}_i(t)) = exp\{\beta_0 + \beta_1\eta_{1i}(t) + \beta_2\eta_{2i}(t)\}$. Then, we approximate the survival function as

$$p(T_i, \delta_i|\mathcal{H}_i(t)) = h_i(T_i|\mathcal{H}_i(t))^{\delta_i} exp\left\{-\int_0^{T_i} h_i(s)ds\right\} \tag{4.3}$$

$$\approx h_i(T_i|\mathcal{H}_i(t))^{\delta_i} exp\left\{-\sum_{m=1}^{M_i} \frac{1}{2}[h_i(s_{m-1}) + h_i(s_m)](s_m - s_{m-1})\right\} \tag{4.4}$$

where $s_m, m = 1, ..., M_i$ correspond to times on a fine grid of M_i points going from $s_0 = 0$ to $s_{M_i} = T_i$. Note that the integral in Equation 4.3 would be straightforward to evaluate if the latent factors were modeled using a mixed model with a linear term for time. Because we use a continuous time OU stochastic process to model the evolution of the latent factors, we must integrate over a complicated function of time and so we use this midpoint approach to approximate the integral instead. In practice, this grid is made up of both measurement

times and additional grid points. We discuss how to determine the density of this grid and how to distribute each individual's set of M_i points from 0 to T_i later in Section 4.4.1.

4.4 Simulation Study

To investigate the empirical performance of our proposed method, we assess the bias of point estimates and coverage of credible intervals via simulation. We use the design of the mHealth smoking cessation study described in Section 4.2 to inform our simulation study. We set the sample size of a single simulated dataset to $N = 200$ individuals. We assume that $K = 4$ longitudinal outcomes are measured repeatedly over time, where the maximum follow-up time is 28 days and the specific pattern of measurements varies across four difference scenarios. For each of the four measurement scenarios, we generate data under two different sets of true parameters (called setting 1 and setting 2). Setting 1 corresponds to a true OU process with higher correlation and setting 2 corresponds to a true OU process with lower correlation. All data-generating parameters are given in Section C.2.2.

All individuals are assumed to have one measurement occasion at baseline. We assume that the $K = 4$ observed longitudinal outcomes, $\boldsymbol{Y}_i$, are measurements of two latent factors, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$, where $\boldsymbol{Y}_i(t) \sim N_K\left(\boldsymbol{\Lambda}\boldsymbol{\eta}_i(t) + \boldsymbol{u}_i, \boldsymbol{\Sigma}_\epsilon\right)$. Our placement of the structural zeros within $\boldsymbol{\Lambda}$ means that $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are measurements of $\boldsymbol{\eta}_1$ and that $\boldsymbol{Y}_3$ and $\boldsymbol{Y}_4$ are measurements of $\boldsymbol{\eta}_2$. We assume that the true hazard model that underlies our observed events is $h_i(t) = \exp\left\{\beta_0 + \beta_1\eta_{1i}(t) + \beta_2\eta_{2i}(t)\right\}$.

The number of times that the longitudinal outcomes are observed varies by setting (i.e., true parameter values) and by measurement pattern, as summarized below.

Pattern 1: Measurements occur frequently and with constant probability. Individuals have an average of 19 and 25 longitudinal measurements in setting 1 and 2, respectively.

Pattern 2: Measurements occur less frequently but still with constant probability. Individuals have an average of 6 and 5 longitudinal measurements in setting 1 and 2, respectively.

Pattern 3: Measurements are distributed according to the measurement times bootstrapped from the motivating mHealth study, CARE. Individuals have an average of 34 and 38 longitudinal measurements in setting 1 and 2, respectively.

Pattern 4: Measurements are clustered together and distributed according to probabilities following a truncated cosine function of time. Individuals have an average of 30 and 21 longitudinal measurements in setting 1 and 2, respectively.

Across the 100 simulated datasets in each setting with measurement patterns 1, 2, and 4, the observed event rates are, on average, 75% in setting 1 and 71% in setting 2. Simulated datasets with measurement pattern 3 have a slightly higher observed event rate: the average in setting 1 is 85% and in setting 2 is 81%.

4.4.1 Discrete Approximation of the Survival Function

Recall that the midpoint rule for evaluating the cumulative hazard function requires defining a grid of M_i points from 0 to T_i for each individual. This grid can vary in both the density and the distribution of the points. A finer grid corresponds to a more accurate approximation of the cumulative hazard but requires increased computation time; on the other hand, a coarser grid potentially decreases the accuracy of this approximation but is less computationally intensive. In our simulation study, we consider various grid densities (i.e., grids that vary in the average gap in time between grid points). We also consider strategically placing the grid points with increased density in areas where the (estimated or true) hazard is higher. Through simulations, we find that the posterior distributions of our parameters are not sensitive to the distribution of the grid points (i.e., we placed the grid points closer together where the *true* hazard function is higher) and so we opt to take the simpler approach of placing these grid points at equally spaced intervals. In the following simulations, we vary the width of the grid between 0.2, 0.8, and 1.2 days. These grid widths are used to define the spacing of additional points that are added to the longitudinal measurement times; together, these added points and the longitudinal measurement times make up the M_i points used to approximate the survival function. When defining the grid, we require that grid points are specified at all event/censoring times but drop any other grid points that are too close to the measurement occasions, where too close is defined as within 30% of the specified grid distance (e.g., for a grid of 1.2, any added grid points within 0.36 units of time of a measurement occasion would be dropped). In our simulation study, we also consider a scenario in which we only add grid points at event/censoring times and not at intermediate time points.

4.4.2 Simulation Results

For each simulated dataset generated under setting 1 and 2 with measurement patterns 1-4, we run the HMC sampler using 1 chain for 3,000 iterations and discard the first 2,000 iterations as burn-in. The sampler allows the user to specify initial parameter estimates; we specify reasonable initial values that have the correct sign and approximately correct order of magnitude. Exact initial values, along with prior distributions, are given in Sections C.2.3 and C.2.4. To ensure that the OU process in our structural submodel is mean-reverting, we

implement the constraints on $\boldsymbol{\theta}_{OU}$ that are derived in Tran et al. (2021) [112] and summarized here in Section C.1.2. We assess convergence via trace plots and find satisfactory mixing. To summarize our point estimates, we present the distribution of the posterior medians across the 100 simulated datasets in each setting and measurement scenario in Figure 4.2, assuming a grid width of 0.8 when fitting the model. To assess the coverage of the 90% credible intervals, we summarize the average coverage rate for each parameter in Figure 4.3. As the number of added grid points increases, computation time increases substantially (see Figure C.5).

We find that our approach, which uses the discrete approximation of the survival function, recovers unbiased estimates of the parameters and returns posterior distributions of appropriate width. We also find that in our setting of ILD, the point estimates and credible intervals are not particularly sensitive to the choice of grid density (see Figure C.6 and C.8 for summaries of posterior medians and coverage rates across all grid widths). Given that grid points would not be needed if we were to fit only the longitudinal submodels (i.e., jointly fit the measurement and structural submodels), we do not expect these parameter estimates to be sensitive to the grid width. In a few instances in which the measurement occasions are irregular, however, adding grid points can help with convergence of the longitudinal submodel parameters; see Section C.3 for more discussion. For the survival submodel parameters, adding grid points at intermediate time points and not only at the event/censoring times does appear to slightly improve our estimates when the longitudinal measurement occasions are infrequent and the correlation of the true OU process decays quickly (i.e., setting 2 under measurement pattern 2), as shown in Figure 4.4.

4.5 Analysis of Smoking Cessation Data

We illustrate our method by using it to jointly model the self-reported intensity of nine different emotions recorded longitudinally and the instantaneous risk of a lapse in smoking cessation after attempted quit. We assume that the three positive emotions—enthusiastic, happy, and relaxed—are measurements of the latent psychological state of positive affect and that six negative emotions—sad, angry, anxious, restless, stressed, and bored—are measurements of the latent psychological state of negative affect. We then model time until first lapse as a function of the current values of positive and negative affect. We also adjust for two baseline covariates: pre-quit smoking history and partner status. Pre-quit smoking history is defined here as a binary variable based on the average number of cigarettes smoked per day, where more than 20 cigarettes/day corresponds to heavy smoking. We adjust for this baseline covariate because tobacco dependence is likely associated with the risk of lapse after

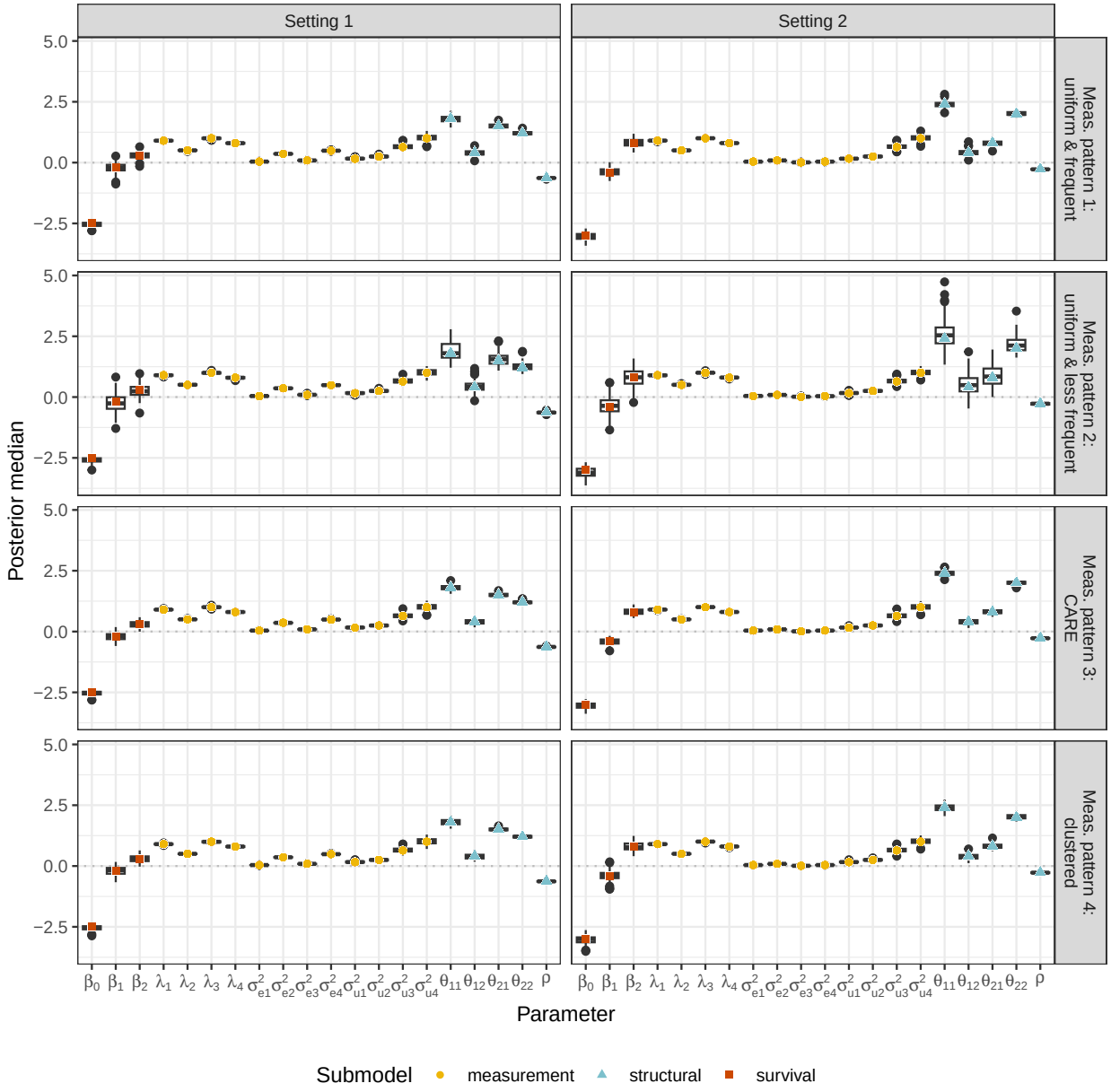


Figure 4.2: For data generated under settings 1 and 2 with each of the four measurement patterns, we use box plots to summarize the distribution of the **posterior medians for all parameters** across the 100 simulated datasets. When fitting the model, we assume a **grid width of 0.8**. True parameter values are indicated with colored dots.

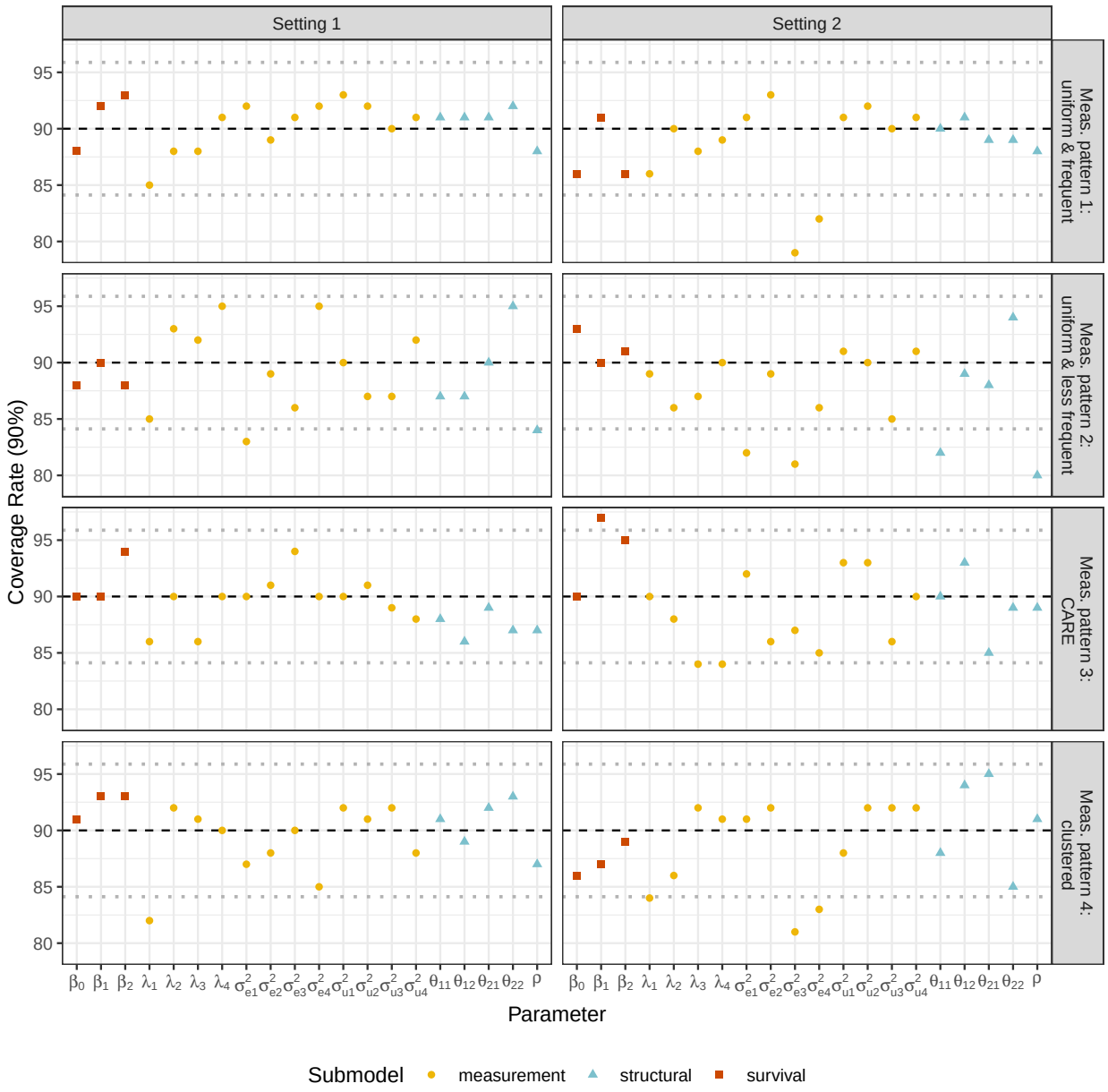


Figure 4.3: For data generated under settings 1 and 2 with each of the four measurement patterns, we summarize the **coverage rate of 90% credible intervals** across the 100 simulated datasets with the colored dots. The black horizontal dashed lines indicate target coverage and the dotted grey lines corresponds to the upper and lower bounds of a 90% binomial proportion confidence interval for a probability of 0.9.

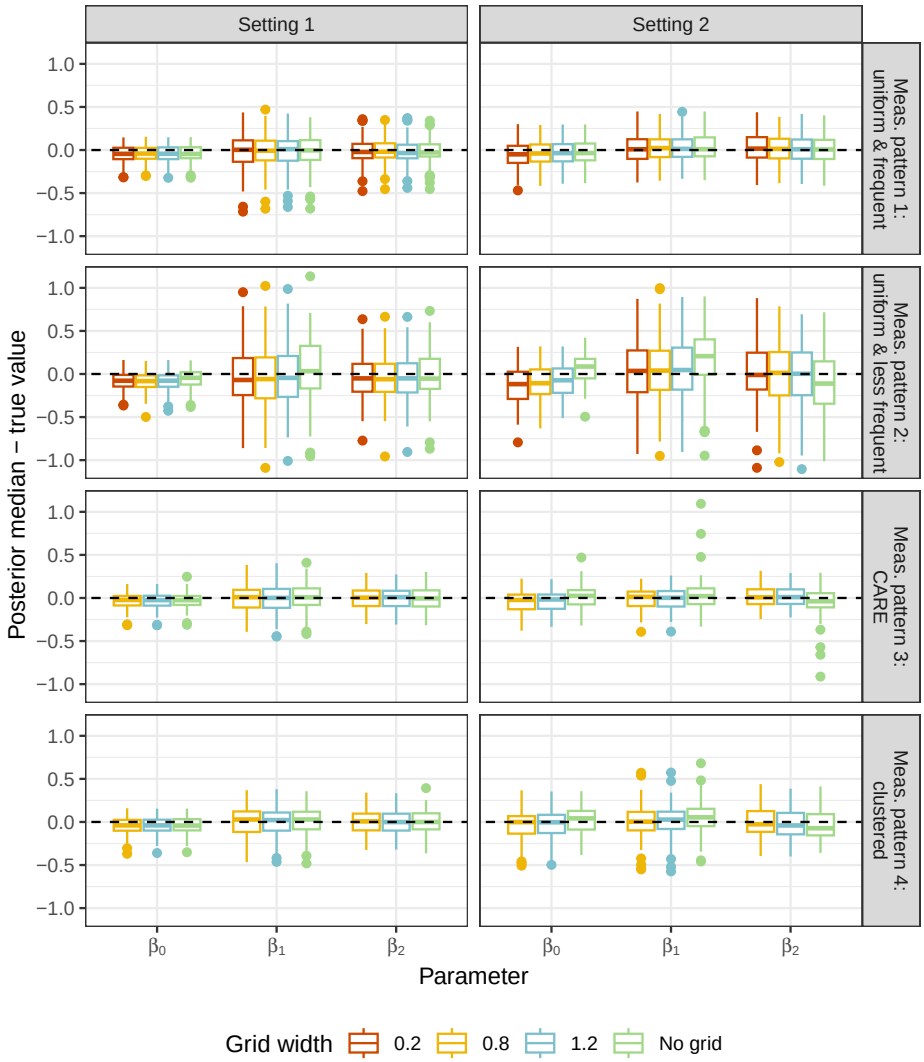


Figure 4.4: For data generated under settings 1 and 2 with each of the four measurement patterns, we use box plots to summarize the distribution of the **difference between the posterior medians and true values by grid width for the survival submodel parameters** across the 100 simulated datasets. We fit the model using a grid of 0.2, 0.8, 1.2, and no grid; we do not use the finest grid when fitting the model to data generated under measurement patterns 3 and 4 due the high computation times and lack of bias using the coarser grids.

attempted quit. Prior studies have found positive associations between partner involvement and outcomes of smoking cessation attempts [11] and so we also adjust for partner status here in our survival submodel.

Our survival submodel is $h_i(t) = h_0(t)exp\{\beta_1\eta_{1i}(t) + \beta_2\eta_{2i}(t) + \alpha_1 D_i + \alpha_2 P_i\}$. We specify a flexible piecewise constant baseline hazard for $h_0(t)$; $\eta_{1i}(t)$ and $\eta_{2i}(t)$ are the time-varying latent factors interpreted as positive affect and negative affect, respectively; D_i is the baseline measure of pre-quit smoking history (1 = 20 or more cigarettes per day, 0 = less than 20 cigarettes per day); and P_i is an indicator variable for partner status (1 = lives with a partner or spouse, 0 = everyone else). More details on the specification of this baseline hazard, along with priors, are given in Section C.4.1 and C.4.2.

We initialize parameter estimates using a two-stage approach: we first fit only the longitudinal submodel (via Stan) and use posterior samples of the latent process to fit the hazard regression model (via `flexsurv` [41]). For simplicity, we assume a constant (exponential) baseline hazard during initialization. Posterior medians—for the longitudinal submodel parameters—and maximum likelihood estimates—for survival submodel parameters—are used as initial parameter values for joint estimation. To fit the joint model, we run the HMC sampler with 4 chains for 4,000 iterations and discard the first 3,000 samples as burn-in. We assess mixing via trace plots (see Figure C.10). We also considered a survival submodel with a Weibull baseline hazard, but after comparing the goodness-of-fit of these two joint models via the distribution of predicted survival probabilities, we concluded that the piecewise constant baseline hazard better fit our data. More details on our approach to assessing goodness-of-fit are given in Section C.4.3. We present results for the joint model with the piecewise constant baseline hazard below.

In Figure 4.5, we plot posterior medians and 95% credible intervals for the parameters in each submodel. From our structural submodel, we see that our two latent factors representing positive affect (η_1) and negative affect (η_2) have a negative correlation of approximately -0.54. We also find that the posterior estimates of the parameters in the structural submodel show fairly symmetric behavior across both positive and negative affect; that is, the correlation shows similar patterns of decay as positive and negative affect are measured across increasing intervals of time. From the measurement submodel, we find that measurements of happy have the largest loading onto the latent factor representing positive affect, measurements of stressed have the largest loading onto the latent factor representing negative affect, and measurements of bored have the smallest loading onto negative affect. Finally, from the survival submodel, we find that a one-standard deviation increase in negative affect is associated with a 1.87-times increase (95% CI: 1.03-3.10) in the hazard of a lapse. Neither of our baseline covariates are significantly associated with changes in the hazard of lapse.

We can also use posterior estimates from our model to examine the trajectory of the latent factors and understand how these latent psychological states of positive and negative affect are linked with the risk of lapse after attempted quit. In Figure 4.6, we plot the posterior samples of the two latent factors for positive and negative affect, and the posterior estimates of the cumulative hazard of lapse for four study participants. For periods of follow-up during which the measurement occasions are less frequent, we see increases in the range of values covered by the 25-75% percentiles of our posterior samples, demonstrating our model's ability to capture the increased uncertainty. We also see the symmetry of our fitted structural submodel reflected in this plot: both positive and negative affect tend to vary in similar ways but in opposite directions. The estimated cumulative hazard functions for these individuals show that the instantaneous risk of lapse is highest immediately after quit time and that the cumulative hazard increases more gradually as time since quit increases.

4.6 Discussion

Motivated by ILD of self-reported emotions collected in an mHealth study of smoking cessation, we propose a joint longitudinal time-to-event model appropriate for modeling ILD. We summarize the multiple longitudinal outcomes as a smaller number of time-varying latent factors using a dynamic factor model with a structure informed by scientific context. These latent factors summarize the multiple longitudinal outcomes (e.g., emotions) and capture vulnerability to an event-time outcome (e.g., risk of lapse). This dimension-reduction approach both simplifies computation and interpretation of the factors associated with altered risk of an event. To fit our model, we use Stan [18]. We integrate out a subset of the latent parameters and leverage a discrete approximation for the survival function to make fitting this model computationally feasible. This proposed approach fills a gap in the literature as a method suitable for modeling multivariate ILD jointly with a time-to-event outcome. While we present models with only two latent factors in this paper, a different (but small) number of factors could also be considered. The choice of number of latent factors could be determined by either domain knowledge or deviance information criterion (DIC). Simulated data and code are available at `github.com/madelineabbott/OUF_JM`.

Computational cost is a major concern when jointly fitting a multivariate longitudinal-survival model. In our case, we incorporate a continuous-time stochastic process as a time-varying covariate in our hazard model, increasing the complexity of our likelihood. A Bayesian approach allows us to avoid directly evaluating the complex integrals present in our likelihood, but still requires substantial time due to the repeated sampling within the HMC algorithm. Fitting the joint model in Section 4.5 does require about 17 hours (with 4

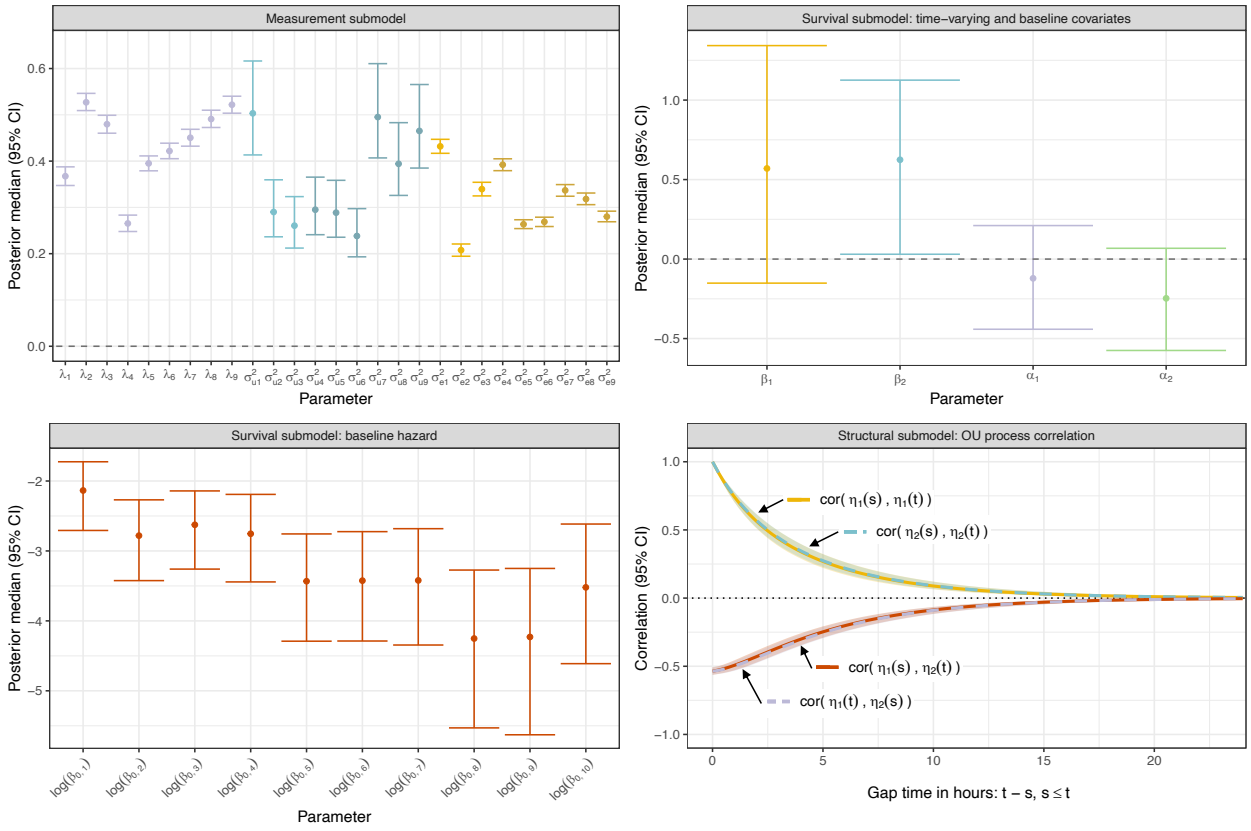


Figure 4.5: Plot of posterior medians and 95% credible intervals for parameters in the joint model with a piecewise constant baseline hazard fit to data from the mHealth smoking cessation study. Structural submodel parameters are presented via the estimated correlation decay in latent factors across increasing time intervals (see Section C.4.4 for more details on the construction of this subplot). Subscripts index the measured emotions as: 1 = enthusiastic, 2 = happy, 3 = relaxed, 4 = bored, 5 = sad, 6 = angry, 7 = anxious, 8 = restless, 9 = stressed. $\eta_1(t)$ is interpreted as positive affect and $\eta_2(t)$ is interpreted as negative affect.

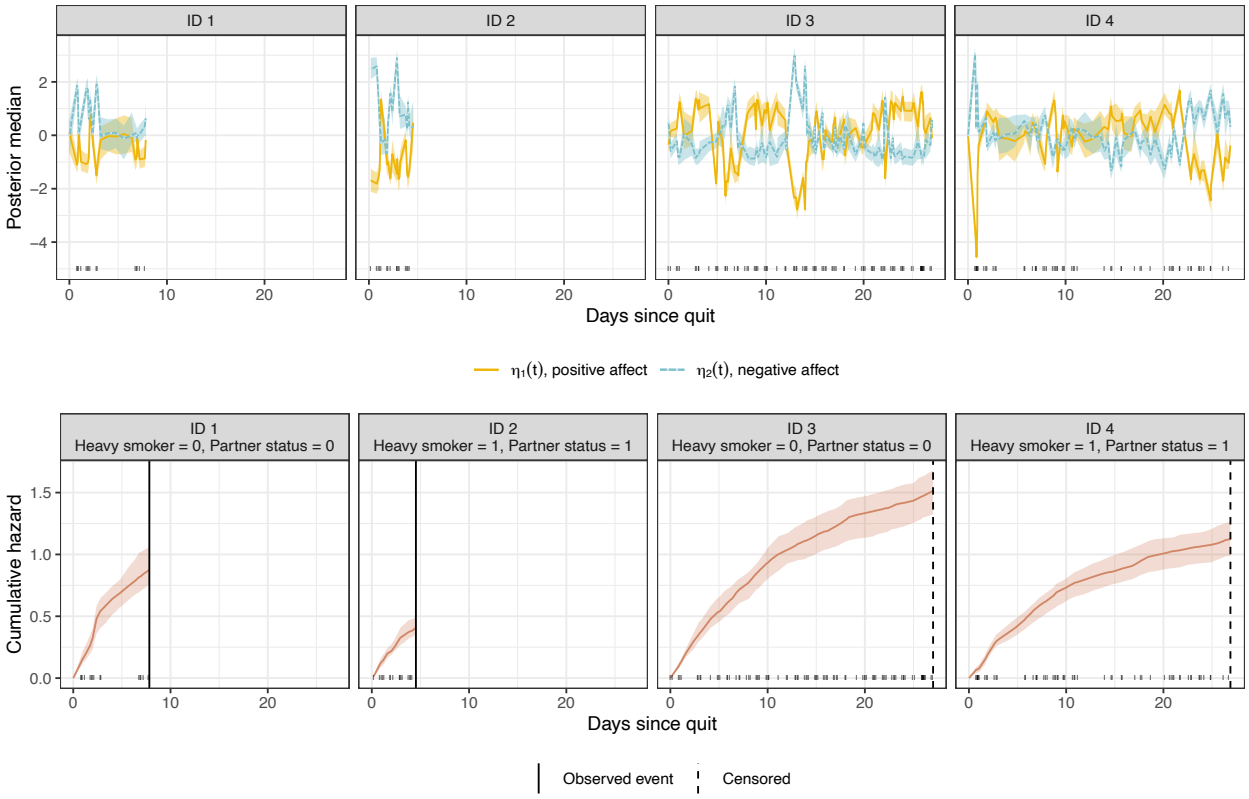


Figure 4.6: Posterior samples of the latent factors (interpreted as positive and negative affect) and cumulative hazards for four individuals in the mHealth smoking cessation study. Posterior estimates of the latent factors and cumulative hazards are summarized using posterior medians; shaded bands denote the range of values between the 25th and 75th percentiles. The black tick marks along the x-axis of the plots correspond to the longitudinal measurement occasions observed in the data; in the cumulative hazard plots, the vertical lines indicate the timing of observed events or censoring.

chains run in parallel on 4 cores) and so further investigation of alternative computational strategies that increase the speed of modeling fitting is an important area of future research. For example, the strategies used in Murray and Philipson (2022) [62] and Rustand et al. (2023) [98] could potentially be adapted to work in our setting.

In our approach, we use a discrete approximation of the cumulative hazard function. We show via simulation that simply assuming a somewhat sparse and uniform grid for this approximation works well in the setting of ILD. Furthermore, when ILD is used, our ability to recover good point estimates is not sensitive to the width of the grid. Our specific setting is one in which the ILD captures rapidly varying outcomes and so measurement occasions must occur frequently so that large abrupt fluctuations are not missed. Since measurement occasions are close together, the placement and number of additional grid points are not as important. In other settings in which the longitudinal outcomes change more smoothly, less frequent measurement occasions could still capture important changes over time. When measurement occasions are farther apart, sensitivity to the choice of grid may increase. But, if the longitudinal outcome changes more slowly, linearly interpolating the intermediate values of the longitudinal outcome within the discrete approximation of the hazard function might not be so problematic. Overall, we find that in our ILD setting, the grid width is not particularly important. But one could imagine an alternative scenario where the placement of grid points might matter more. In this scenario, further investigation of the location and number of grid points may be warranted and methods, such as that described in Fernández et al. (2016) [27], could be adapted to place grid points according to the intensity of the hazard function. We leave investigation of this alternative scenario as future work.

A weakness of our approach is our assumption of non-informative measurement occasions. Although this assumption is commonly made in joint longitudinal-survival models, self-reported longitudinal outcomes are likely susceptible to informative missingness. Responses to EMA questionnaires may be missing for many reasons, ranging from non-response due to poor mood or lack of cell phone reception. Important future work could include incorporating an additional submodel that accounts for important patterns in the timing of the measurement occasions.

Finally, in our motivating mHealth data, individuals generally experience repeated smoking lapses after attempted quit. We modeled the time until first lapse after attempted quit, but our model could be adapted for the recurrent event of repeated lapses. Additionally, current smokers who attempt to quit often progress through phases in which they are actively attempting to quit before potentially relapsing back into their prior smoking habits. Multi-state models have previously been proposed to model transitions in long-term smoking habits; e.g., [12]. Our joint model could potentially be extended to model short-term changes

in cigarette use. Finally, temporal trends could also be incorporated into the longitudinal submodel in order to account for systematic changes over time in the measured longitudinal outcomes, which may vary according to the current state of smoking.

CHAPTER 5

Estimation of Time-Varying Treatment Effects in a Joint Model for Longitudinal and Recurrent Event Outcomes in Mobile Health Data

5.1 Introduction

As described in earlier chapters, mHealth technology enables frequent measurements of multiple outcomes over time. The resulting ILD potentially contain information on underlying behavioral, psychological, or other health-related states that are indirectly measured through this larger number of observed outcomes. Not only does mHealth technology allow researchers to collect rich data on study participants, but it also facilitates the delivery of repeated low-cost treatments directly to individuals. Often, these treatments take the form of JITAIs [69]. This type of intervention is generally delivered multiple times per day and tailored—in timing, type, and intensity—to each individual's context and state. JITAIs can be developed or optimized using MRTs in which individuals are randomized—possibly multiple times per day—to either be sent or not be sent a treatment.

MRTs have gained popularity over the past few years, after their proposal in 2015 [45, 51]. This type of trial has been used to inform the development of effective JITAIs in a variety of different settings, which range from promoting physical activity among sedentary adults [46] to improving data collection in the context of substance use [82]. In an MRT, researchers are often primarily interested in understanding the effect of the treatment on an outcome measured shortly after randomization. For example, researchers might be interested in how sending notifications affect substance use within the next 12 hours.

The treatment effect of interest—specifically, of the repeatedly sent treatment on the repeatedly measured outcome—may account for the time-varying nature of the treatment,

possibly other covariates, and a repeatedly-measured outcome. Boruvka et al. (2018) [9] propose an estimator of treatment effect, specifically a causal excursion effect, that accounts for the time-varying nature of the treatment, the outcome, and other time-varying potential moderators. This estimator, which can be conditional on all past treatments or a different set of potential moderators (e.g., past engagement with notifications sent to individuals' smartphones), is designed to capture the effect of the treatment on a future measurement of the outcome under different treatment scenarios. This treatment effect is marginal across all data except the set of potential moderators on which the effect is conditioned. To estimate these MRT treatment effects, approaches based on generalized estimating equations are often used; namely, weighted and centered least-squares [9]. Various extensions of this estimating-equation based method have also been proposed; e.g., [80] and [101].

Rather than using an estimating equation-based approach, we propose a method for obtaining model-based estimates of treatment effect. Model-based estimates of treatment effect have been previously used in the joint longitudinal-survival model setting in which understanding the hazard of an event as a function of a time-varying predictor is of interest. When a longitudinal and survival outcome are modeled jointly, rather than separately, the risk of bias in estimates of treatment effect decreases and statistical efficiency can increase, as this joint approach considers dependencies between the longitudinal and survival processes [40]. Joint models often assume that the longitudinal outcome is an intermediate variable, meaning that the treatment impacts the risk of an event by altering the longitudinal process. Much work in this area has been motivated specifically by observational data (e.g., [129, 111, 92]). In observational data, the decision to provide treatment often depends on the value of the longitudinal outcome. When both the risk of an event and the decision to provide treatment depend on the value of the longitudinal outcome, the longitudinal process is a so-called time-dependent confounder and extra considerations must be made (for more details, see [111] and [92]). In our MRT setting, however, the decision to deliver treatment is random, and so we avoid the situation in which the longitudinal outcome is also a time-dependent confounder. Instead, we must consider the fact that treatment can be delivered to participants multiple times per day; this pattern of treatment is standard among other MRTs.

A primary advantage of using this model-based approach to estimating treatment effects is that we can use it to disentangle the effect of treatment on both a longitudinal latent process measured through multiple longitudinal outcomes and on the hazard of recurrent events. Our work is motivated by an MRT that aimed to use behavioral strategies to promote smoking cessation among current smokers attempting to quit. The design of this study is identical to that described in Nahum-Shani et al. (2021) [68]. Over the course of the study, reminders

to engage in certain behavioral strategies (i.e., treatments) are sent multiple times per day to participants' phones via app-based notifications. Multiple times per day, information on participants' emotions (i.e., ILD) and repeated instances of substance use (i.e., recurrent events) is also recorded. We propose incorporating these repeated treatments into a joint longitudinal-recurrent event model that links latent longitudinal psychological states with the hazard of recurrent events of substance use. While this approach has similarities to mediation analysis, we view our contribution as useful for exploratory analyses of associations of interest. Developing a framework to simultaneously model both the time-varying effect of the adaptive interventions on the risk of recurrent events and the underlying latent process related to these events has the potential to provide useful insights that can help scientists better understand how JITAIs may be impacting health events of interest and thus inform the design of improved JITAIs.

The main contribution of this work is a model-based approach for estimating treatment effects—on both the risk of recurrent events and the trajectory of the latent factors—in an MRT. We consider two different mechanisms by which the treatments potentially impact the longitudinal latent process, which we model using a continuous-time multivariate stochastic process. Specifically, we allow treatment to impact the latent process through an additive shift to its mean, or we allow treatment to impact the underlying dynamics of the latent process by altering the rate at which it reverts towards the mean (i.e., as time-varying drift). Useful consequences of this model-based approach are threefold:

1. We can disentangle the impact of treatment directly on the event outcome and through the latent factors.

2. We can use the fitted model to predict the risk of experiencing an event within a fixed interval of time, given the trajectory of the latent factors and different treatment patterns.

3. We can use the fitted model to inform the development of improved JITAIs or other treatments through increased understanding of the pathways by which treatment may impact the recurrent event outcome.

When analyzing the motivating MRT data, we also demonstrate how information criteria can be used to compare the fit of models that make different assumptions about the relationship between the recurrent events, longitudinal latent process, and repeated treatment effects.

The remainder of this paper is organized as follows: in Section 5.2, we describe the MRT motivating this work; in Section 5.3, we present the joint longitudinal-recurrent event models that incorporate time-varying treatment effects; in Section 5.4, we demonstrate the

statistical properties of our method via simulation; in Section 5.5, we analyze MRT data in a case study; and in Section 5.6, we provide a discussion.

5.2 Motivating Data

This work is motivated by data from the Affective Science MRT. This study is still ongoing and so we use data available at the time of drafting this manuscript from currently enrolled participants. The design of this study is identical to the one described in Nahum-Shani et al. (2021) [68] and uses EMAs to collect self-reported information on the intensity of a variety of different positive and negative emotions, along with recent substance use. Study participants are randomized to be sent interventions delivered via prompts to their smartphones up to six times per day. These prompts are aimed at improving their engagement in behavioral and self-regulatory activities known to decrease vulnerability to substance use. These prompts, which are sent randomly with a probability of 1/2, are sent approximately 1 hour before the EMAs are sent. A simplified diagram of the MRT design is provided in Figure 5.1. More details on the design can be found in Nahum-Shani et al. (2021) [68].

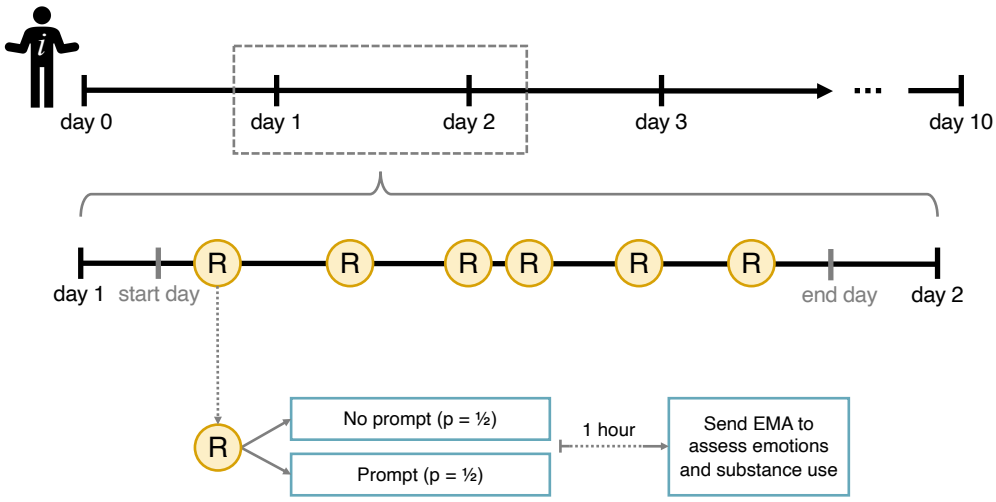


Figure 5.1: Simplified diagram of the micro-randomized trial design. During each day, a participant is randomized six times to potentially receive treatment, which comes in the form of an app-based prompt. Approximately one hour after randomization, participants are prompted to respond to an EMA that includes questions that assess the current intensity of multiple emotions and the approximate time of substance use since the prior EMA, among other things. This figure is adapted from Nahum-Shani et al. (2021) [68].

Our longitudinal outcome of interest is a set of 15 different emotions—5 positive and 10 negative—that are measured via self-report on a 5-point Likert scale at each EMA. The

specific emotions are: happy, proud, relaxed, grateful, enthusiastic, angry, ashamed, irritable, guilty, lonely, anxious, sad, restless, bored, and hopeless. On average, emotions are measured 28 times per person (min. = 3, max. = 53).

Our recurrent event of interest is poly-substance use, which is also measured using self-report in the EMAs. Poly-substance use is defined as either using marijuana, vaping, or smoking cigarettes. At each EMA, individuals are asked to report substance use since the prior EMA; they are also asked to respond to additional questions related to the time of use. Defining an outcome appropriate for modeling when data are collected from multiple questions asked in EMAs is a complex challenge often encountered when analyzing data from mHealth studies, as discussed in Potter et al. (2023) [74]. We defined a set of deterministic rules that we use to consolidate information across the substance use-related questions to approximate the time of recurrent events of poly-substance use. Our data wrangling approach is described in more detail in Section D.1.1.

The within-individual average EMA completion rate is 47%, but this completion rate varies widely across individuals (range: 5% - 88%). Although many individuals respond frequently to the EMAs, some individuals go for multiple days without responding to an EMA. The event submodel assumes that individuals are always at risk of experiencing an event (up until censoring) and so to help reduce the impact of non-response on our event submodel, we censor individuals at the time of their most recent longitudinal measurement if they fail to respond to an EMA for a period of more than 48 hours. Otherwise, we censor individuals at the time of their final completed EMA. Additionally, we exclude any individuals who fail to respond to an EMA within the first 48 hours of the study, resulting in an analytic sample size of $N = 64$ individuals.

The observed data, which consist of the longitudinal outcomes, recurrent events, and treatment timings, are plotted in Figure 5.2 for a subset of individuals in the motivating MRT. Across all individuals, we observe 1,162 events of poly-substance use, which corresponds to an average of 19 events per person spread over a maximum of 10 days. Additional plots detailing the timing of events are given in Section D.1.1.

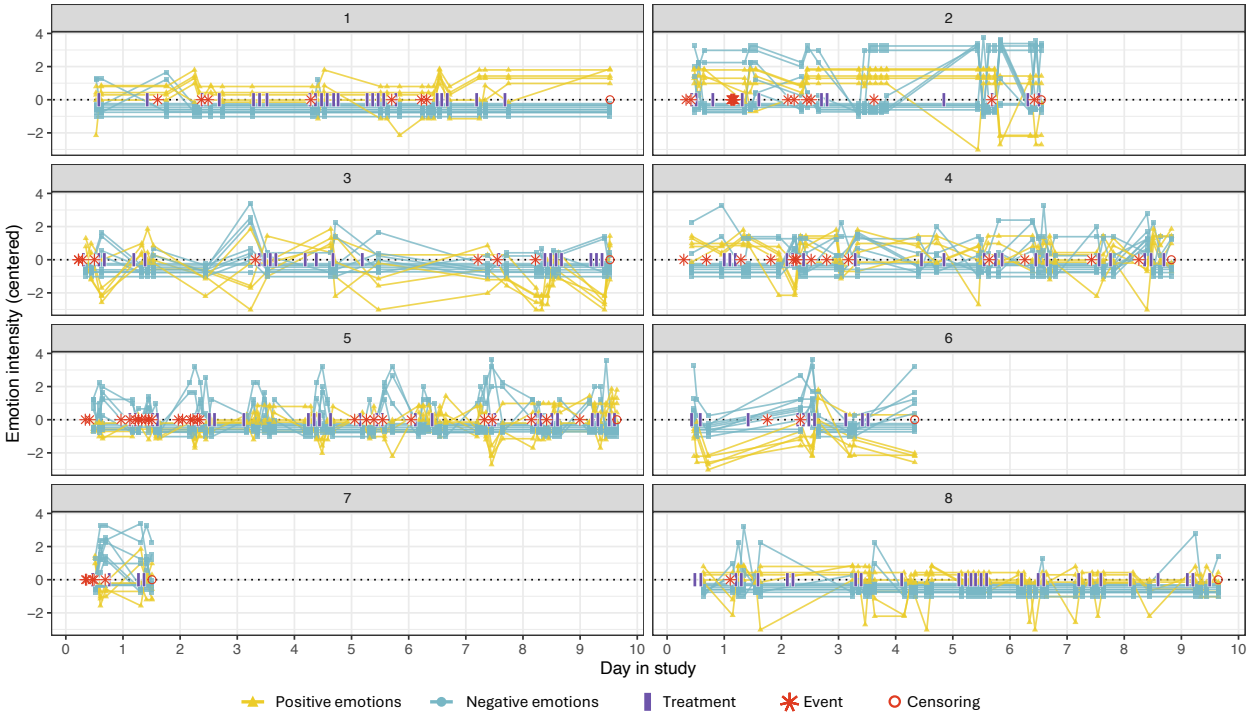


Figure 5.2: Responses to emotion-related questions, timing of recurrent poly-substance use events, and timing of treatments for eight different participants in the motivating MRT.

5.3 Methods

To model the relationship between repeated treatments, a multivariate longitudinal latent process, and the risk of recurrent events, we use a joint longitudinal-recurrent event model. The model extends the joint longitudinal-survival model previously presented in Chapter 4 to allow for recurrent events and to include a model for the effect of treatment on the latent process and the hazard of events.

We first introduce our joint longitudinal-recurrent event model in a setting without treatment effect models. Let $i = 1, ..., N$ index the independent individuals in our dataset, and let $j = 1, ..., n_i$ index the longitudinal measurement occasions for individual i. This joint model consists of the following three submodels:

Measurement submodel: The measured k longitudinal outcomes, represented by k-length vector $\boldsymbol{Y}_i(t_{ij})$, are assumed to be noisy observations of a p-dimensional latent process $\boldsymbol{\eta}_i(t_{ij})$, where $p < k$. We model the measured longitudinal outcomes using a dynamic factor model,

$$\boldsymbol{Y}_i(t_{ij}) = \boldsymbol{\Lambda}\boldsymbol{\eta}_i(t_{ij}) + \boldsymbol{u}_i + \boldsymbol{\epsilon}_i(t_{ij}). \tag{5.1}$$

$\boldsymbol{\Lambda}$ is a $k \times p$ time-invariant loadings matrix that captures the association between the measured longitudinal outcomes and the time-varying latent process. $\boldsymbol{u}_i$ is a k-length vector of outcome-specific random intercepts that account for baseline differences in the measured longitudinal outcomes across individuals; we assume $\boldsymbol{u}_i \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ where $\boldsymbol{\Sigma}_u$ is a diagonal matrix. $\boldsymbol{\epsilon}_i(t_{ij})$ captures measurement error; we assume $\boldsymbol{\epsilon}_i(t_{ij}) \sim N_k(0, \boldsymbol{\Sigma}_\epsilon)$ and that $\boldsymbol{\Sigma}_\epsilon$ is diagonal. In the motivating data, $\boldsymbol{Y}_i(t_{ij})$ is the set of 15 emotions measured at time t_{ij} for individual i.

Structural submodel: The p-dimensional latent process $\boldsymbol{\eta}_i(t)$ is assumed to be a multivariate OU stochastic process. This process can be thought of as a continuous-time version of a VAR process and captures correlated change within and between multiple latent factors over time. The stochastic differential equation definition of the OU process is

$$d\boldsymbol{\eta}_i(t) = [\boldsymbol{\mu} - \boldsymbol{\theta}\boldsymbol{\eta}_i(t)]\, dt + \boldsymbol{\sigma} dW_i(t). \tag{5.2}$$

$\boldsymbol{\theta}$ is a $p \times p$ parameter matrix that captures the speed at which the latent process reverts towards a constant mean $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ is a $p \times p$ parameter matrix that captures the volatility of the process. $W_i(t)$ is a Wiener process. Often, the constant mean is assumed to be 0 ($\boldsymbol{\mu} = 0$). Later, when we introduce interventions, we allow $\boldsymbol{\mu}$ to depend on time. For more discussion of the OU process, see Brigo and Mercurio (2007)[10]. In the motivating data, we summarize the 15 measured emotions as a two time-varying latent factors using a bivariate ($p = 2$) OU process. These latent factors can be interpreted as capturing different aspects of psychological state, namely positive affect and negative affect.

Event-time submodel: The hazard of the r^{th} event for individual i can depend on the history of the latent process $\mathcal{H}_i(t) = \{\boldsymbol{\eta}_i(s), 0 \leq s \leq t\}$, the history of recurrent events $\mathcal{R}_i(t) = \{T_{ir} < t, r = 1, 2, ...\}$, and possibly baseline covariates $\boldsymbol{X}_i$. Very generally, we define the hazard for recurrent event r as

$$h_{ir}\{t|\mathcal{H}_i(t), \mathcal{R}_i(t), X_i\} = h_0(t) \exp\big\{f(\mathcal{H}_i(t); \boldsymbol{\beta}_H) + g(\mathcal{R}_i(t); \boldsymbol{\beta}_R) + \boldsymbol{X}_i^\top \boldsymbol{\gamma}\big\}. \tag{5.3}$$

$h_0(t)$ is the baseline hazard, which may be time-varying. We take a clock-reset approach in which we use the hazard to model the time between each recurrent event and thus reset the clock for the baseline hazard (but not the time-dependent predictors) to 0 after the occurrence of an event. We then account for associations between repeated events through $g(\mathcal{R}_i(t); \boldsymbol{\beta}_R)$; for example, $g(\mathcal{R}_i(t); \boldsymbol{\beta}_R)$ could be some transformation

of the time since the most recent prior event, $T_{i,r-1}$. $\boldsymbol{X}_i$, which is a vector containing baseline covariates, can help capture and adjust for pre-quit substance use or other time-independent risk factors. We discuss more concrete forms of f and g later in the simulation study.

Due to the complexity of our model and the MRT data, we could potentially adapt our joint model to capture the effect of these randomized treatments in up to three different ways. Each treatment could be associated with changes in the measured longitudinal outcome, changes in the latent process, or modifications directly to the risk of a recurrent event. We summarize these possible pathways for treatment effect in Figure D.10. In the remainder of this paper, we focus on the scenarios in which treatment directly impacts (a) the latent process and (b) the risk of a recurrent event. We could potentially also allow the treatment to directly impact the measured longitudinal outcome, but this modeling decision would imply that the treatment changes the observed longitudinal outcome without modifying the latent process that the longitudinal outcome is assumed to measure. This mechanism for treatment effect is less scientifically relevant than (a) or (b) because we view the observed longitudinal outcomes simply as noisy measurements of the latent factors of interest. Our measurement error perspective implies that any effect of treatment on the measured outcomes should result from changes to the latent factors themselves. Thus, our main contribution is the development of treatment effect models for (a) and (b).

5.3.1 Modeling the Impact of Treatment on the Latent Process

We now build on the models defined in the previous section to model the effect of treatment interventions. Let $a_i(t_{ij})$ denote the decision to treat individual i at time t_{ij} and $\mathcal{A}_i(t) := \{a_i(s), 0 \leq s \leq t\}$ be their treatment history.

We can model the impact of treatment directly on the latent process by incorporating a model for the treatment effect into our structural submodel. The prompts in the Affective Science MRT target engagement in behavioral activities known to decrease vulnerability to smoking and so, through this model formulation, we assume that these vulnerability-related behavioral states are represented by the low-dimensional latent process. This assumption about the treatment effect implies that treatment could also impact the risk of a recurrent event and the measured longitudinal outcomes via changes in the trajectory of the latent process.

We consider two different approaches to modeling the impact of treatment on the trajectory of the latent process, which are based on different assumptions about how the treatment might affect the latent process. In this first approach, we assume that the impact of treat-

ment is additive and directly shifts the OU process away from the constant mean for a short window of time after the treatment. Returning to the setting of the motivating MRT, this approach would assume that sending a prompt to an individual results in a shift in the level of their behavioral state (e.g., positive and negative affect) on the scale of the state itself. Let $\boldsymbol{\eta}^*(t)$ denote the OU process without treatment effect (i.e., Equation 5.2 with $\boldsymbol{\mu} = 0$). Then, the impact of treatment is modeled as

$$\boldsymbol{\eta}_i(t) = \boldsymbol{\eta}_i^*(t) + \boldsymbol{\mu}_i(t) \tag{5.4}$$

where $\boldsymbol{\mu}_i(t)$ is a simple function that describes the short-term impact of treatment. Here, we assume that, for each individual i, $\boldsymbol{\mu}_i(t)$ is a simple deterministic function:

$$\boldsymbol{\mu}_i(t) = \sum_{t_{ia} \in \mathcal{A}_i(t)} \boldsymbol{\tau} \left(1 - \frac{t - t_{ia}}{\delta_a}\right)_+ \tag{5.5}$$

where $\mathcal{A}_i(t)$ contains the time t_{ia} of all treatments sent to individual i prior to time t, $\boldsymbol{\tau}$ is a length-p vector that captures the maximum impact of the treatment at the time at which it is delivered, and δ_a defines the window over which each treatment is active. We assume that δ_a is known, but that $\boldsymbol{\tau}$ is estimated. Note that the deterministic function $\boldsymbol{\mu}_i(t)$ implicitly conditions on the treatment history $\mathcal{A}_i(t)$. This form for $\boldsymbol{\mu}_i(t)$ assumes that the effect of treatment on the latent process is largest at the time of treatment; then, this effect decays linearly until no effect remains δ_a units of time after the treatment was delivered. If multiple treatments are delivered in rapid succession, then the cumulative impact of these treatments is additive. This model for treatment is illustrated in Figure 5.3a. Due to the complexity of our joint model, we assume a fairly simple form for $\boldsymbol{\mu}_i(t)$ here, but an analyst could specify a different form for $\boldsymbol{\mu}_i(t)$ better suited to their specific setting.

Assuming this additive treatment effect in Equation 5.4, we can write the conditional distribution for our structural submodel as follows: if $\boldsymbol{\eta}_i(0) \sim N_p(\mathbf{0}, \boldsymbol{V})$ where $\boldsymbol{V} = vec^{-1}\left\{(\boldsymbol{\theta} \oplus \boldsymbol{\theta})^{-1} vec(\boldsymbol{\sigma}\boldsymbol{\sigma}^\top)\right\}$, then for times t and s, $t > s$,

$$\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s) \sim N_p\left(\boldsymbol{\mu}_i(t) + e^{-\boldsymbol{\theta}(t-s)}\left(\boldsymbol{\eta}_i(s) - \boldsymbol{\mu}_i(s)\right), \boldsymbol{V} - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t-s)}\right) \tag{5.6}$$

where e is the matrix exponential.

As an alternative to modeling the treatment effect as an additive shift to the mean of the OU process, we can instead model treatment as impacting the dynamics of the latent process through a time-varying drift term on the derivative scale. In the setting of the motivating MRT, this approach assumes that sending a prompt to an individual alters the rate at which

their behavioral states (e.g., positive and negative affect) change over time. For example, sending a prompt could increase the rate at which levels of negative affect revert towards their average level. The standard OU process assumes a constant value for $\boldsymbol{\mu}$, but the Hull-White process, which is often used in financial math applications, extends the OU process to allow for time-varying drift. For more discussion of the OU process and Hull-White model, see Brigo and Mercurio (2007) [10]. The SDE for the Hull-White model is

$$d\boldsymbol{\eta}_i(t) = [\boldsymbol{\mu}_i(t) - \boldsymbol{\theta}\boldsymbol{\eta}_i(t)]\, dt + \boldsymbol{\sigma} dW_i(t) \tag{5.7}$$

where $\boldsymbol{\mu}_i(t)$ is still a simple function that describes the short-term treatment effect. From this SDE, it follows that the conditional distribution of $\boldsymbol{\eta}_i(t)$ with time-dependent drift is

$$\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s) \sim N_p\left(e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) + \int_s^t e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\mu}_i(u)du, \boldsymbol{V} - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t-s)}\right) \tag{5.8}$$

The mean of this distribution requires integrating across $e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\mu}_i(u)$ as a function of u. Depending on the specific formulation of $\boldsymbol{\mu}_i(t)$, an analytic solution to this integral may or may not exist. If we assume that the treatment effect model takes the linear form given in Equation 5.5, we can derive the analytic solution to the integral in the conditional mean in Equation 5.8. In a setting in which only a single treatment impacts the drift of this latent process, integration would be straight forward; in our setting, however, we must carefully account for overlapping active treatments. We use $\mathcal{A}_i(s-\delta_a, t)$ to denote the set of times at which treatments were sent to individual i between time $s-\delta_a$ and time t; this set of treatment times corresponds to all treatments that are active between times s and t. If we solve the integral in Equation 5.8, then we can re-write the distribution in an analytic form:

$$\begin{aligned}\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s) \sim N_p\Bigg(& e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) \\ &+ \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \left[\left(1 - \frac{u-t_{ia}}{\delta_a}\right)e^{-\boldsymbol{\theta}(t-u)}\theta^{-1} + \frac{1}{\delta_a}e^{-\boldsymbol{\theta}(t-u)}\right]\boldsymbol{\tau}\Bigg|_{u=max(t_{ia},s)}^{u=min(t,t_{ia}+\delta_a)}, \\ &\boldsymbol{V} - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t-s)}\Bigg)\end{aligned} \tag{5.9}$$

More details on this derivation are given in the Appendix (Section D.3).

If we compare the conditional distribution of $\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s)$ when the treatment effect is modeled as an additive term (Eq. 5.6) to the conditional distribution when treatment effect

is modeled as impacting the dynamics of the latent process through the drift (Eq. 5.9), we see that the variance terms are the same but that the treatment function $\boldsymbol{\mu}_i(t)$ shows up differently in the conditional means:

Additive treatment effect:

$$\mathbb{E}\left[\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s)\right] = e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) + \boldsymbol{\mu}_i(t) - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\mu}_i(s)$$

Drift treatment effect:

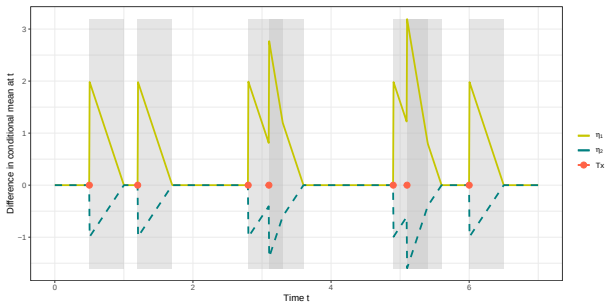
$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s)\right] =& e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) \\ &+ \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \left[\left(1 - \frac{u-t_{ia}}{\delta_a}\right)e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\theta}^{-1} + \frac{1}{\delta_a}e^{-\boldsymbol{\theta}(t-u)}\right]\boldsymbol{\tau}\Bigg|_{u=max(t_{ia},s)}^{u=min(t,t_{ia}+\delta_a)}\end{aligned}$$

In the additive version, the mean reversion parameter $\boldsymbol{\theta}$ and treatment effect function $\boldsymbol{\mu}_i(t)$ show up clearly in the terms in the sum. Using the definition of the latent process under additive treatment effect as given in Equation 5.4, we can re-write the conditional mean as $\mathbb{E}\left[\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s)\right] = e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i^*(s) + \boldsymbol{\mu}(t)$. Writing the conditional mean in this format makes the additive impact of treatment quite clear. When treatment is incorporated into the drift of the latent process, the impact of treatment on the conditional expectation is still additive; however, the impacts of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_i(t)$ on the trajectory of the latent process are linked together in a much more complicated way (see Figure 5.3b).

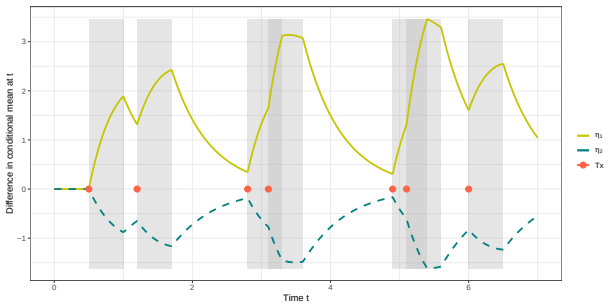
In some settings, it might also be reasonable to assume that the occurrence of events alters the latent process. The impact of events on the latent process could be modeled as a drift or additive term, similar to how we modeled the impact of treatment on the latent process.

5.3.2 Modeling the Impact of Treatment on the Hazard of Recurrent Events

In the previous section, we described how treatment can modify the trajectory of the latent process. In this section, we discuss modeling the direct impact of treatment on the hazard of an event. Modeling the impact of treatment through a term included in the hazard model implies that the treatment alters the risk of a recurrent event through some mechanism that is not captured by the latent process. If the app-based notifications in the motivating study target engagement in certain behaviors, then, by using this model formulation, we assume that these certain behavioral states are different from those captured by the latent



(a) The difference in means at time t between a latent process with treatment effect modeled as an **additive** shift to the mean and a latent process without a treatment effect is $\boldsymbol{\mu}_i(t) - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\mu}_i(s)$, where $s = 0$.



(b) The difference in means at time t between a latent process with treatment effect modeled on the derivative scale as time-varying **drift** and a latent process without a treatment effect is $\int_s^t e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\mu}_i(u)du$, where $s = 0$.

Figure 5.3: Treatment models for the longitudinal process. The plots show the difference between the latent process with a treatment effect model—either additive or drift—and a latent process with a constant mean of 0. We plot this difference across times t, with s fixed at 0, if treatments were to be sent at various times between 0 and t, as indicated by the red dots. The shaded grey bars highlight the (potentially overlapping) regions over which treatments may have some non-zero effect.

factors. To model this treatment effect, we can include a treatment-related term $\tilde{\mu}_i(t)$ in the event-time submodel for individual i's r^{th} recurrent event:

$$h_{ir}\{t|\mathcal{H}_i(t), \mathcal{R}_i(t), X_i\} = h_0(t)\exp\left\{f(\mathcal{H}_i(t); \boldsymbol{\beta}_H) + g(\mathcal{R}_i(t); \boldsymbol{\beta}_R) + \tilde{\mu}_i(t) + \boldsymbol{X}_i^\top \boldsymbol{\gamma}\right\} \quad (5.10)$$

For simplicity, we assume that $\tilde{\mu}_i(t)$ has the same form as the treatment model for the latent process (Equation 5.5). In a different setting, the analyst may want to specify an alternative form for $\tilde{\mu}_i(t)$, but here we use:

$$\tilde{\mu}_i(t) = \sum_{t_{ia} \in \mathcal{A}_i(t)} \tilde{\tau}\left(1 - \frac{t - t_{ia}}{\delta_b}\right)_+ \quad (5.11)$$

where the impact of treatment on the hazard is captured through the (scalar) parameter $\tilde{\tau}$, which is to be estimated. As before, we assume that the duration of the treatment effect δ_b is known.

5.3.3 Inference

We take a Bayesian approach to fitting our joint longitudinal-recurrent event model. Combining the measurement submodel (Equation 5.1), the hazard submodel (Equation 5.10),

and the structural submodel with the treatment effect modeled either as an additive shift (Equation 5.6) or as drift (Equation 5.8), our likelihood is

$$
\begin{aligned}
p(\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{Y}; \boldsymbol{\Theta}_R, \boldsymbol{\Theta}_M, \boldsymbol{\Theta}_S) = \\
\prod_{i=1}^{N} \prod_{r=1}^{R_i} \int \Bigg[& p(T_{ir}, \delta_{ir} | \mathcal{H}_i(T_{ir}), \mathcal{R}_i(T_{i,r-1}), \mathcal{A}_i(T_{ir}); \boldsymbol{\Theta}_R) \\
& \times \prod_{j=1}^{n_i} p(Y_i(t_{ij}) | \mathcal{H}(T_{ir})_i, \mathcal{A}(T_{ir})_i; \boldsymbol{\Theta}_M) p(\boldsymbol{\eta}_i; \boldsymbol{\Theta}_S) d\boldsymbol{\eta}_i \Bigg]
\end{aligned}
\tag{5.12}
$$

where R_i is the number of events for individual i, $\boldsymbol{\Theta}_R = (\boldsymbol{\beta}_H, \boldsymbol{\beta}_R, \tilde{\tau})$, $\boldsymbol{\Theta}_M = (\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon)$, and $\boldsymbol{\Theta_S} = (\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\tau})$. Recall that we have defined $\mathcal{H}_i(t)$ as this history of the latent process until time t, $\mathcal{R}_i(t)$ as the history of recurrent events until time t, and $\mathcal{A}_i(t)$ as the treatment history until time t. The likelihood does depend on the treatment history, but we can factor out this term after conditioning on the observed treatment history and so we omit it from our likelihood definition above. We do not include baseline covariates in this definition of the likelihood, but incorporating them is straightforward.

The likelihood contribution of the recurrent events is given by

$$
p(T_{ir}, \delta_{ir} | \mathcal{H}_i(T_{ir}), \mathcal{R}_i(T_{i,r-1}), \mathcal{A}_i(T_{ir}); \boldsymbol{\Theta}_R) = h_{ir}(T_{ir} | \mathcal{H}_i(T_{ir}))^{\delta_{ir}} exp \left\{ - \int_{T_{i,r-1}}^{T_{ir}} h_{ir}(s) ds \right\}.
$$

Note that in this definition of the joint likelihood, we opt to write the distribution of the observed longitudinal outcome conditional only on the latent factors and integrate out the random intercepts in order to decrease the number of unknown parameters that must be sampled within the Bayesian algorithm. This distribution for the measured longitudinal outcome is

$$
\boldsymbol{Y}_i(t_{ij}) | \mathcal{H}_i(T_{ir}), \mathcal{A}_i(T_{ir}); \boldsymbol{\Theta}_M \sim N(\boldsymbol{\Lambda} \eta_i(t_{ij}), \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_\epsilon).
$$

The distribution of the latent process $p(\boldsymbol{\eta}_i; \boldsymbol{\Theta}_S)$ is the product of p-dimensional conditional Gaussian distributions with either the form in Equation 5.6 or 5.8, depending on how treatment is assumed to impact the latent process.

When fitting this joint model, we must consider (i) how to ensure both the latent process parameters and the loadings matrix are identifiable and (ii) how to calculate the cumulative hazard, which requires integrating over the multivariate OU process. To make sure that the parameters in the longitudinal submodel are identifiable, we model the latent factors on the correlation scale. This strategy is common when fitting factor models and was previously described in Tran et al. (2021) [112] in the context of a similar factor model with a latent

OU process. By forcing the latent factors to have a stationary variance of 1, we fix the amount of variability in the process. The loadings matrix $\boldsymbol{\Lambda}$ then rescales the latent factors to capture their association with the measured longitudinal outcomes $\boldsymbol{Y}$. To implement this identifiability constraint, we follow the approach described in Tran et al. (2021) [112] and reparameterize the OU process: instead of estimating parameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$, we estimate $\boldsymbol{\theta}$ and the off-diagonal elements $\boldsymbol{\rho}$ of the stationary correlation matrix of the OU process $\boldsymbol{V}$ (this matrix was previously defined in Section 5.3.1 as $\boldsymbol{V} = vec^{-1}\big\{(\boldsymbol{\theta} \oplus \boldsymbol{\theta})^{-1} vec(\boldsymbol{\sigma}\boldsymbol{\sigma}^\top)\big\}$. For a bivariate ($p = 2$) OU process, the matrix that we estimate is $\boldsymbol{V} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. In addition to requiring that the OU process has a stationary variance of 1, we require that $\boldsymbol{\theta}$ has eigenvalues with positive real parts. This constraint is discussed in more detail in Tran et al. (2021) [112]. We also require that $\boldsymbol{\Lambda}$ contains structural zeros and that the location of these structural zeros are known. This constraint means that we know which of the observed longitudinal outcomes are measurements of which of the latent factors. In practice, domain knowledge can help inform the placement of these structural zeros. All non-zero elements of $\boldsymbol{\Lambda}$ must be positive.

Calculating the cumulative hazard function as written in the likelihood in Equation 5.12 would require integrating over the multivariate continuous-time OU process. To avoid this complex integration, we take a discrete approximation based on a midpoint rule. Specifically, within the Bayesian algorithm, we generate values of the latent process on a fine grid and then sum the hazard across this fine grid using a midpoint rule that allows us to closely approximate the integral with a sum. We previously used this strategy in the context of a joint longitudinal-survival model (for single time-to-event outcomes) and found that, in the ILD setting, parameters' posterior distributions were not sensitive to the choice of grid width. This sensitivity analysis, along with a more detailed description of this midpoint approximation to the cumulative hazard, can be found in Chapter 4.

We use Stan, a software that carries out an HMC sampling algorithm, to fit our model [18]. The priors that we use are given in Section D.4.

After fitting the joint model, it may be of interest to compare models that specify, for example, different mechanisms for the effect of treatment or different structures for the loadings matrix and latent factors. To do so, DIC and Watanabe-Akaike information criterion (WAIC) [123] can be used. Gelman et al. (2014) [30] define these information criteria as consisting of two terms: one term for the log-likelihood and another term that captures the effective number of parameters. Letting $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_R, \boldsymbol{\Theta}_M, \boldsymbol{\Theta}_S)$, DIC is defined as:

$$DIC = -2logp(Y, T, \delta|\hat{\boldsymbol{\Theta}}) + 2p_{DIC}$$

where $\hat{\boldsymbol{\Theta}}$ is the posterior mean; p_{DIC} is the effective number of parameters, with $p_{\text{DIC}} = 2\left(logp(Y,T,\delta|\hat{\boldsymbol{\Theta}}) - \frac{1}{S}\sum_{s=1}^{S} logp(Y,T,\delta|\hat{\boldsymbol{\Theta}}^s)\right)$; and s indexes posterior samples. Similarly, WAIC is computed as:

$$WAIC = -2\widehat{\text{lppd}} + 2p_{\text{WAIC}}$$

where $\widehat{\text{lppd}} = \sum_{i=1}^{N} log\left(\frac{1}{S}\sum_{s=1}^{S} p(Y_i,T_i,\delta_i|\boldsymbol{\Theta}_R^s,\boldsymbol{\Theta}_M^s,\boldsymbol{\Theta}_S^s)\right)$ and the effective number of parameters is $p_{\text{WAIC}} = \sum_{i=1}^{N} V_{s=1}^{S}\left(logp(Y_i,T_i,\delta_i|\boldsymbol{\Theta}_R^s,\boldsymbol{\Theta}_M^s,\boldsymbol{\Theta}_S^s)\right)$, with $V_{s=1}^{S}(a_s) = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$. Both of these information criteria rely on evaluating the log-likelihood, either at the posterior mean $\hat{\boldsymbol{\Theta}}$ or each of the posterior samples $\boldsymbol{\Theta}^s, s = 1, ..., S$. When fitting the model in Stan, we rely on the conditional likelihood, $logp(Y_i,T_i,\delta_i|\mathcal{H}_i;\boldsymbol{\Theta})$, rather than a version of the likelihood that is marginalized over the latent process. The marginal version, however, is generally recommended when comparing the fits of latent variable models [89, 60]. To compute DIC and WAIC, we must integrate the conditional likelihood in Equation 5.12 over the continuous-time multivariate stochastic process in the longitudinal and event submodels. Specifically, we use a Monte Carlo-based approach to sample values of the latent process and approximate the integral, as described in Section D.6. This approach allows us to estimate the value of the marginal log-likelihood at both the posterior mean $\hat{\boldsymbol{\Theta}}$ and each of the posterior samples $\boldsymbol{\Theta}^s$, which we then use to calculate DIC and WAIC.

5.4 Simulation Study

The goal of the simulation study is to assess the statistical properties of correctly specified models fit to simulated datasets informed by other mHealth studies of smoking cessation. When generating data, we use two sets of true parameter values—called setting 1 and setting 2—which are informed by parameter estimates from observational mHealth studies similar to the motivating mHealth MRT. Our simulated data, however, are slightly simpler than the data from these mHealth studies to modulate the computational cost of conducting the simulation study.

5.4.1 Data Generation

A single simulated dataset consists of $N = 100$ individuals who are followed for 14 days. At four random times each data, we generate observations of our measured longitudinal outcomes, $\boldsymbol{Y}$, which consists of $k = 4$ observed outcomes. We assume that these 4 observed outcomes are measurements of $p = 2$ latent factors. Treatments are sent randomly once per day and the effect of each treatment lasts for half a day ($\delta_a = 0.5, \delta_b = 0.5$). Depending

on the exact time at which each treatment is delivered, two treatments may be active at the same time. As discussed earlier, we do not consider the setting in which treatment directly impacts the longitudinal outcome, so the measurement submodel always takes the form described in Equation 5.1. For the structural submodel, we assume that the treatment impacts the latent process as either an additive shift to the mean (Eq. 5.4) or through a drift term (Eq. 5.7). The treatment function $\boldsymbol{\mu}_i(t)$ is defined as in Equation 5.5. For each version of the structural submodel, we consider two different event-time submodels:

1. Treatment impacts the hazard of event r through the latent process and treatment modifies the hazard directly.

$$h_{ir}(t) = h_0 \exp\left\{\beta_1 \eta_{1i}(t) + \beta_2 \eta_{2i}(t) + \tilde{\mu}_i(t)\right\}$$

2. Treatment impacts the hazard of event r through the latent process and treatment modifies the hazard directly. The hazard also depends on the time since the most recent prior $(r-1)^{th}$ event, where the function relating the two is specified such that experiencing an event (i.e., engaging in substance use) temporarily increases the risk of a subsequent event, but this temporary increase decays to 0 after a certain amount of time.

$$h_{ir}(t) = h_0 \exp\left\{\beta_1 \eta_{1i}(t) + \beta_2 \eta_{2i}(t) + \beta_3 g(t - t_{i,r-1}) + \tilde{\mu}_i(t)\right\}$$

For setting 1, $g(x) = \frac{1}{1+\exp\{4(x-2)\}}$ and for setting 2, $g(x) = \frac{1}{1+\exp\{1.5(x-2)\}}$. We assume that these functions are known when fitting the model. $\tilde{\mu}_i(t)$ is given in Equation 5.11 with $\delta_b = 0.5$. In both of these hazard models, we assume that the baseline hazard is constant, $h_0 = \exp(\beta_0)$.

When simulating the data, we assume that the treatment has a positive impact on one of the latent factors and a negative impact on the other. In the motivating case study, we might expect that receiving a prompt would temporarily increase positive affect (corresponding to $\tau_1 = 2$) while decreasing negative affect (corresponding to $\tau_2 = -1$). These effects are interpreted as either occurring on the scale of the mean (if the treatment effect is modeled additively) or on the scale of the derivative (if the treatment effect is modeled as drift). We also assume that receiving a prompt decreases the hazard of an event (via $\tilde{\tau} = -0.8$), which is what the prompts in the MRT aim to do by encouraging engagement in certain behavioral strategies. True values for the other model parameters, which are informed by models fit to other mHealth smoking cessation studies, are given in Section D.5.

In Section D.5, we provide some plots of the simulated data. The number of observed

events ranges from an average of 1.9 events per person in setting 2 when the treatment effect modeled through the drift term and the hazard takes the form of model 2, to an average of 4.7 events per person in setting 2 when the treatment effect modeled as an additive shift to the latent process and the hazard takes the form of model 1.

5.4.2 Results

For each parameterization of the treatment effect on the latent process (additive or drift) and for each hazard model (1 or 2), we simulate 100 datasets and fit the model. We repeat this process for the true parameter values corresponding to both setting 1 and setting 2. When fitting the joint model, we initialize parameters at values with approximately the correct order of magnitude and with the correct sign. In practice, a two-stage approach could be used to determine reasonable starting values when true values are not known. For each dataset, we run 1 chain for 2000 iterations and discard the first 1000 iterations as burn-in.

To assess bias and variance, we compare the posterior medians to the data-generating values (see Figure 5.4) and evaluate the coverage of 95% credible intervals (see Figure 5.5). We find that the posterior medians are close to the true values and that the posterior distributions have close-to-nominal coverage. In hazard model 2, point estimates for β_3—the parameter that captures the association between the hazard and a known function of time since the most recent event—show some bias and, as a result of this bias, lower-than-nominal coverage. We further investigate the bias in this parameter with some additional simulations, which we present in more detail in Section D.5.1. These supplemental simulations suggest that this bias is likely related to fitting a rather complicated model to a fairly small dataset; they also suggest that the higher temporal correlation in the latent process in setting 1, compared to setting 2, makes recovering unbiased estimates of β_3 more difficult. Overall, however, the magnitude of the bias in our point estimates for β_3 is still small and reasonable given the complexity of this joint model.

5.5 Analysis of the MRT Data

We use our joint longitudinal recurrent event model to analyze data from the motivating MRT. We use the longitudinal submodel to summarize the observed 15 emotions as two latent factors that represent positive affect and negative affect. We consider two models for the repeated treatments (i.e., the app-based prompts): one in which treatment has an additive impact directly on the mean of the latent factors and another in which treatment impacts the dynamics of the latent factors on the derivative scale through the time-varying

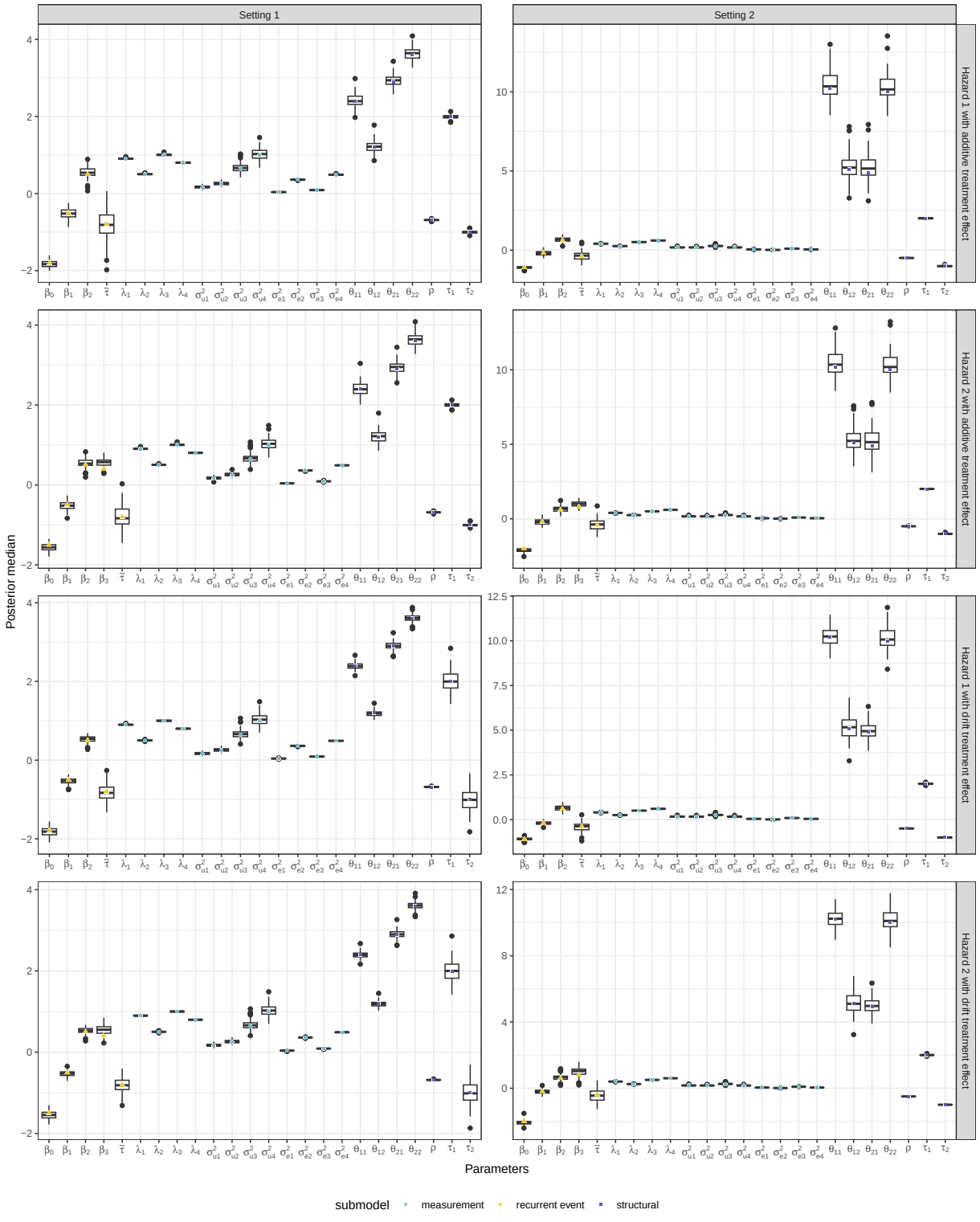


Figure 5.4: For data generated under settings 1 and 2 with different hazards and treatment effect models, we use box plots to summarize the distribution of the **posterior medians for all parameters** across the 100 simulated datasets. When fitting the model, we assume that the grid used in the midpoint approximation of the cumulative hazard function has a width of 0.5 days. True parameter values are indicated with colored dots.

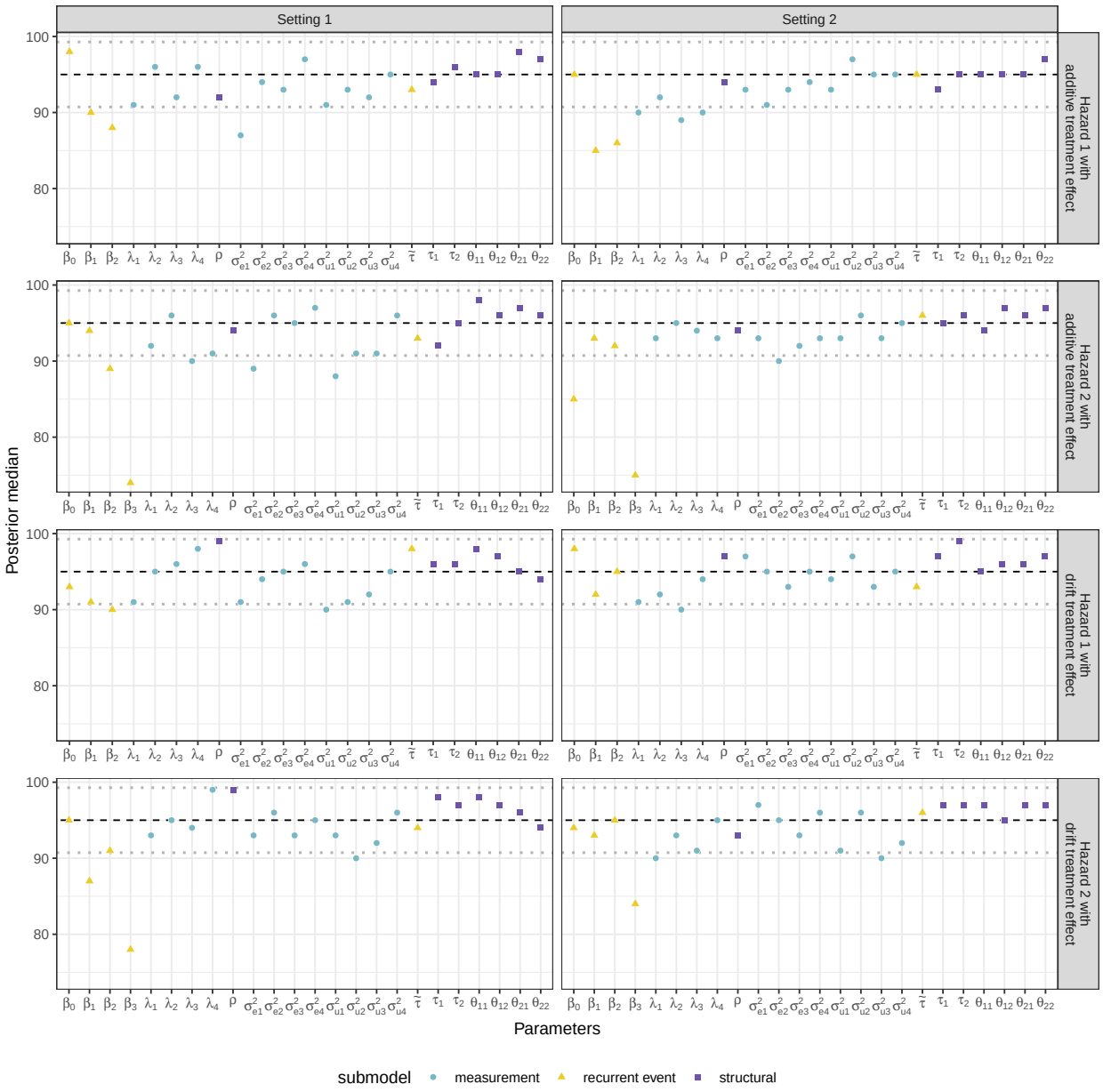


Figure 5.5: For data generated under settings 1 and 2 with different hazards and treatment effect models, we summarize the **coverage rate of 95% credible intervals** across the 100 simulated datasets with the colored dots. The black horizontal dashed lines indicate target coverage and the dotted grey lines corresponds to the upper and lower bounds of a 95% binomial proportion confidence interval for a probability of 0.95.

drift term. In the recurrent event submodel, we model the hazard of recurrent poly-substance use events using the time-varying predictors of positive affect, negative affect, and repeated treatment effects. The hazard model is:

$$h_{ir}(t) = h_0(t) \exp\left[\beta_1 \eta_{1i}(t) + \beta_2 \eta_{2i}(t) + \tilde{\mu}_i(t)\right]$$

We assume that the duration of the treatment effect is 8 hours for both submodels ($\delta_a = \delta_b = 8$ hours). We specify a Weibull baseline hazard; we use a clock-reset approach and reset the baseline hazard after the occurrence of each event. The time-varying latent factors, $\eta_{1i}(t)$ and $\eta_{2i}(t)$, are still functions of time since the start of the study. In the motivating case study, participants are instructed to quit smoking on day 4; we account for the change in frequency of substance use before and after day 4 by specifying a separate baseline hazard for the pre- and post-quit periods. When modeling the impact of treatment directly on the hazard (i.e., when specifying $\tilde{\mu}_i(t)$), we consider one version that assumes the treatment parameter $\tilde{\tau}$ is constant across the entire study and another other version that allows this treatment parameter to differ between the pre- and post-quit periods (i.e., we have $\tilde{\tau}_{\text{pre}}$ and $\tilde{\tau}_{\text{post}}$). Further details on these hazard models are given in Section D.1.2.

One additional advantage of using a Weibull baseline hazard is that it is potentially flexible enough to allow the underlying risk of an event to change as a function of the time since the prior event. In an attempt to avoid specifying an overly-complex model given our limited sample size, we opt to account for potential changes in the hazard of an event as a function of time since the previous event through this Weibull baseline hazard with the clock reset, rather than by modeling the hazard as a separate function of time since the most recent event (i.e., by including a $\beta_3 g(\cdot)$ term in our hazard model).

When fitting the model, we rescale follow-up in the study so that the 10-day interval has time units in the range of 0 to 1 (this approach is suggested in Tran et al. (2021) [112] to help deal with potentially oscillating OU processes). This rescaling of the time interval impacts the interpretation of the $\boldsymbol{\theta}$ parameter in the structural submodel and the Weibull shape and scale parameters in the baseline hazard.

We use a two-stage approach to set initial parameter values: we first fit the longitudinal submodels and initialize these parameters at their posterior medians. Using the posterior values of η, we fit a simpler hazard model to get initial estimates of the recurrent event submodel parameters. To fit the full joint model, we use 4 chains with 3,000 samples and discard the first 1,000 as burn-in. We find that fitting a joint model that assumes treatment impacts the latent process in an additive way, rather on the derivative scale, is easier for a dataset of this size. For more details on model convergence, please see Section D.1.3.

We compare the fits of these different joint models using DIC and WAIC, which are calculated using the marginal log-likelihood as described in Section 5.3.3. We focus on WAIC, as this measure is generally preferred over DIC [31], and present WAIC for each model in Table 5.1. DIC is provided in the Appendix (see Table D.3) and supports the same conclusions as WAIC. The values of these information criteria indicate that a time-varying drift treatment effect model for the latent process and two treatment-related parameters in the hazard model fit our data the best. The WAIC for this model, however, is only very slightly better (smaller) than the WAIC for the joint model with the same hazard but an additive treatment model for the latent process. Due to the sampling-based approach we use when calculating the marginal log-likelihood, some uncertainty does exist in the exact value of WAIC and so we emphasize that our main motivation for presenting two different approaches for modeling the impact of treatment on the latent process is that these approaches correspond to different scientific beliefs about the mechanisms by which treatment impacts the latent process. Given the somewhat limited sampled size of this MRT data ($N = 64$ individuals), it may be difficult to assess which treatment model is preferred.

Posterior medians and 95% credible intervals are shown for all four joint models in Figure 5.6. We see that the estimated correlation (ρ) between the latent factor for positive affect and the latent factor for negative affect is negative, as expected. Across all models, posterior means for this parameter ρ range from -0.48 to -0.45. We also see that the positive emotions have larger variance estimates for their item-specific random intercepts, compared to those for the negative emotions. When we examine the data directly, we see that the empirical variability in individual-specific averages for each emotion tends to be slightly higher for positive emotions than for negative emotions, supporting this result (see Section D.1.3.3 in the Appendix for more details).

Across all joint models, we find that the estimated effect of treatment ($\boldsymbol{\tau}$) in the treatment model for the latent process is near zero. This result is the same whether we model the impact of treatment on the latent process as an additive shift to the mean or on the derivative scale. When assuming a time-varying drift treatment effect model on the latent process, WAIC indicates that allowing the coefficient on the treatment function in the hazard model to differ between the pre- and post-periods improves the fit of the model, compared to assuming a single treatment-related parameter in the hazard. When examining posterior parameter estimates, we see a slightly stronger protective effect of sending a prompt to an individual after they have attempted to quit ($\tilde{\tau}_{\text{post}}$), compared to before quit ($\tilde{\tau}_{\text{pre}}$).

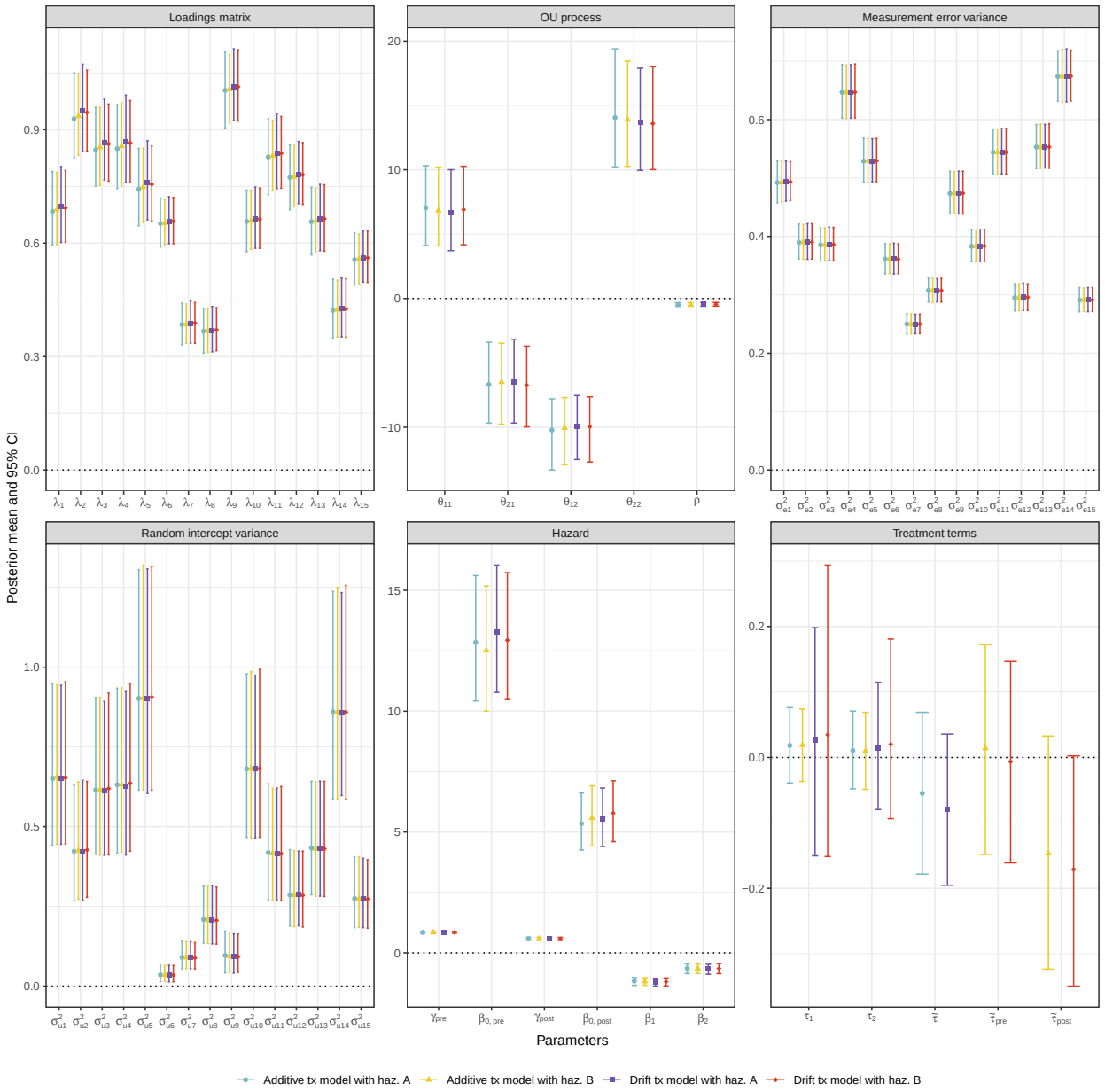


Figure 5.6: Posterior means and 95% credible intervals for parameters in the four different joint models that make different assumptions about how the treatment impacts the latent process and hazard. Hazard model A assumes a single treatment-related parameter in the hazard model; hazard model B allows for separate pre- and post-quit treatment parameters. Emotions are indexed as: 1 = grateful, 2 = happy, 3 = proud, 4 = relaxed, 5 = enthusiastic, 6 = angry, 7 = ashamed, 8 = guilty, 9 = irritable, 10 = lonely, 11 = anxious, 12 = sad, 13 = restless, 14 = bored, 15 = hopeless.

WAIC	Impact of treatment on latent process	
Hazard model	Additive	Drift
Single treatment parameter	61,861.17	62,224.93
Separate pre- and post-quit treatment parameters	61,723.92	**61,712.76**

Table 5.1: WAIC for the joint models fit to the motivating MRT data. When approximating the marginal log-likelihood, we subsampled every 5th iteration of the final 1,000 posterior samples to use in our calculations. We used a value of $M = 25$ when generating marginal posterior samples of the latent process. A lower value of WAIC is preferred. The lowest value of WAIC is indicated in bold text.

5.6 Discussion

In this paper, we propose a method for jointly modeling multivariate ILD, a recurrent event outcome, and the effect of repeatedly-delivered treatments, as often found in data from MRTs. We present two different ways to model the impact of treatment on the longitudinal latent process—one in which the treatment effect is modeled through an additive shift to the mean and other in which the treatment model alters the latent process on the derivative scale through a time-varying drift term. We consider both the additive and drift versions in this work, as our overall goal is to present different ways to model the impact of a JITAI on a low-dimensional stochastic latent process, which here is observed through a larger number of measured outcomes.

In the longitudinal submodel, we only consider modeling the impact of treatments on the latent process. In addition to allowing treatments to alter the trajectory of the latent process, it might be reasonable to assume that the latent process is modified by the occurrence of an event. In this case, one could use the same approach that we describe for modeling the impact of treatment on the latent process, but substitute treatment times with event times. The impact of treatments and events on the latent process could also be modeled simultaneously. We do not consider this variation of the longitudinal submodel in our simulations or case study, but mention it now as it may be of interest in other settings.

A limitation of our joint model is that this approach is computationally demanding. For example, fitting the joint models to the case study data required between 38 hours (when the model for treatment effect on the latent process was additive and the hazard had two treatment-related coefficients) and 121 hours (when the model for treatment effect on the latent process assumed the drift form and the hazard had two treatment-related coefficients) when running four chains of 3000 samples each in parallel across four cores. Additional time (on a scale similar to that required for model fitting) is also needed to compute the marginal log-likelihood used in DIC/WAIC calculations, which adds to the

overall computational burden of this approach. Parallelization can substantially speed up DIC and WAIC calculations so that they take under an hour from the analyst's perspective. Although our total computation time is still small compared to the amount of time required to design and conduct an MRT, improving the speed of inference is a practically important area for future work. Improving computation time will also facilitate additional model comparisons and sensitivity analyses.

In the case study, we demonstrate how DIC and WAIC can be used to compare models that make different assumptions about how the longitudinal, recurrent event, and treatment-related components of the joint model are associated with each other. We focus on using these measures to compare joint models as a whole; Zhang et al. (2017) [131] propose a version of DIC for joint longitudinal-survival models where DIC is decomposed into separate contributions from the longitudinal and survival submodels. Adapting this version of DIC to our setting could be a potentially interesting and useful extension.

When modeling treatment in the longitudinal submodel and the recurrent event submodel via the functions $\boldsymbol{\mu}_i(t)$ and $\tilde{\mu}_i(t)$, we make assumptions about both the specific shape of the treatment effect over time and the duration of active treatment, δ_a and δ_b. In both our simulations study and the case study, we assume that the effect of treatment is at its maximum at the time of treatment and that the treatment "dosage" decreases with the time since delivery. We focus on estimating the maximum effect of treatment ($\boldsymbol{\tau}$ and $\tilde{\tau}$) at the time of delivery but assume that the form of its decay is known. We do so largely because the MRT motivating this work does not have a particularly large sample size ($N = 64$). Here, DIC and WAIC can be used to compare joint models with different pre-specified treatment decay functions, but in other settings where more data are available, we could try to estimate additional parameters that characterize the treatment function (e.g., δ_a or δ_b).

When analyzing the MRT, we assume that the effect of treatment is non-zero for 8 hours following delivery of the notification ($\delta_a = \delta_b = 8$ hours). This MRT was designed such that longitudinal outcomes (i.e., emotions) were measured via EMA approximately 1 hour after individuals were randomized to potentially receive treatment (i.e., a notification). As a result, the data should contain enough information to detect possible associations between treatment and changes in the latent process. One advantage of modeling the effect of treatment on the latent process, as opposed to only the hazard, is that the measurement frequency of the longitudinal outcomes—which are assumed to be generated by the latent process—is determined, at least partially, by the design of the study. That is, in the motivating MRT, EMAs were intended to be sent to participants up to six times per day; furthermore, the EMAs were specified to be sent 1 hour after each randomization. As a result of the frequent measurement occasions, we can potentially capture short term treatment effects that only

last, for example, for a third of a day. When modeling the impact of treatment directly on the hazard of recurrent events, our ability to detect a potential treatment effect is somewhat limited by the frequency of these recurrent events and their proximity to the time of treatment. That is, if we assume that the effect of treatment on the hazards is non-zero for only a third of a day and we rarely observe events occurring within half a day of treatment delivery, then capturing the direct effect of treatment on the hazard will be difficult. In the motivating MRT, events of poly-substance use are observed quite frequently: 1,162 total events across 64 individuals, with 566 of the events occurring within eight hours of treatment. In a different setting in which events are rarer, estimating $\tilde{\tau}$ may be more difficult. However, as long as the longitudinal outcomes are measured frequently enough, then one may still be able to capture an indirect effect of treatment on the hazard through the latent process.

So far, we have assumed that treatment is binary; that is, either a notification is sent or not sent. In reality, treatment in the motivating MRT can take three different levels: no notification, a notification encouraging engagement in a behavioral strategy requiring low effort, or a notification encouraging engagement in a behavioral strategy requiring high effort. Extending the models presented in this paper to account for multiple levels of treatment would be straightforward, but would require more parameters. Additionally, our current definition of treatment depends only on the outcome of randomization. This definition corresponds to an intent-to-treat analysis. Information on subjects' engagement with the intervention (i.e., whether or not they carried out the behavioral strategy suggested by the notification) is self-reported in the EMA delivered one hour after randomizaton. Engagement with the intervention, called treatment compliance in other settings, could also be incorporated into the definition of treatment to enable an as-treated analysis.

Finally, in our hazard model, we use responses from the substance use-related EMA questions to approximate the time of recurrent use. In practice, many EMA questionnaires are designed such that event outcomes are measured in an interval censored way; that is, questions are often phrased as, "Since the last EMA, have you [experience some sort of event or engaged in some sort of behavior]". Sometimes, as in the case of the substances considered in our case study, this question is followed up with questions that prompt an individual to provide the approximate time of the events or behaviors. Quite often, however, this subsequent set of questions is not asked and so the data only contain information on whether or not an event occurred within the time interval between consecutive EMAs. Thus, extending our method to handle this type of interval censored data could be practically useful.

CHAPTER 6

Conclusion

In this dissertation, we propose four different methods for modeling ILD and event data collected in mHealth studies. We present an approach for jointly modeling ILD; for modeling ILD and cumulative event outcomes; for modeling ILD and survival times; and for modeling ILD and recurrent event times with repeated treatment effects.

Advancements in joint modeling methods have been traditionally motivated by problems in cancer and HIV. In these settings, a single longitudinal outcome is generally of interest. Most often, these outcomes are measured on the scale of weeks or months and so modeling gradual changes in their levels—and how these levels relate to the hazard of an event—is both scientifically interesting and useful. Modeling smooth changes in the longitudinal outcomes is straightforward with methods such as linear mixed models or generalized linear mixed models. Splines can also be used to account for sharp transitions in the smooth trends. Although estimation is not trivial, integrating across linear, or possibly quadratic, functions of time in the longitudinal submodel is straightforward and can sometimes be done analytically. In this traditional setting, a combination of high computational cost and limited availability of user-friendly software previously inhibited the use of joint models in practice. Over the past decade or so, high-powered computing resources have become more accessible and, as a result, the computational cost of fitting traditional joint models is no longer prohibitive. Development of well-documented and comprehensive `R` packages, such as `JM` [87], `JMBayes2` [91], and `brms` [15, 16, 17], has also enabled analysts to use joint models while avoiding the need to implement code for estimation and inference themselves.

Although the longitudinal data considered in this dissertation are higher dimensional and more rapidly-varying than the type of data that has traditionally motivated joint models, our ILD are not truly big data in today's sense. EMAs facilitate frequent collection of a sizable number of outcomes (e.g., measurements 4-6 times per days of 10-20 emotions, plus additional state and context variables), but the frequency at which they can be sent to study participants is limited by the burden on the participants, as manually filling out these many-question surveys does require both time and effort. In addition to using EMAs, some

studies use sensors (e.g., accelerometers or GPS) to collect additional data. Longitudinal sensor-based measurements, which may include functional data, are generally much higher dimensional than the EMA-based ILD considered here. As a result, sensor data may contain a huge amount of information about complex and granular risk factors associated with vulnerability to events of interest. GPS data recorded by smartphones also facilitates the collection of network data, which may contain information on social interactions known to have an important role in influencing substance use [26]. Together with traditional EMA data, geographical momentary assessment may provide additional insights into substance use patterns and help researchers identify factors that increase or decrease the risk of substance use [25]. The complexity and size of this type of data creates many opportunities for the development of efficient and flexible methods for modeling. Developing methods to incorporate this type of information into joint models could allow researchers to further investigate how vulnerability to substance use or other health-related events of interest fluctuates as a function of social networks, movement, stress, or other factors related to participants' current state and context. Developing approaches that are computationally scalable to this type of data, however, would be challenging.

Not only do sensors enable collection of new types of predictors, they can also improve the measurement of events of interest. For example, Saleheen et al. (2015) [99] developed an algorithm that uses data collected from wrist sensors that record arm movements and respiration patterns to determine when individuals are likely to be smoking. Sensor data can contain a lot of noise, and so complex rules for data processing are required in order to define event outcomes suitable for modeling. Developing a joint model that can handle complex sensor-based data as an event outcome could be an interesting direction for future work. Such a method might allow analysts to avoid developing complicated deterministic rules in the data processing phase.

Importantly, the complex data collected from EMAs and sensors may not satisfy the traditional parametric assumptions of longitudinal submodels. For example, the noise in the longitudinal measurements may not be normally distributed or responses may be severely skewed. Some work has been done on developing model-checking approaches specifically for joint models (e.g., [94]), but the availability of new methods will become increasingly important as data become more and more complicated. If joint models are to be used to inform the development and assessment of interventions, for example, then it is crucial that a model accurately represents the data. Thus, as the complexity of data increases, and the models used for analysis do too, having an appropriate suite of tools to easily and accurately asses the validity of fitted models is important.

Beyond traditional biostatistical literature, the fields of finance math and econometrics

have traditionally used a variety of different stochastic time series models. For example, the Hull-White model used in Chapter 5 is often used to model stock market data but is less commonly applied in biomedical settings. Future research could draw more from financial modeling literature for additional flexible longitudinal models. Often, finance math focuses on univariate outcomes so some work would need to be done to adapt these models to the multivariate ILD setting. Appropriate model comparison and diagnostic measures would also need to be determined to ensure the most appropriate choice of model.

Although it is statistically interesting to build additional submodels into joint models, from a practical perspective, such complicated models are less likely to be used as they may require days of computation time. With enough patience, an analyst can reasonably model ILD without needing to put a huge amount of effort into the development of novel methods for computationally efficient estimation and inference; the same cannot be said for functional data. As mentioned throughout this dissertation, improving the computational efficiency of the algorithms used to fit these models is a major area for future work. Advancements in computational efficiency would also expand not only the types of data that could be modeled, but also the ways in which joint models could be used in practice—for example, for prediction purposes.

mHealth studies often include apps with dashboards that provide real-time feedback to participants on their progress towards certain goals or their engagement in certain behaviors. Joint models are useful for making predictions in exploratory work; for example, they can be used to examine the risk of certain events over time under hypothetical trajectories of a longitudinal process or under hypothetical treatment patterns. If joint models are to be used to provide real-time feedback in an mHealth app, then computational efficiency must be improved. Although the frequentist framework was used to fit the earliest versions of joint models, Bayesian approaches have become more popular as computing has improved, as they handle latent variables quite naturally. Approximate Bayesian algorithms, which are much faster than traditional sample-based Bayesian methods, have been recently proposed for the joint model setting (e.g., [98, 130, 116]). These types of methods could be explored further in the future. Computation time improvements would also facilitate sensitivity analyses to assess the assumptions required by model-based approaches for estimating treatment effects, including the one proposed in Chapter 5. If computation time could be addressed, then developing a flexible `R` package to fit the types of stochastic process joint models described in this dissertation could be useful.

While computation is one area with room for major developments, missing data is another. In mHealth studies, reasons for missing values can range from intentional pauses in data collection due to low battery power to unplanned software malfunctions. These causes of

missingness apply to both self-reported outcomes and sensor-based outcomes. Self-reported outcomes, however, have their own missing data challenges, as these data are dependent on individuals filling out surveys. These surveys can be long and repetitive and can quickly become somewhat burdensome to study participants. While some types of missingness may be ignorable, other types may be quite informative. In mHealth studies, the devices used for data collection often offer a small amount of additional information about why values may be missing. For example, many mHealth studies that prompt participants to respond to EMAs include rules that require participants to be available before sending a notification; defining "availability" prevents prompts from being sent in inconvenient or potentially dangerous situations, such as when someone is driving. Information on availability may be recorded in the data. The app may also detect and record when phone battery levels are low, as this may also impact whether or not an EMA is sent. On the other hand, participants may not respond to surveys because they do not feel like responding. This type of nonresponse that is associated with the level of the longitudinal outcome is distinct from the previously described types of missingness and so would ideally be treated as so in the model. Sensor-based measurements may help reduce some amount of missingness (for example, the work in Saleheen (2015) [99] could help reduce missingness in reported episodes of smoking). Sensors still malfunction, however, and so modeling missingness mechanisms is another important direction for future research. Some work has been done to develop an imputation-based approach for addressing missingness in longitudinal mHealth data [61], but additional contributions to the field are needed.

Finally, while the models that we present in each chapter of this dissertation build on those in previous chapters to become more complicated, at the same time the motivating datasets shrink in size. In Chapter 5, the size of the dataset ($N = 64$ individuals) makes both fitting the joint longitudinal-recurrent event model and estimating treatment effects challenging. We do not detect a statistically significant association between treatment and the longitudinal latent process; this could be because the app-based notifications do not impact the latent psychological states of positive and negative affect, or it could be because we have limited power to detect effects given the complexity of our model and the size of our dataset. The MRT, called the Affective Science MRT, that motivated the work in Chapter 5, is identical in design to another MRT, the Mobile Assistance for Regulating Smoking MRT [68], and so integrating data across these two studies could help increase power or improve precision. The non-interventional observational mHealth studies described in Chapters 2-4 are slightly different in design but share many similarities with the MRT. Data integration methods could be developed for use with the types of latent variable stochastic process joint models described in this dissertation. These data integration methods could potentially help

improve our ability to identify the pathways and mechanisms by which treatment impacts the longitudinal and event processes. However, increasing the size of the datasets through data integration methods could bring additional computational challenges. This computational concern reinforces the importance of future work focusing efficient approaches for estimation and inference.

Overall, the combination of advances in computing power and the increasing availability of rich data collected by mHealth technology creates interesting opportunities and important challenges for the area of joint modeling. In this dissertation, we describe four methods motivated by the intersection of ILD and joint modeling, but substantial opportunities and challenges remain.

APPENDIX A

Supplementary Material for: A Continuous-Time Dynamic Factor Model for Intensive Longitudinal Data Arising from Mobile Health Studies

A.1 Derivation of the Analytic Form of the Conditional Covariance Function of the OU Process

Assume $\eta(t)$ is a p-dimensional OU stochastic process with a marginal mean of 0. From Vatiwutipong and Phewchean (2019) [117], if we assume that the initial state $\eta(t_0 = 0)$ is known, then the cross-covariance function of the OU process at times s and t is

$$Cov\{\eta(s), \eta(t)|\eta(t_0 = 0)\} = \int_0^{\min(s,t)} e^{-\theta(s-u)}\sigma\sigma^\top e^{-\theta^\top(t-u)}du$$

where e^A is the matrix exponential. Note that we can assume that $t_0 = 0$ without loss of generality because this stochastic process is stationary. Using the identity for matrices A, B, and C that $vec(ABC) = (C^\top \otimes A)vec(B)$, we can re-write the vectorized version of the cross-covariance function as

$$vec\{Cov\{\eta(s), \eta(t)|\eta(t_0)\}\} = \int_0^{\min(s,t)} e^{-\theta(t-u)} \otimes e^{-\theta(s-u)}\ du\ vec\{\sigma\sigma^\top\}$$

We can also use the identity that $e^A \otimes e^B = e^{A\oplus B}$, so

$$vec\{Cov\{\eta(s), \eta(t)|\eta(t_0)\}\} = \int_0^{\min(s,t)} e^{[-\theta(t-u)]\oplus[-\theta(s-u)]}du\ vec\{\sigma\sigma^\top\} \tag{A.1}$$

Next, we simplify Equation A.1 by pulling all the u's into a single term. For now, focus

on the term in the exponential:

$$\begin{aligned}[-\theta(t-u)] \oplus [-\theta(s-u)] &\overset{(a)}{=} -\theta(t-u) \otimes I + I \otimes (-\theta(s-u)) \\ &= -t(\theta \otimes I) + u(\theta \otimes I + I \otimes \theta) - s(I \otimes \theta) \\ &= -(t\theta \oplus s\theta) + u(\theta \oplus \theta)\end{aligned}$$

where equality (a) is by the definition of the Kronecker sum; $A \oplus B = A \otimes I_B + I_A \otimes A$, where I_A and I_B are identity matrices with dimensions of A and B, respectively. Now, substituting this new term back into the exponential term in Equation A.1, we get

$$e^{[-\theta(t-u)] \oplus [-\theta(s-u)]} = e^{-(t\theta \oplus s\theta) + u(\theta \oplus \theta)} \tag{A.2}$$

We can simplify this further using the identity $e^{A+B} = e^A e^B$ if A and B commute. Letting $A = (t\theta) \oplus (s\theta)$ and $B = (\theta \oplus \theta)$, we first show that these terms commute:

$$\begin{aligned}A \cdot B =& [(t\theta) \oplus (s\theta)] \cdot [\theta \oplus \theta] \\ =& [t\theta \otimes I + I \otimes s\theta] \cdot [\theta \otimes I + I \otimes \theta] \\ =& (t\theta \otimes I)(\theta \otimes I) + (t\theta \otimes I)(I \otimes \theta) + (I \otimes s\theta)(\theta \otimes I) + (I \otimes s\theta)(I \otimes \theta) \\ =& (t\theta \otimes I)(\theta \otimes I) + (I \otimes \theta)(t\theta \otimes I) + (\theta \otimes I)(I \otimes s\theta) + (I \otimes s\theta)(I \otimes \theta) \\ =& (\theta \otimes I)\left[(t\theta \otimes I) + (I \otimes s\theta)\right] + (I \otimes \theta)\left[(t\theta \otimes I) + (I \otimes s\theta)\right] \\ =& \left[(\theta \otimes I) + (I \otimes \theta)\right] \cdot \left[(t\theta \otimes I) + (I \otimes s\theta)\right] \\ =& \left[(\theta \oplus \theta)\right] \cdot \left[(t\theta \oplus s\theta)\right]\end{aligned}$$

where line 4 uses the mixed-product property of the Kronecker product.

Referring back to Equation A.2, we now have

$$e^{-(t\theta \oplus s\theta) + u(\theta \oplus \theta)} = e^{-(t\theta \oplus s\theta)} e^{u(\theta \oplus \theta)}$$

We can substitute this term into Equation A.1 to get

$$\begin{aligned}vec\{Cov\{\eta(s), \eta(t) | \eta(t_0 = 0)\}\} &= \int_0^{\min(s,t)} e^{-(t\theta \oplus s\theta)} e^{u(\theta \oplus \theta)} du \; vec\{\sigma\sigma^\top\} \\ &= \int_0^{\min(s,t)} e^{u(\theta \oplus \theta)} du \; e^{-(t\theta \oplus s\theta)} vec\{\sigma\sigma^\top\}\end{aligned}$$

Now that we have rewritten the conditional cross-covariance function in this form, the

only term that we need to integrate is $e^{u(\theta\oplus\theta)}$. We find

$$\int_0^{min(s,t)} e^{u(\theta\oplus\theta)}du = (\theta\oplus\theta)^{-1}\big[e^{min(s,t)(\theta\oplus\theta)} - I\big]$$

We now have an integral-free analytic form of the conditional cross-covariance function:

$$vec\{Cov\{\eta(s),\eta(t)|\eta(t_0=0)\}\} = (\theta\oplus\theta)^{-1}\big[e^{min(s,t)(\theta\oplus\theta)} - I\big]e^{-(t\theta\oplus s\theta)}vec\{\sigma\sigma^\top\}$$

Note that if $s = t$, then the conditional cross-covariance function simplifies to the conditional covariance function given in Vatiwutipong and Phewchean (2019) [117].

A.2 Derivation of the Analytic Form of the Marginal Covariance Function of the OU Process

The analytic form of the conditional covariance function, given in Section 2.3.3, is based on the assumption that the initial state $\eta(t_0)$, with $t_0 = 0$ is *known*. We now derive the analytic form of the unconditional cross-covariance function that accounts for the additional uncertainty of an unknown initial state. From Vatiwutipong and Phewchean (2019) [117], if $\eta(t_0)$, with $t_0 = 0$, is known, then

$$\mathbb{E}\{\eta(t)|\eta(t_0)\} = e^{-\theta t}\eta(t_0)$$

Assuming that $s \le t$, from Lemma 1, we have

$$Cov\{\eta(s),\eta(t)|\eta(t_0)\} = vec^{-1}\Big\{(\theta\oplus\theta)^{-1}\Big[e^{(\theta\oplus\theta)s} - I\Big]e^{-(\theta t\oplus\theta s)}vec\{\sigma\sigma^\top\}\Big\}$$

If $\eta(t_0)$ is *unknown* and $t_0 = 0$, then using the Law of Total Covariance we can calculate

$$\begin{aligned}
Cov\{\eta(s),\eta(t)\} &= \mathbb{E}\big\{Cov\big(\eta(s),\eta(t)|\eta(t_0)\big)\big\} + Cov\big\{\mathbb{E}\big(\eta(s)|\eta(t_0)\big),\mathbb{E}\big(\eta(t)|\eta(t_0)\big)\big\} \\
&= vec^{-1}\Big\{(\theta\oplus\theta)^{-1}\Big[e^{(\theta\oplus\theta)s} - I\Big]e^{-(\theta t\oplus\theta s)}vec\{\sigma\sigma^\top\}\Big\} \\
&\quad + Cov\big\{e^{-\theta s}\eta(t_0), e^{-\theta t}\eta(t_0)\big\} \\
&= vec^{-1}\Big\{(\theta\oplus\theta)^{-1}\Big[e^{(\theta\oplus\theta)s} - I\Big]e^{-(\theta t\oplus\theta s)}vec\{\sigma\sigma^\top\}\Big\} \\
&\quad + e^{-\theta s}Var\big\{\eta(t_0)\big\}[e^{-\theta t}]^\top
\end{aligned}$$

If we assume that $\eta(t_0)$ is drawn from the stationary distribution, then $Var(\eta(t_0)) =$

$vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\}$. Then, we have

$$\begin{aligned} Cov\{\eta(s), \eta(t)\} =& vec^{-1}\Big\{(\theta \oplus \theta)^{-1}\Big[e^{(\theta\oplus\theta)s} - I\Big]e^{-(\theta t \oplus \theta s)} vec\{\sigma\sigma^\top\}\Big\} \\ &+ e^{-\theta s} vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\}[e^{-\theta t}]^\top \end{aligned}$$

Now we simplify this function. Consider the terms involving θ in the first term of the sum,

$$(\theta \oplus \theta)^{-1}\Big[e^{(\theta\oplus\theta)s} - I\Big]e^{-(\theta t \oplus \theta s)}$$

We can simplify this expression using the fact that $e^A e^B = e^B e^A$ in our setting. This property means that both

$$(\theta \oplus \theta)^{-1}\big[e^{s(\theta\oplus\theta)} - I\big]e^{-(t\theta \oplus s\theta)} = e^{-(t\theta\oplus s\theta)}(\theta \oplus \theta)^{-1}\big[e^{s(\theta\oplus\theta)} - I\big] \tag{A.3}$$

and

$$(\theta \oplus \theta)^{-1}\big[e^{s(\theta\oplus\theta)} - I\big]e^{-(t\theta \oplus s\theta)} = (\theta \oplus \theta)^{-1}e^{-(t\theta\oplus s\theta)}\big[e^{s(\theta\oplus\theta)} - I\big] \tag{A.4}$$

Setting Equations A.3 and A.4 equal and cancelling the final term implies that

$$e^{-(t\theta \oplus s\theta)}(\theta \oplus \theta)^{-1} = (\theta \oplus \theta)^{-1}e^{-(t\theta\oplus s\theta)}$$

We will use this proof of the commutative property later and now return to our expression for the unconditional cross-covariance function, $Cov\{\eta(s), \eta(t)\}$,

$$\begin{aligned} Cov\{\eta(s), \eta(t)\} =& vec^{-1}\Big\{(\theta \oplus \theta)^{-1}\Big[e^{(\theta\oplus\theta)s} - I\Big]e^{-(\theta t \oplus \theta s)} vec\{\sigma\sigma^\top\}\Big\} \\ &+ e^{-\theta s} vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\}[e^{-\theta t}]^\top \end{aligned} \tag{A.5}$$

Consider the second term in the sum,

$$e^{-\theta s} vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\}[e^{-\theta t}]^\top$$

By applying the identity $vec(ABC) = (C^\top \otimes A)vec(B)$, we can rewrite the vectorized form of the expression as

$$\begin{aligned} vec\big\{e^{-\theta s} vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\}[e^{-\theta t}]^\top\big\} &= e^{-\theta t} \otimes e^{-\theta s} vec\big\{vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\} \\ &= e^{-\theta t} \otimes e^{-\theta s}(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\} \\ &= e^{-(\theta t \oplus \theta s)}(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\} \end{aligned}$$

Reversing the vectorization operation and applying the commutative property, we then get

$$\begin{aligned} e^{-\theta s} vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\}[e^{-\theta t}]^\top &= vec^{-1}\big\{e^{-(\theta t \oplus \theta s)}(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\} \\ &= vec^{-1}\big\{(\theta \oplus \theta)^{-1} e^{-(\theta t \oplus \theta s)} vec\{\sigma\sigma^\top\}\big\} \end{aligned}$$

Plugging the term above into the second term of Equation A.5, the cross-covariance function becomes

$$\begin{aligned} Cov\{\eta(s), \eta(t)\} &= vec^{-1}\Big\{(\theta \oplus \theta)^{-1}\Big[e^{(\theta \oplus \theta)s} e^{-(\theta t \oplus \theta s)} - e^{-(\theta t \oplus \theta s)}\Big] vec\{\sigma\sigma^\top\}\Big\} \\ &\quad + vec^{-1}\Big\{(\theta \oplus \theta)^{-1} e^{-(\theta t \oplus \theta s)} vec\{\sigma\sigma^\top\}\Big\} \\ &= vec^{-1}\Big\{(\theta \oplus \theta)^{-1} e^{(\theta \oplus \theta)s} e^{-(\theta t \oplus \theta s)} vec\{\sigma\sigma^\top\} - (\theta \oplus \theta)^{-1} e^{-(\theta t \oplus \theta s)} vec\{\sigma\sigma^\top\}\Big\} \\ &\quad + vec^{-1}\Big\{(\theta \oplus \theta)^{-1} e^{-(\theta t \oplus \theta s)} vec\{\sigma\sigma^\top\}\Big\} \\ &= vec^{-1}\Big\{(\theta \oplus \theta)^{-1}\big[e^{(\theta \oplus \theta)s} e^{-(\theta t \oplus \theta s)}\big] vec\{\sigma\sigma^\top\}\Big\} \\ &= vec^{-1}\Big\{(\theta \oplus \theta)^{-1}\big[e^{(\theta \oplus \theta)s - (\theta t \oplus \theta s)}\big] vec\{\sigma\sigma^\top\}\Big\} \end{aligned} \tag{A.6}$$

Equation A.6 is the marginal cross-covariance function of the OU process when the initial state at time $t_0 = 0$ is *unknown*.

A.3 Derivation of the Precision Matrix for the OU Process

We derive the sparse precision matrix for the multivariate OU process assuming an unknown initial state. This sparsity results from the Markov property. We use Ω to represent the precision matrix and Ψ for the covariance matrix.

First, we start in the simplest setting in which we assume a stationary univariate OU process with evenly spaced measurement occasions. The spacing of the measurement times is given by $|t_j - t_{j-1}| =: d > 0$. The covariance matrix takes the form,

$$\Psi = \frac{\sigma^2}{2\theta}\begin{bmatrix} 1 & e^{-\theta d} & \dots & e^{-\theta(n-2)\cdot d} & e^{-\theta(n-1)\cdot d} \\ e^{-\theta d} & 1 & \dots & e^{-\theta(n-3)\cdot d} & e^{-\theta(n-2)\cdot d} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-\theta(n-2)\cdot d} & e^{-\theta(n-3)\cdot d} & \dots & 1 & e^{-\theta d} \\ e^{-\theta(n-1)\cdot d} & e^{-\theta(n-2)\cdot d} & \dots & e^{-\theta d} & 1 \end{bmatrix}$$

We know that the univariate OU process is equal to the AR(1) process when measure-

ments are evenly spaced, so the OU process precision matrix (assuming evenly spaced measurements) can be expressed as

$$\Omega = \frac{2\theta}{\sigma^2}\frac{1}{1-e^{-2\theta d}}\begin{bmatrix} 1 & -e^{-\theta d} & \dots & 0 & 0 \\ -e^{-\theta d} & 1+e^{-2\theta d} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1+e^{2\theta d} & -e^{-\theta d} \\ 0 & 0 & \dots & -e^{-\theta d} & 1 \end{bmatrix}$$

Now, consider a more general setting in which measurements do not necessarily occur at evenly spaced intervals. Assume that $t_1 < t_2 < \cdots < t_{n-1} < t_n$. Then, the covariance matrix takes the form

$$\Psi = \frac{\sigma^2}{2\theta}\begin{bmatrix} 1 & e^{-\theta|t_2-t_1|} & \dots & e^{-\theta|t_{n-1}-t_1|} & e^{-\theta|t_n-t_1|} \\ e^{-\theta|t_2-t_1|} & 1 & \dots & e^{-\theta|t_{n-1}-t_2|} & e^{-\theta|t_n-t_2|} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-\theta|t_{n-1}-t_1|} & e^{-\theta|t_{n-1}-t_2|} & \dots & 1 & e^{-\theta|t_{n-1}-t_n|} \\ e^{-\theta|t_n-t_1|} & e^{-\theta|t_n-t_2|} & \dots & e^{-\theta|t_n-t_{n-1}|} & 1 \end{bmatrix}$$

and the precision matrix can be expressed as

$$\Omega = \frac{2\theta}{\sigma^2}\begin{bmatrix} \frac{1}{1-e^{-2\theta|t_2-t_1|}} & -\frac{e^{-\theta|t_2-t_1|}}{1-e^{-2\theta|t_2-t_1|}} & \dots & 0 & 0 \\ -\frac{e^{-\theta|t_2-t_1|}}{1-e^{-2\theta|t_2-t_1|}} & \frac{1-e^{-2\theta|t_3-t_1|}}{(1-e^{-2\theta|t_2-t_1|})(1-e^{-2\theta|t_3-t_2|})} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1-e^{-2\theta|t_n-t_{n-2}|}}{(1-e^{-2\theta|t_2-t_1|})(1-e^{-2\theta|t_3-t_2|})} & -\frac{e^{-2\theta|t_n-t_{n-1}|}}{1-e^{-2\theta|t_n-t_{n-1}|}} \\ 0 & 0 & \dots & -\frac{e^{-2\theta|t_n-t_{n-1}}}{1-e^{-2\theta|t_n-t_{n-1}|}} & \frac{1}{1-e^{-2\theta|t_n-t_{n-1}|}} \end{bmatrix}$$

Next, we move from the one-dimensional case to the two-dimensional case. We start by re-arranging the terms in the definition of the cross-covariance function for the bivariate OU process.

$$
\begin{aligned}
Cov\{\eta(s),\eta(t)\} &= vec^{-1}\big\{(\theta\oplus\theta)^{-1}e^{s\wedge t(\theta\oplus\theta)-(\theta t)\oplus(\theta s)}vec(\sigma\sigma^\top)\big\} \\
&\overset{(a)}{=} vec^{-1}\big\{e^{s\wedge t(\theta\oplus\theta)}e^{-(\theta t)\oplus(\theta s)}(\theta\oplus\theta)^{-1}vec(\sigma\sigma^\top)\big\} \\
&= vec^{-1}\big\{\big[e^{s\wedge t\theta}\otimes e^{s\wedge t\theta}\big]\big[e^{-\theta t}\otimes e^{-\theta s}\big](\theta\oplus\theta)^{-1}vec(\sigma\sigma^\top)\big\} \\
&= vec^{-1}\big\{\big[e^{s\wedge t\theta}e^{-\theta t}\big]\otimes\big[e^{s\wedge t\theta}e^{-\theta s}\big](\theta\oplus\theta)^{-1}vec(\sigma\sigma^\top)\big\} \\
&= vec^{-1}\big\{\big[e^{-\theta(t-s\wedge t)}\big]\otimes\big[e^{-\theta(s-s\wedge t)}\big](\theta\oplus\theta)^{-1}vec(\sigma\sigma^\top)\big\} \\
&\overset{(b)}{=} vec^{-1}\big\{\big[e^{-\theta(t-s\wedge t)}\big]\otimes I(\theta\oplus\theta)^{-1}vec(\sigma\sigma^\top)\big\} \\
&= vec^{-1}\big\{(\theta\oplus\theta)^{-1}vec(\sigma\sigma^\top)\big\}e^{-\theta^\top|t-s|} \\
&:= V\cdot e^{-\theta^\top|t-s|}
\end{aligned}
$$

where equality (a) is because these terms commute and equality (b) holds when we assume that $min(s,t) = s$. We can make this assumption without loss of generality because the matrices are symmetric. When $min(s,t) = t$, $Cov\{\eta(s),\eta(t)\} = e^{-\theta|t-s|}V^\top = e^{-\theta|t-s|}V$. Then, the covariance matrix is given by

$$
\Psi = \begin{bmatrix}
V & Ve^{-\theta^\top|t_2-t_1|} & \dots & Ve^{-\theta^\top|t_{n-1}-t_1|} & Ve^{-\theta^\top|t_n-t_1|} \\
e^{-\theta|t_2-t_1|}V & V & \dots & Ve^{-\theta^\top|t_{n-1}-t_2|} & Ve^{-\theta^\top|t_n-t_2|} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
e^{-\theta|t_{n-1}-t_1|}V & e^{-\theta|t_{n-1}-t_2|}V & \dots & V & Ve^{-\theta^\top|t_{n-1}-t_n|} \\
e^{-\theta|t_n-t_1|}V & e^{-\theta|t_n-t_2|}V & \dots & e^{-\theta|t_n-t_{n-1}|}V & V
\end{bmatrix}
$$

By the definition of the OU process, we know that the precision matrix, $\Omega = \Psi^{-1}$, is block tri-diagonal. We start by solving for two blocks, Ω_{11} and Ω_{12}. We assume that $\Omega_{11} = A^{-1}$ and $\Omega_{12} = A^{-1}B$, based on the form of the precision matrix in the case of the univariate OU process. Based on patterns seen when multiplying the AR(1) precision and covariance matrices, we assume that, for the OU process, the first row of blocks in the precision matrix, $[\Omega_{11},\Omega_{12},0,\dots,0]$ times the second column of blocks in the covariance matrix, $[Ve^{-\theta^\top(t_2-t_1)},V,\dots]^\top$, is equal to 0. So,

$$
\begin{aligned}
0 &= \Omega_{11}Ve^{-\theta^\top(t_2-t_1)} + \Omega_{12}V \\
\implies 0 &= A^{-1}Ve^{-\theta^\top(t_2-t_1)} + A^{-1}BV \\
\implies 0 &= Ve^{-\theta^\top(t_2-t_1)} + BV \\
\implies BV &= -Ve^{-\theta^\top(t_2-t_1)} \\
\implies B &= -Ve^{-\theta^\top(t_2-t_1)}V^{-1}
\end{aligned}
$$

By similar logic, the first row of blocks in the precision matrix times the first column of

blocks in the covariance matrix is equal to the identity matrix. So,

$$\begin{aligned} I &= \Omega_{11}V + \Omega_{12}e^{-\theta(t_2-t_1)}V \\ \implies I &= A^{-1}V + A^{-1}Be^{-\theta(t_2-t_1)}V \\ \implies A &= V + Be^{-\theta(t_2-t_1)}V \end{aligned}$$

We know that $B = -Ve^{-\theta^\top(t_2-t_1)}V^{-1}$ so

$$\begin{aligned} A &= V - Ve^{-\theta^\top(t_2-t_1)}V^{-1}e^{-\theta(t_2-t_1)}V \\ \implies A^{-1} &= [V - Ve^{-\theta^\top(t_2-t_1)}V^{-1}e^{-\theta(t_2-t_1)}V]^{-1} \end{aligned}$$

Now we have

$$\begin{aligned} \Omega_{11} &= \left[V - Ve^{-\theta^\top(t_2-t_1)}V^{-1}e^{-\theta(t_2-t_1)}V\right]^{-1} \\ \Omega_{12} &= -\left[V - Ve^{-\theta^\top(t_2-t_1)}V^{-1}e^{-\theta(t_2-t_1)}V\right]^{-1}Ve^{-\theta^\top(t_2-t_1)}V^{-1} \end{aligned}$$

Continuing with this logic, we can check the first row of blocks in Ω against all other columns of Ψ and see that

$$\begin{aligned} 0 &= \Omega_{11}Ve^{-\theta^\top(t_k-t_1)} + \Omega_{12}Ve^{-\theta^\top(t_k-t_2)} \\ &= A^{-1}Ve^{-\theta^\top(t_k-t_1)} + A^{-1}BVe^{-\theta^\top(t_k-t_2)} \\ &= Ve^{-\theta^\top(t_k-t_1)} + BVe^{-\theta^\top(t_k-t_2)} \\ &= Ve^{-\theta^\top(t_k-t_1)} - Ve^{-\theta^\top(t_2-t_1)}V^{-1}Ve^{-\theta^\top(t_k-t_2)} \\ &= Ve^{-\theta^\top(t_k-t_1)} - Ve^{-\theta^\top(t_2-t_1)}e^{-\theta^\top(t_k-t_2)} \\ &= Ve^{-\theta^\top(t_k-t_1)} - Ve^{-\theta^\top(t_k-t_1)} \\ &= 0 \end{aligned}$$

Now we move to the second row of blocks in Ω. Because $\Omega = \Omega^\top$, we also know that $\Omega_{21} = \Omega_{12}^\top$. This symmetry means that we only need to derive Ω_{22} and Ω_{23}. Based on previous results, we have

$$\Omega_{23} = -\left[V - Ve^{-\theta^\top(t_3-t_2)}V^{-1}e^{-\theta(t_3-t_2)}V\right]^{-1}Ve^{-\theta^\top(t_3-t_2)}V^{-1}$$

Then we find the form of Ω_{22} by once again using the same logic to say that the second

row of blocks in Ω times the second column of blocks in Ψ will be equal to an identity matrix:

$$\begin{aligned}
I &= \Omega_{21} V e^{-\theta^\top (t_2 - t_1)} + \Omega_{22} V + \Omega_{23} e^{-\theta(t_3 - t_2)} V \\
\Rightarrow V^{-1} &= \Omega_{21} V e^{-\theta^\top (t_2 - t_1)} V^{-1} + \Omega_{22} + \Omega_{23} e^{-\theta(t_3 - t_2)} \\
\Rightarrow \Omega_{22} &= V^{-1} + V^{-1} e^{-\theta(t_2 - t_1)} V \left[V - V e^{-\theta^\top (t_2 - t_1)} V^{-1} e^{-\theta(t_2 - t_1)} V \right]^{-1^\top} V e^{-\theta^\top (t_2 - t_1)} V^{-1} \\
&\quad + \left[V - V e^{-\theta^\top (t_3 - t_2)} V^{-1} e^{-\theta(t_3 - t_2)} V \right]^{-1} V e^{-\theta^\top (t_3 - t_2)} V^{-1} e^{-\theta(t_3 - t_2)}
\end{aligned}$$

The final terms are then given by:

$$\begin{aligned}
I &= \Omega_{n,n-1} V e^{-\theta^\top (t_n - t_{n-1})} + \Omega_{nn} V \\
\Rightarrow I &= -V^{-1} e^{-\theta(t_n - t_{n-1})} V \left[V - V e^{\theta^\top (t_n - t_{n-1})} V^{-1} e^{-\theta(t_n - t_{n-1})} V \right]^{-1} V e^{-\theta^\top (t_n - t_{n-1})} + \Omega_{nn} V \\
\Rightarrow \Omega_{nn} &= V^{-1} + V^{-1} e^{-\theta(t_n - t_{n-1})} V \left[V - V e^{-\theta^\top (t_n - t_{n-1})} V^{-1} e^{-\theta(t_n - t_{n-1})} V \right]^{-1} V e^{-\theta^\top (t_n - t_{n-1})} V^{-1}
\end{aligned}$$

Thus, the precision matrix Ω is block tri-diagonal with the following entries (indexed by j) for $1 < j < n$:

$$\begin{aligned}
V &:= vec^{-1}\big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\big\} \\
\Omega_{11} &= \big[V - V e^{-\theta^\top (t_2 - t_1)} V^{-1} e^{-\theta(t_2 - t_1)} V\big]^{-1} \\
\Omega_{j,j+1} &= \Omega_{j+1,j}^\top = -\big[V - V e^{-\theta^\top (t_{j+1} - t_j)} V^{-1} e^{-\theta(t_{j+1} - t_j)} V\big]^{-1} V e^{-\theta^\top (t_{j+1} - t_j)} V^{-1} \\
\Omega_{jj} &= V^{-1} + V^{-1} e^{-\theta(t_j - t_{j-1})} V \big[V - V e^{-\theta^\top (t_j - t_{j-1})} V^{-1} e^{-\theta(t_j - t_{j-1})} V\big]^{-1} V e^{-\theta^\top (t_j - t_{j-1})} V^{-1} \\
&\quad + \big[V - V e^{-\theta^\top (t_{j+1} - t_j)} V^{-1} e^{-\theta(t_{j+1} - t_j)} V\big]^{-1} V e^{-\theta^\top (t_{j+1} - t_j)} V^{-1} e^{-\theta(t_{j+1} - t_j)} \\
\Omega_{nn} &= V^{-1} + V^{-1} e^{-\theta(t_n - t_{n-1})} V \big[V - V e^{-\theta^\top (t_n - t_{n-1})} V^{-1} e^{-\theta(t_n - t_{n-1})} V\big]^{-1} V e^{-\theta^\top (t_n - t_{n-1})} V^{-1}
\end{aligned}$$

A.4 Identifiability Constraint: Re-Scaling the OU Process

Let (θ^*, σ^*) be a pair of OU process parameters satisfying the identifiability constraint that the stationary variance of the OU process is equal to 1; that is, $diag\{\Psi(\theta^*, \sigma^*)\} = 1$, where Ψ is the covariance matrix of the OU process. We show that we can always find a pair of (θ^*, σ^*) that defines a valid mean-reverting OU process with stationary variance of 1 that has the same correlation structure as the original unconstrained OU process defined by (θ, σ). As an example, consider the stochastic differential equation definition of the bivariate

OU process. For an arbitrary mean-reverting OU process, $\eta(t)$,

$$d\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix} = -\begin{bmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\end{bmatrix}\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix}dt + \begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}d\begin{bmatrix}W_1(t)\\ W_2(t)\end{bmatrix}$$

We could equivalently define this OU process $\eta(t)$ as

$$\begin{aligned} d\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix} &= -\begin{bmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\end{bmatrix}\begin{bmatrix}c_1 & 0\\ 0 & c_2\end{bmatrix}\begin{bmatrix}1/c_1 & 0\\ 0 & 1/c_2\end{bmatrix}\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix}dt + \begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}d\begin{bmatrix}W_1(t)\\ W_2(t)\end{bmatrix}\\ &= -\begin{bmatrix}c_1\theta_{11} & c_2\theta_{12}\\ c_1\theta_{21} & c_2\theta_{22}\end{bmatrix}\begin{bmatrix}\frac{1}{c_1}\eta_1(t)\\ \frac{1}{c_2}\eta_2(t)\end{bmatrix}dt + \begin{bmatrix}\sigma_{11} & 0\\ 0 & \sigma_{22}\end{bmatrix}d\begin{bmatrix}W_1(t)\\ W_2(t)\end{bmatrix} \end{aligned}$$

Let $\eta^*(t)$ be a scaled version of η where

$$\begin{bmatrix}\eta_1^*(t)\\ \eta_2^*(t)\end{bmatrix} = \begin{bmatrix}\frac{1}{c_1}\eta_1(t)\\ \frac{1}{c_2}\eta_2(t)\end{bmatrix}$$

and

$$\begin{bmatrix}\theta_{11}^* & \theta_{12}^*\\ \theta_{21}^* & \theta_{22}^*\end{bmatrix} = \begin{bmatrix}c_1\theta_{11} & c_2\theta_{12}\\ c_1\theta_{21} & c_2\theta_{22}\end{bmatrix}$$

and assume that $\eta^*(t)$ has a stationary variance equal to 1. Then,

$$\begin{aligned} d\eta^*(t) &= -\begin{bmatrix}\theta_{11}^* & \theta_{21}^*\\ \theta_{12}^* & \theta_{22}^*\end{bmatrix}\begin{bmatrix}\eta_1^*(t)\\ \eta_2^*(t)\end{bmatrix}dt + \begin{bmatrix}\sigma_{11}^* & 0\\ 0 & \sigma_{22}^*\end{bmatrix}d\begin{bmatrix}W_1(t)\\ W_2(t)\end{bmatrix}\\ &= -\begin{bmatrix}\theta_{11}^* & \theta_{21}^*\\ \theta_{12}^* & \theta_{22}^*\end{bmatrix}\begin{bmatrix}c_1 & 0\\ 0 & c_2\end{bmatrix}\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix}dt + \begin{bmatrix}\sigma_{11}^* & 0\\ 0 & \sigma_{22}^*\end{bmatrix}d\begin{bmatrix}W_1(t)\\ W_2(t)\end{bmatrix}\\ &= -\begin{bmatrix}c_1\theta_{11}^* & c_2\theta_{21}^*\\ c_1\theta_{12}^* & c_2\theta_{22}^*\end{bmatrix}\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix}dt + \begin{bmatrix}\sigma_{11}^* & 0\\ 0 & \sigma_{22}^*\end{bmatrix}d\begin{bmatrix}W_1(t)\\ W_2(t)\end{bmatrix} \end{aligned}$$

Looking back at the original OU process $\eta(t)$,

$$\begin{aligned} d\eta(t) &= d\begin{bmatrix}\frac{1}{c_1} & 0\\ 0 & \frac{1}{c_2}\end{bmatrix}\eta^*(t)\\ &= -\begin{bmatrix}\frac{1}{c_1} & 0\\ 0 & \frac{1}{c_2}\end{bmatrix}\begin{bmatrix}c_1\theta_{11}^* & c_2\theta_{12}^*\\ c_1\theta_{21}^* & c_2\theta_{22}^*\end{bmatrix}\eta(t)dt + \begin{bmatrix}\frac{1}{c_1}\sigma_{11}^* & 0\\ 0 & \frac{1}{c_2}\sigma_{22}^*\end{bmatrix}dW(t)\\ &= -\begin{bmatrix}\frac{c_1}{c_1}\theta_{11}^* & \frac{c_2}{c_1}\theta_{12}^*\\ \frac{c_1}{c_2}\theta_{21}^* & \frac{c_2}{c_2}\theta_{22}^*\end{bmatrix}\eta(t)dt + \begin{bmatrix}\frac{1}{c_1}\sigma_{11}^* & 0\\ 0 & \frac{1}{c_2}\sigma_{22}^*\end{bmatrix}dW(t) \end{aligned}$$

Finally, we see that the parameters for $\eta(t)$ can easily be re-scaled to satisfy our identifiability assumption:

$$\begin{bmatrix} \theta_{11} & \frac{c_1}{c_2}\theta_{12} \\ \frac{c_2}{c_1}\theta_{21} & \theta_{22} \end{bmatrix} = \begin{bmatrix} \theta^*_{11} & \theta^*_{12} \\ \theta^*_{21} & \theta^*_{22} \end{bmatrix}$$

and

$$\begin{bmatrix} c_1\sigma_{11} & 0 \\ 0 & c_2\sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma^*_{11} & 0 \\ 0 & \sigma^*_{22} \end{bmatrix}$$

Thus, we have shown that for a mean-reverting bivariate OU process defined by θ and σ with covariance matrix $\Psi(\theta, \sigma)$ and correlation matrix $\Psi^*(\theta, \sigma)$, we can re-scale this OU process to have stationary variance equal to 1 by scaling θ_{12}, θ_{21} and σ_{11}, σ_{22} by a pair of positive scalar constants, (c_1, c_2). This proof can easily be extended to higher dimensional OU processes.

A.5 Derivation of the Analytic Gradients for the Measurement Submodel

We have previously defined the log-likelihood for a single subject i as

$$\ell_i = -\frac{1}{2} log|\Sigma^*_i| + Y_i^\top \Sigma^{*-1}_i Y_i \tag{A.7}$$

where we ignore the constant terms and

$$\Sigma^*_i = (I_{n_i} \otimes \Lambda)\Psi_i(I_{n_i} \otimes \Lambda)^\top + J_{n_i} \otimes \Sigma_u + I_{n_i} \otimes \Sigma_\epsilon \tag{A.8}$$

Gradient with Respect to the Loadings: We first take the derivative of ℓ_i with respect to the elements of the loadings matrix Λ, λ_k, $k = 1, ..., p \times K$. The first element of the loadings matrix is parameterized on the log scale in order to restrict this element to positive values for identifiability purposes and so the gradient of this element looks slightly different. For $k > 1$, we have

$$\frac{\partial \ell_i}{\partial \lambda_k} = -\frac{1}{2}\left[tr\left\{ \Sigma^{*-1}_i \frac{\partial \Sigma^*_i}{\partial \lambda_k} \right\} - Y_i^\top \Sigma^{*-1}_i \frac{\partial \Sigma^*_i}{\partial \lambda_k} \Sigma^{*-1}_i Y_i \right] \tag{A.9}$$

where

$$\frac{\partial \Sigma^*_i}{\partial \lambda_k} = (I_{n_i} \otimes \Lambda)\Psi_i(I_{n_i} \otimes J^{k^\top}) + (I_{n_i} \otimes J^k)\Psi_i(I_{n_i} \otimes \Lambda^\top) \tag{A.10}$$

We use J^k as an indicator matrix that has the same dimension as Λ but contains all zeros

except for a single 1 indicating the location of element λ_k. For $k = 1$, we apply the chain rule and have

$$\frac{\partial \ell_i}{\partial log(\lambda_k)} = \frac{\partial \ell_i}{\partial \lambda_k}\left[\frac{\partial log(\lambda_k)}{\partial \lambda_k}\right]^{-1} = \frac{\partial \ell_i}{\partial \lambda_k}\lambda_k \tag{A.11}$$

Gradient with Respect to the Random Effects: Next, we take the gradient of ℓ_i with respect to the elements of R_u where R_u comes from the Cholesky decomposition of the random effects covariance matrix, $\Sigma_u = R_u^\top R_u$. For $p, q = 1, ..., K$ and $p \neq q$,

$$\frac{\partial \Sigma_i^*}{\partial r_{pq}} = J_{n_i} \otimes (J^{k^\top} R_u + R_u^\top J^k) \tag{A.12}$$

$$\frac{\partial \ell_i}{\partial r_{pq}} = -\frac{1}{2}\left[tr\left\{\Sigma_i^{*-1}\frac{\partial \Sigma_i^*}{\partial r_{pq}}\right\} + Y_i^\top \Sigma_i^{*-1}\frac{\partial \Sigma_i^*}{\partial r_{pq}}\Sigma_i^{*-1}Y_i\right] \tag{A.13}$$

where again J^k is an indicator matrix of the same dimensions as Σ_u.

For $p, q = 1, ..., K$ and $p = q$,

$$\frac{\partial \ell_i}{\partial log(r_{u_{pp}})} = \frac{\partial \ell_i}{\partial r_{u_{pp}}}\left[\frac{\partial log(r_{u_{pp}})}{\partial r_{u_{pp}}}\right]^{-1} = \frac{\partial \ell_i}{\partial r_{u_{pp}}} r_{u_{pp}} \tag{A.14}$$

Note that if we assume only random intercepts (i.e., a diagonal covariance matrix) then we can avoid the Cholesky decomposition by estimating σ_u on the log scale. In this case, the gradient simplifies to the form given below for the measurement error.

Gradient with Respect to the Measurement Error: Finally, we take the gradient of ℓ_i with respect to the elements of the measurement error covariance matrix, Σ_ϵ. For $k = 1, ..., K$, we have

$$\frac{\partial \Sigma_i^*}{\partial \sigma_{\epsilon_k}} = I_{n_i} \otimes 2\sigma_{\epsilon_k} J^k \tag{A.15}$$

$$\frac{\partial \ell_i}{\partial \sigma_{\epsilon_k}} = -\frac{1}{2}\left[tr\left\{\Sigma_i^{*-1}\frac{\partial \Sigma_i^*}{\partial \sigma_{\epsilon_k}}\right\} - Y_i^\top \Sigma_i^{*-1}\frac{\partial \Sigma_i^*}{\partial \sigma_{\epsilon_k}}\Sigma_i^{*-1}Y_i\right] \tag{A.16}$$

$$\frac{\partial \ell_i}{\partial log(\sigma_{\epsilon_k})} = \frac{\partial \ell_i}{\partial \sigma_{\epsilon_k}}\left[\frac{log(\sigma_{\epsilon_k})}{\partial \sigma_{\epsilon_k}}\right]^{-1} = \frac{\partial \ell_i}{\partial \sigma_{\epsilon_k}}\sigma_{\epsilon_k} \tag{A.17}$$

where J^k is an indicator matrix of the same dimensions as Σ_ϵ.

A.6 Parameterization of the Log-Likelihood for Standard Error Estimation

To make our OUF model identifiable, we impose a constraint on the scale of the OU process by forcing the stationary variance equal to 1 via a set of p positive scalar constants. These constants are functions of OU parameters θ and σ.

When the log-likelihood is allowed to vary as a function all parameters, rather than just a single block of parameters as in our block coordinate descent algorithm, our model is no longer identifiable. To estimate standard errors, we take advantage of the fact that under the identifiability constraint, σ can be written as a function of θ, as shown here:

Recall that the stationary variance of the OU process is $V := vec^{-1}\Big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^\top\}\Big\}$. Assuming a bivariate OU process, under the identifiability constraint, V takes the form $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ where the off-diagonal element ρ is the correlation. Then,

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = vec^{-1}\Big\{(\theta \oplus \theta)^{-1} vec\{\sigma\sigma^{-1}\}\Big\} \Longrightarrow \begin{bmatrix} 1 \\ \rho \\ \rho \\ 1 \end{bmatrix} = (\theta \oplus \theta)^{-1} \begin{bmatrix} \sigma_{11}^2 \\ 0 \\ 0 \\ \sigma_{22}^2 \end{bmatrix}.$$

Letting

$$(\theta \oplus \theta)^{-1} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix},$$

where each element x_{ij} is some function of the elements of θ, we can solve for σ in the (θ, σ) pair that satisfies the identifiability constraint via

$$1 = x_{11}\sigma_{11}^2 + x_{14}\sigma_{22}^2$$
$$1 = x_{41}\sigma_{11}^2 + x_{44}\sigma_{22}^2$$

By constraining σ to be a function of θ, we take an alternative approach to identification and no longer require use of the scaling constants here.

A.7 Choice of True OU Process in Simulation Study

In the simulation study described in the main text (Section 2.4.1-2.4.2), we generate data in three different settings in which the true OU process has varying degrees of auto-correlation. We present the true OU process parameters here:

Setting 1:

$$\theta = \begin{bmatrix} 1 & 0.6 \\ 4 & 5 \end{bmatrix} \text{ and } \sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

Setting 2:

$$\theta = \begin{bmatrix} 1.0 & 0.4 \\ 1.8 & 3.0 \end{bmatrix} \text{ and } \sigma = \begin{bmatrix} 1.25 & 0 \\ 0 & 2.00 \end{bmatrix}$$

Setting 3:

$$\theta = \begin{bmatrix} 1 & 0.5 \\ 2 & 5 \end{bmatrix} \text{ and } \sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

In the simulation study assessing use of AIC and BIC to select the correct number of latent factors in a model (described in the main text in Section 2.4.3-2.4.4), the true parameters were set to the values listed below. The true values used for Σ_u and Σ_ϵ were the same as in the original simulation study (see Section 2.4.1 in the main text).

One-Factor Model:

$$\Lambda = \begin{bmatrix} 1.2 \\ 1.8 \\ -0.4 \\ 2 \end{bmatrix}, \quad \theta = 0.8, \quad \sigma = 1$$

Two-Factor Model with Low Signal:

$$\Lambda = \begin{bmatrix} 1.2 & 0 \\ 1.8 & 0 \\ 0 & -0.4 \\ 0 & 2 \end{bmatrix}, \quad \theta = \begin{bmatrix} 2 & 0.5 \\ 0.4 & 4 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

Two-Factor Model with High Signal:

$$\Lambda = \begin{bmatrix} 1.2 & 0 \\ 1.8 & 0 \\ 0 & -0.4 \\ 0 & 2 \end{bmatrix}, \quad \theta = \begin{bmatrix} 1 & 1.5 \\ 2 & 5 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

Three-Factor Model with Low Signal:

$$\Lambda = \begin{bmatrix} 1.2 & 0 & 0 \\ 1.8 & 0 & 0 \\ 0 & -0.4 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \theta = \begin{bmatrix} 2 & 0.2 & 0.4 \\ 0.8 & 1.1 & 0.5 \\ 0.7 & 0.5 & 1.2 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 1.2 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.4 \end{bmatrix}$$

Three-Factor Model with High Signal:

$$\Lambda = \begin{bmatrix} 1.2 & 0 & 0 \\ 1.8 & 0 & 0 \\ 0 & -0.4 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \theta = \begin{bmatrix} 1 & 0.4 & 0.6 \\ 1.8 & 3 & 0.9 \\ 0.9 & 1 & 1.2 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 1.2 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.4 \end{bmatrix}$$

A.8 Discussion of Numerical Issues in Simulation Results

Simulation Study: Bias and Variance The estimation algorithm failed to converge due to numerical issues when applied to a few of the simulated datasets generated in the simulation study described in Section 2.4.1-2.4.2. The failures were caused by a singular V matrix at the start of the first block update of the structural submodel parameters. Slightly altering the values at which the OU process parameters were initialized resolved this issue. Point estimates were ultimately calculate for all 1000 simulated datasets in each setting. In Setting 3, an invalid variance for the measurement submodel parameter $\sigma^2_{\epsilon_4}$ was estimated

from one dataset. In this instance, the variance estimated for this parameter was negative. We attribute this issue to the numerical approximation used to calculate the Hessian when applied to these this dataset of size $N = 200$. We anticipate that a larger dataset would improve the approximation of the numerical Hessian but chose to simulate a dataset of this size in order to assess model performance in a realistic setting similar to that encountered in the motivating data application. In practical application, if a negative variance were to be estimated, it could be rounded to 0. In the results presented in the main text, we ignore the variance estimate for this one $\sigma^2_{\epsilon_4}$.

Simulation Study: Model Selection In our simulation studies, we aimed to assess simulated datasets with sample sizes similar to that of our motivating dataset. For datasets of fixed size ($N = 200$ subjects), we found that convergence speeds decrease and estimation becomes more difficult as the number of factors in the model increases. We found that point estimates of the diagonal elements of θ_{OU} hit the lower bound of 1×10^{-4} less than 1% of the time. To improve convergence, we slightly altered the set of default parameter values considered during the initialization steps of the block-wise estimation algorithm for a subset of datasets. However, when assessing AIC and BIC as model selection criteria (see Section 2.4.3-2.4.4), we very occasionally encountered numerical issues and so failed to calculate parameter estimates for a subset of models applied to the simulated datasets. The results reported in the main paper correspond to a comparison of AIC and BIC across datasets for which the algorithm used to fit all three models (the one-factor, two-factor, and three-factor models) either converged or reached the maximum number of iterations prior to convergence. We assessed whether or not including results in which the maximum number of iterations was reached prior to convergence impacted our model selection results and found no substantial changes. Table A.1 shows the equivalent version of Table 2.1 if only results from datasets that had converged were shown.

In Table A.2, we summarize the number (out of 100) of datasets (in each setting) for which the algorithm converged (using $\delta = 1 \times 10^{-6}$) or reached the maximum number of iterations prior to convergence. When this total number does not add up to 100, the remaining datasets correspond to situations in which the algorithm failed due to numerical issues (e.g., current OU parameter estimates corresponded to a singular stationary covariance matrix).

After loosening the convergence criteria across the block-wise iterations, we did not find substantially different results when evaluating AIC and BIC as model selection criteria when compared to results under the original convergence criteria. For example, if we categorized convergence using $\delta \leq 1 \times 10^{-3}$, rather than only the original $\delta = 1 \times 10^{-6}$, the algorithm would have converged when fitting almost every model to almost every dataset (see Table

True Model		# Factors in Fitted Model with Best AIC			# Factors in Fitted Model with Best BIC		
# Factors	Signal	1	2	3	1	2	3
1	-	**99**	0	1	**100**	0	0
2	Low	0	**93**	7	4	**96**	0
2	High	0	**100**	0	0	**100**	0
3	Low	0	0	**100**	0	8	**92**
3	High	0	0	**100**	0	0	**100**

Table A.1: For datasets generated under each true model, we summarize the percent of times that the model-selection metric chose the fitted model with the indicated number of factors. The settings in which the fitted model has the same number of factors as the true data-generating model are emphasized with bold orange text. These results are presented for datasets on which the algorithm converged prior to reaching the maximum number of iterations (200) for all three models.

		# Factors in Fitted Model					
True Model		1		2		3	
# Factors	Signal	convergence	iteration limit	convergence	iteration limit	convergence	iteration limit
1	-	100	0	100	0	79	20
2	Low	100	0	100	0	96	4
2	High	100	0	100	0	98	2
3	Low	100	0	99	1	100	0
3	High	99	0	100	0	99	1

Table A.2: For datasets generated under each true model, we summarize the number of datasets (out of 100) on which the algorithm converged or reached the maximum number of block-wise iterations prior to convergence (when $\delta = 1 \times 10^{-6}$). For totals that do not sum to 100, the remaining cases correspond to instances in which the algorithm failed due to numerical issues prior to converging or reaching the maximum number of block-wise iterations (200).

True Model		# Factors in Fitted Model					
		1		2		3	
# Factors	Signal	convergence	iteration limit	convergence	iteration limit	convergence	iteration limit
1	-	100	0	100	0	94	5
2	Low	100	0	100	0	100	0
2	High	100	0	100	0	100	0
3	Low	100	0	99	1	100	0
3	High	99	0	100	0	100	0

Table A.3: For datasets generated under each true model, we summarize the number of datasets (out of 100) on which the algorithm converged or reached the maximum number of block-wise iterations prior to convergence (when $\delta \leq 1 \times 10^{-3}$). For totals that do not sum to 100, the remaining cases correspond to instances in which the algorithm failed due to numerical issues prior to converging or reaching the maximum number of block-wise iterations (200).

True Model		# Factors in Fitted Model with Best AIC			# Factors in Fitted Model with Best BIC		
# Factors	Signal	1	2	3	1	2	3
1	-	**99**	0	1	**100**	0	0
2	Low	0	**93**	7	4	**96**	0
2	High	0	**100**	0	0	**100**	0
3	Low	0	0	**100**	0	8	**92**
3	High	0	0	**100**	0	0	**100**

Table A.4: For datasets generated under each true model, we summarize the percent of times that the model-selection metric chose the fitted model with the indicated number of factors. The settings in which the fitted model has the same number of factors as the true data-generating model are emphasized with bold orange text. These results are presented for datasets on which the algorithm converged (using $\delta \leq 1 \times 10^{-3}$) prior to reaching the maximum number of iterations (200) for all three models.

A.3) but the model selection results would not have changed (see Table A.4).

We expect that increasing the size of the simulated dataset would increase the rate at which we successfully fit models with more factors.

A.9 Application to mHealth Emotion Data

A.9.1 OUF Model with One Factor

In this model, we assume that a single latent factor generates all observed emotions of happy, joyful, enthusiastic, active, calm, determined, grateful, proud, attentive, sad, scared, disgusted, angry, ashamed, guilty, irritable, lonely, and nervous. We plot the point estimates from this model in Figure A.1. Using these estimated parameters, we calculate the auto-correlation half-life of this latent factor as approximately 27 days. This model has a total of

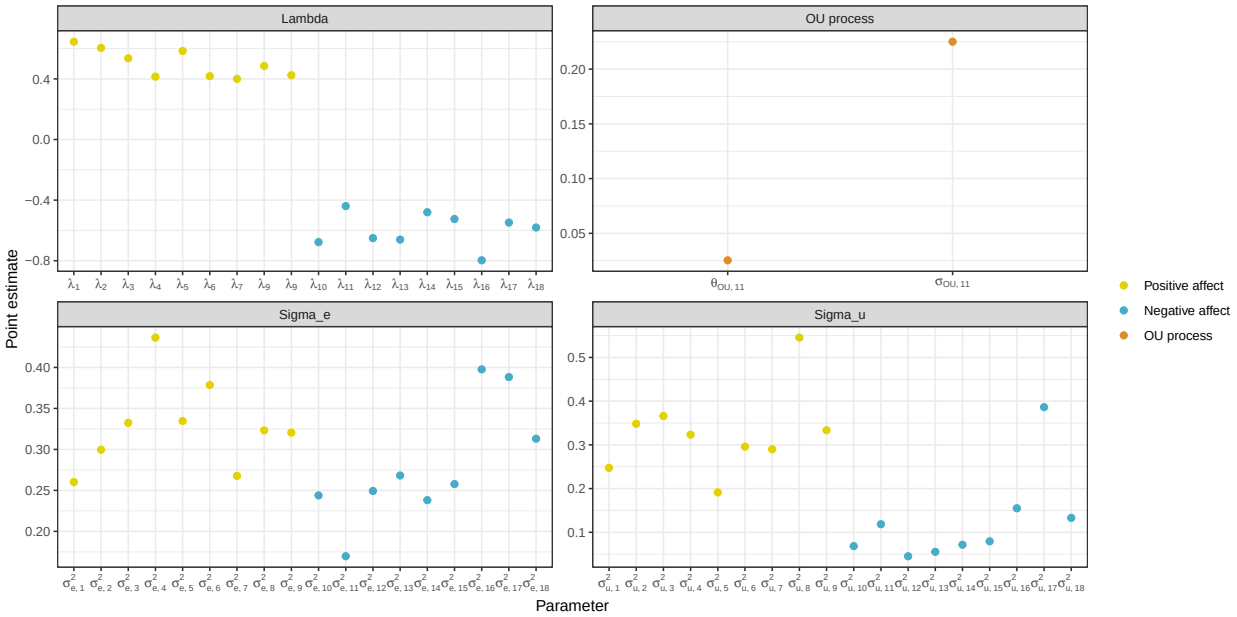


Figure A.1: Point estimates for each of the parameter matrices in our one-factor OUF model. Because we assume structural zeros in the loadings matrix are known, each emotion has only a single loading. Parameter subscripts 1-18 correspond to the emotions as follows: 1 = happy, 2 = joyful, 3 = enthusiastic, 4 = active, 5 = calm, 6 = determined, 7 = grateful, 8 = proud, 9 = attentive, 10 = sad, 11 = scared, 12 = disgusted, 13 = angry, 14 = ashamed, 15 = guilty, 16 = irritable, 17 = lonely, 18 = nervous.

56 free parameters, along with one constraint, which we use when calculating AIC and BIC.

A.9.2 OUF Model with Two Factors

In this model, we assume that two latent factors generate the observed emotions. The latent factors represent positive affect (which underlies happy, joyful, enthusiastic, active, calm, determined, grateful, proud, and attentive) and negative affect (which underlies sad, scared, disgusted, angry, ashamed, guilty, irritable, lonely, and nervous). Results from this fitted model are available in Section 2.5. This model has a total of 60 free parameters, along with two constraints, which we use when calculating AIC and BIC.

A.9.3 OUF Model with Three Factors

We assume that three latent emotional states underlie the emotions observed during this study. The emotions load on to the latent factors as follows:

1. enthusiastic, proud, active, calm, determined, attentive, grateful [η_1 = high arousal positive affect]

Positive affect items	Arousal	Citation
calm	no-to-low	McManus (2019), Gilbert (2008), Remington (2000)
grateful	high	Reisenzein (1994)
proud	high	McManus (2019)
happy	no-to-low	Remington (2000)
joyful	no-to-low	Remington (2000)
enthusiastic	high	McManus (2019), Gilbert (2008), Remington (2000)
active	high	McManus (2019), Gilbert (2008), Remington (2000)
determined	high	McManus (2019)
attentive	high	McManus (2019)

Table A.5: Behavioral science literature supporting the division of the positive emotions into two groups representing no-to-low arousal positive affect and high arousal positive affect.

2. calm, happy, joyful [η_2 = no-to-low arousal positive affect]

3. sad, scared, disgusted, angry, ashamed, guilty, irritable, lonely, nervous [η_3 = negative affect]

We use behavioral science literature and theory—namely the circumplex model of emotion—to inform the division of the positive affect emotions into groups representing high arousal positive affect and no-to-low arousal positive affect [85, 86, 32, 59]. Literature supporting the placement of each positive affect emotion is summarized in Table A.5. Happy and joyful are also commonly placed midway between high and low arousal in the circumplex model of emotion [86] and so we chose to assess the fit of the OUF model when these emotion items load onto the latent factor representing no-to-low arousal positive affect. This model converged after 211 block iterations and we present point estimates in Figure A.2. This model has a total of 66 free parameters, along with three constraints, which we use when calculating AIC and BIC.

A.10 Estimation Algorithm

A.10.1 Parameter Initialization

Due to the complexity of our model, our estimation algorithm is sensitive to the choice of initial estimates. Here we present an approach to estimating reasonable starting values based on simple existing models prior to maximizing the entire likelihood.

1. To initialize the **measurement submodel parameters**, fit a standard cross-sectional factor model to the data collapsed across time (do not include a random intercept but do assume that the positions of the non-zero loadings are known).

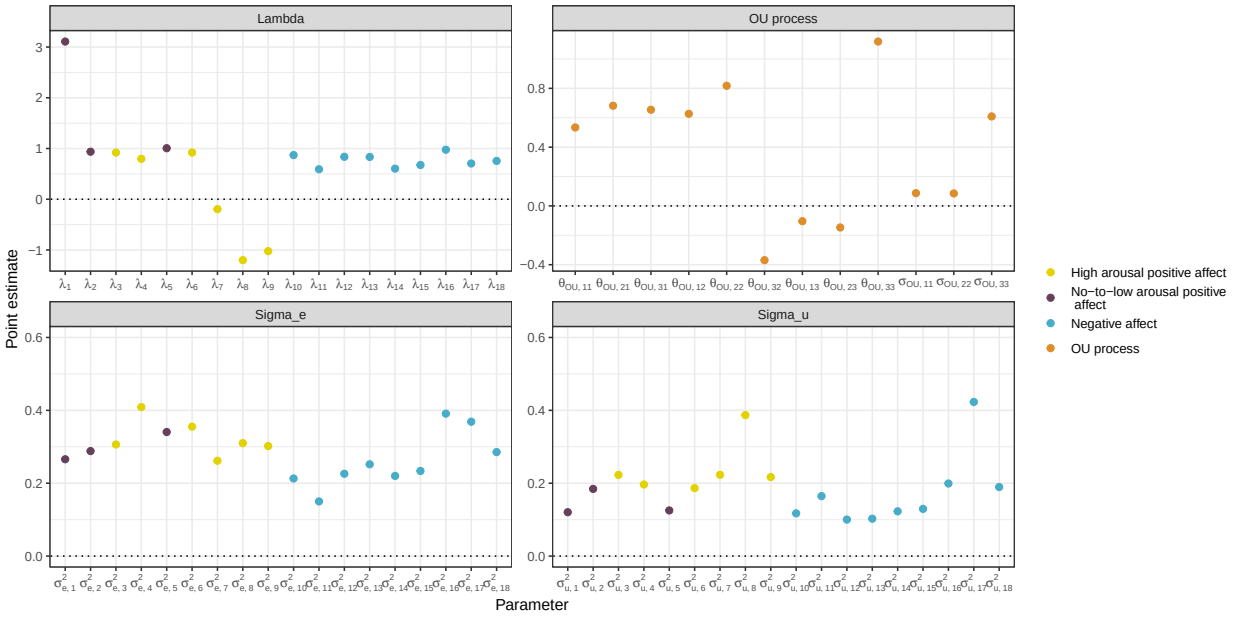


Figure A.2: Point estimates for each of the parameter matrices in our three-factor OUF model. Because we assume structural zeros in the loadings matrix are known, each emotion has only a single loading. Parameter subscripts 1-18 correspond to the emotions as follows: 1 = happy, 2 = joyful, 3 = enthusiastic, 4 = active, 5 = calm, 6 = determined, 7 = grateful, 8 = proud, 9 = attentive, 10 = sad, 11 = scared, 12 = disgusted, 13 = angry, 14 = ashamed, 15 = guilty, 16 = irritable, 17 = lonely, 18 = nervous.

2. Using this fitted factor model, estimate the factor scores (predicted values for η_1 and η_2).

3. Fit four separate linear mixed effects models—one for each of the observed outcomes, $Y_1, ..., Y_K$—including the factor scores as fixed effects and a random intercept for subject. We do not include a fixed effect intercept in these models. For outcome $k = 1, ..., K$, subject $i = 1, ..., N$, and measurement occasion $j = 1, ..., n_i$, the mixed model is given by

$$Y_{kij} = \lambda_k \eta_i(t_j) + u_{k0i} + \epsilon_{kij}$$

where $u_{k0i} \sim N(0, \sigma^2_{u_k})$ and $\epsilon_{kij} \sim N(0, \sigma^2_{\epsilon_k})$.

4. From each of these K mixed models, extract estimates of the coefficient for the fixed effect, the variance for the random intercept, and the residual variance. Use the coefficients of the fixed effects to initialize the non-zero elements of Λ and the variance estimates to initialize the diagonal components of Σ_u and Σ_ϵ. In some cases, the estimated variances were very small, so a lower limit of 0.1 was set for the initial parameter values to avoid extremely negative estimates after logging. We also set the same lower bound for initial values of the elements in the loadings matrix.

5. To initialize the **structural submodel parameters**, we add a term for white noise to the OU process likelihood. This noise term will absorb some of the extra variability in the predicted factor scores and allow for more stable estimation. Let Γ_i be white noise, then $\eta_i \sim N(0, \Psi_i + \Gamma_i)$ where Γ_i is a diagonal matrix (of the same dimension as OU covariance matrix Ψ_i) with constant but unknown diagonal γ. We then maximize this likelihood and use the estimated OU process parameter values as initial values, restricting the maximum initial values of the diagonals of θ_{OU} to be less than 7. This maximum helps deal with instability in the initial estimate of θ.

A.10.2 Maximization of the Marginal Log-Likelihood

To maximize the log-likelihood, we use quasi-Newton optimizers as implemented in the `stats` package in `R` [81]. To prevent the parameter estimates from diverging to infinite values, we set the maximum allowed step size to 10.

Using the initial parameter values estimated via the approach described in the previous section, we iteratively update measurement and structural submodel parameter estimates in blocks:

1. Initialize estimates: $\Lambda^{(0)}, \Sigma_u^{(0)}, \Sigma_\epsilon^{(0)}, \theta^{(0)}, \sigma^{(0)}$. Measurement submodel parameters are

always initialized empirically; for structural submodel parameters, two sets of initial estimates are considered—an empirical set of values estimated as described above and a default set of values that are based on a reasonable guess. The set of values that corresponds to the higher log-likelihood is used.

2. Set $r = 1$ and $\delta = 0$. While $r \leq 200$ and $\delta = 0$,

   (a) Update block of **measurement submodel parameters**:

   $$\Lambda^{(r)}, \Sigma_u^{(r)}, \Sigma_\epsilon^{(r)} = \underset{\Lambda, \Sigma_u, \Sigma_\epsilon}{argmax} \big\{ logL(Y|\theta^{(r-1)}, \sigma^{(r-1)}) \big\}.$$

   We solve this iteratively using `nlm` [81] and analytic gradients with convergence criteria set to `gradtol` $= max(1 \times 10^{-4}/10^r, 1 \times 10^{-8})$ and `steptol` $= max(1 \times 10^{-4}/10^r, 1 \times 10^{-8})$. `gradtol` is the tolerance for the scaled gradient and `steptol` is the tolerance for parameter estimates across iterations. We model the first element of the loadings matrix and the variance parameters on the log scale, since all of these estimates are required to be positive.

   (b) Update block of **structural submodel parameters**:

   $$\theta^{(r)}, \sigma^{(r)} = \underset{\theta, \sigma}{argmax} \big\{ logL(Y|\Lambda^{(r)}, \Sigma_u^{(r)}, \Sigma_\epsilon^{(r)}) \big\}.$$

   We solve this iteratively using `nlminb` and numeric approximations to the gradients. For estimates of θ, the diagonal elements must be positive and the matrix must have eigenvalues with positive real parts. The eigenvalue constraint is implemented by adding a negative penalty term to the likelihood for proposed values of θ that do not satisfy this constraint. The diagonal element of σ are estimated on the log scale, since they are required to be positive.

   (c) Check for block-wise convergence: Let Θ be a vector containing all elements of Λ, Σ_u, Σ_ϵ, θ, and σ. Then, calculate

   $$\delta = \max \Big\{ I\big\{ |\Theta^{(r)} - \Theta^{(r-1)}| / \Theta^{(r)} < 10^{-6} \big\}, I\big\{ logL(\Theta^{(r)}|Y) - logL(\Theta^{(r-1)}|Y) < 10^{-6} \big\} \Big\}$$

   where all operations on Θ are element-wise.

   (d) Rescale OU process parameters so stationary variance is equal to 1 using Equation 5 in the main paper.

   (e) Update r: $r = r + 1$

3. Estimate standard errors using a numerical approximation to the Hessians of the joint negative log-likelihood for $\Lambda^{(r)}, \Sigma_u^{(r)}, \Sigma_\epsilon^{(r)}, \theta^{(r)}$ at the current parameter values. Rather than rescaling the OU parameters so the stationary variance is equal to 1, we assume that σ is a function of θ. See Appendix A.6 for further description of this function. The numeric approximation to the Hessians is carried out using the `optimHess` function in the `stats` package.

4. Estimate confidence interval for OU process parameter σ based on a parametric bootstrap of θ.

A.11 Comparison with Tran et al. (2021)

To illustrate the computational benefits of our proposed block coordinate descent algorithm for estimation relative to the Bayesian approach taken in Tran et al. (2021) [112], we apply both methods to simulated datasets. Because we only consider continuous outcomes in this work, we slightly modify the original model proposed in Tran et al. (2021) [112] and do not estimate the additional parameters used to account for non-continuous outcomes. Tran et al. (2021) [112] also consider two different sets of constraints on the OU process drift matrix (denoted here as θ_{OU}); we use the set of constraints that specify the eigenvalues of θ_{OU} to have positive real parts.

We use the same simulation set-up as described in the main text (Section 2.4.1) with the true OU process parameters corresponding to setting 1 (Section A.7). We make one modification to the true values of the loadings parameters: we restrict *all* elements of the loadings matrix to be positive. This restriction means that $\lambda_3 = 0.4$, rather than the original $\lambda_3 = -0.4$. We make this assumption in order to make identification of parameters more straightforward in this comparison of methods.

We generate 100 replicates of the simulated dataset and fit the OUF model using our proposed estimation algorithm and the algorithm proposed in Tran et al. (2021) [112]. Tran et al. (2021) [112] use a slightly different parameterization of the OU process than we use in this work. In our implementation of the OU process, we restrict the volatility parameter matrix, σ_{OU}, to be a diagonal matrix. Although Tran et al. (2021) [112] do not make this assumption, there is still a one-to-one correspondence between the set of parameters estimated in our work and the set of posterior estimates resulting from their Bayesian method. As a result of these differences in parameters, we do not report estimates of σ_{OU} in the plot below and instead present parameter estimates for ρ, which is the stationary correlation between η_1 and η_2. Tran et al. [112] estimate this parameter directly and we can

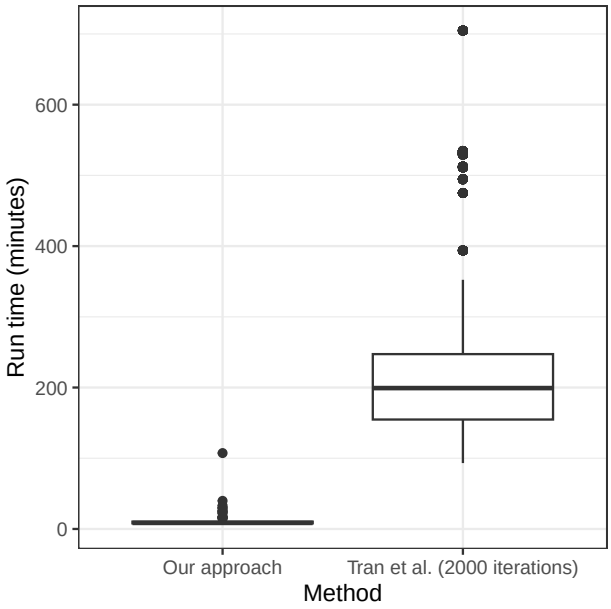


Figure A.3: Computation time (in minutes) for our estimation algorithm and the Bayesian method proposed in Tran et al. (2021) [112]. Box plots summarizes the computation time required to fit the OU factor model using both approaches across 100 simulated datasets. Time required to compute initial parameter estimates is not included in the total above. For our approach, the total time includes both the time required to carry out the block coordinate descent algorithm plus the time required to estimate standard errors.

calculate an estimate for it using $\hat{\theta}_{OU}$ and $\hat{\sigma}_{OU}$.

When applying the Bayesian approach, we use our proposed empirical approach to initializing parameter values, assume 4 chains, and allow the sampler to run for 2,000 iterations. We discard the first half of these samples as burn-in. The computation time of both approaches—excluding time required to compute initial parameter estimates—is shown in Figure A.3. Computing resources are the same across all replicates; we use 4 cores with a total of 4GB of memory for each replicate (this allows gradients to be evaluated or chains to be sampled in parallel, depending on the method). We find that our approach, on average, takes approximately 5% of the time required by the method in Tran et al. (2021) [112].

Point estimates for both estimation approaches are shown in Figure A.4. We present the posterior means for each parameter across the 100 simulated datasets as estimated using the method from Tran et al. (2021) [112]; maximum likelihood estimates resulting from our block coordinate descent algorithm are also summarized across the 100 simulated datasets. Note that the posterior estimates from the Bayesian estimate may be slightly improved by running the sampling algorithm for additional iterations; we limit the MCMC algorithm to 2,000 iterations since our focus is on comparing computation time.

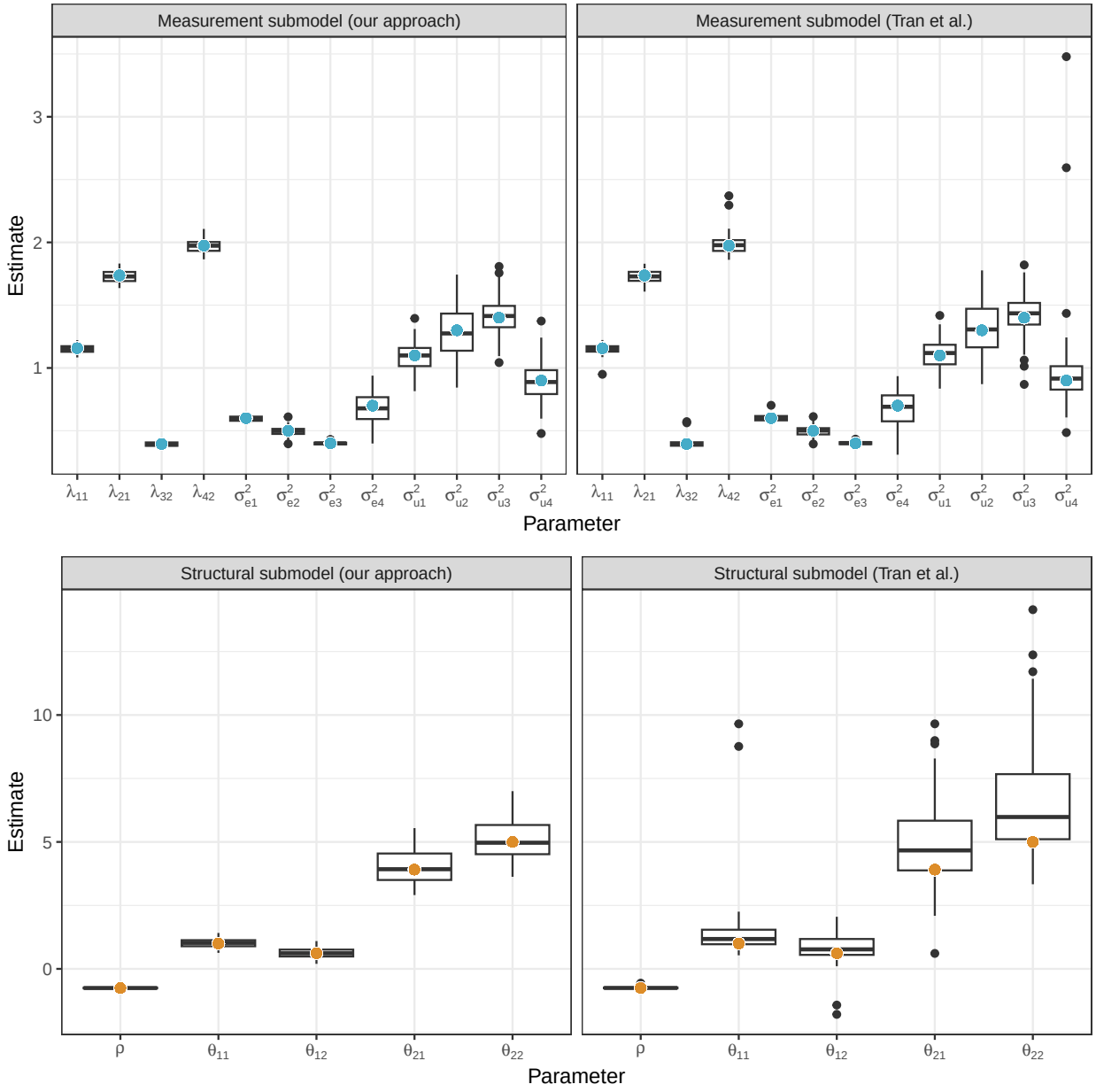


Figure A.4: Final parameter estimates from our block coordinate descent algorithm and the Bayesian method proposed in Tran et al. (2021) [112]. For the Bayesian method, posterior means are used for point estimates. Each box plot summarizes point estimates across the 100 simulated datasets. True parameter values are indicated with colored dots.

APPENDIX B

Supplementary Material for: A Latent Variable Approach to Jointly Modeling Longitudinal and Cumulative Event Data Using a Weighted Two-Stage Method

B.1 Data Structure

The proposed method is motivated by data in which both the longitudinal and event outcomes are partially unobserved. That is, we only observed the longitudinal outcome (e.g., emotions) at measurement occasions (e.g., random ecological momentary assessments or EMAs) and do not directly observe the states that represent risk (e.g., latent emotional states of positive and negative affect). For the event outcome, we only observe the total number of events over some interval of time (e.g., number of cigarettes smoked over a window of time). In Figure B.1, we provide an illustration of the general structure of these data.

B.2 Multivariate Ornstein-Uhlenbeck Stochastic Process

To model the correlated evolution of the multiple latent variables, we use a multivariate Ornstein-Uhlenbeck (OU) process. This process can be viewed as a continuous-time version of the discrete-time vector autoregressive process. Letting η be a (np)-length vector of a p-dimensional OU process observed at n different occasions, the joint distribution of η is

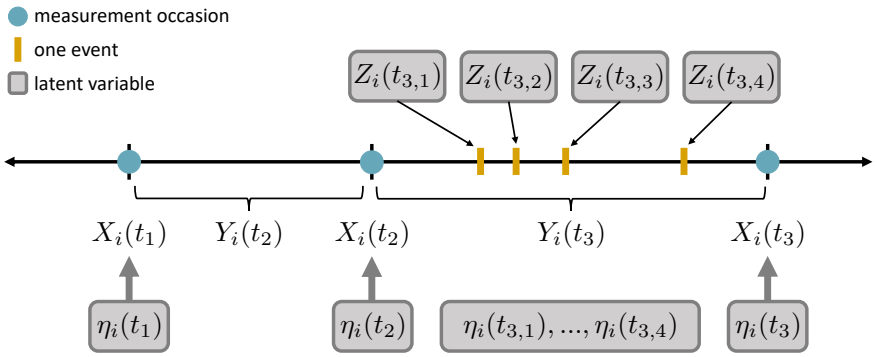


Figure B.1: Illustration of two event intervals as defined by three measurement occasions (i.e., random EMAs). We assume that the longitudinal outcomes, $X_i(t_j)$, are noisy observations of the true latent states, $\eta_i(t_j)$. We also observe the total number of events that occur over the windows of time between measurement occasions, where this cumulative number of events is indicated by $Y_i(t_j)$. In this illustration, we have $Y_i(t_2) = 0$ and $Y_i(t_3) = 4$ events. Each individual event, which we do not observe, is indicated by $Z_i(t_{jl})$. Note that event interval endpoints and measurement occasions are not required to be the same.

$$\eta \sim N(0, \Psi)$$

where Ψ is a $(np) \times (np)$ covariance matrix parameterized by two $p \times p$ matrices, θ and σ. The joint covariance matrix Ψ consists of $p \times p$ blocks that correspond to the covariance of the OU process at measurement times s and t, $s \leq t$, where the blocks take the form

$$Cov\{\eta(s), \eta(t)\} = vec^{-1}\left\{(\theta \oplus \theta)^{-1} vec(\sigma\sigma^\top)\right\} e^{-\theta^\top (t-s)} \tag{B.1}$$

$\oplus$ denotes the Kronecker sum and the $vec\{A\}$ operation stacks the columns of matrix A into a column vector.

B.3 Derivation of the Distribution for the Conditional Mean of the Latent Factors

In this section, we derive the distribution of the average value of the latent factors over an interval of time given known values of the latent factors at each endpoint. We start by considering a single latent factor, η, that follows a stationary univariate OU process. Suppose that we have an interval from time 0 to time t, which we divide into J subintervals of equal

length $\frac{1}{J}$. Then we end up with a grid of points defined by times $0 < t_1 < ... < t_J = t$. Consider observations of the univariate OU process taken across this grid of discrete times. Our first goal is to calculate the variance of the average value of the OU process across this interval, which we do by considering a discrete approximation to the integral,

$$\begin{aligned} Var\left(\frac{1}{t}\int_0^t \eta(s)ds\right) &\approx Var\left\{\frac{1}{J}\sum_{j=1}^{J}\eta(t_j)\right\} \\ &= \sum_{j'=1}^{J}\sum_{j=1}^{J}\left(\frac{1}{J}\right)^2 Cov\left(\eta(t_{j'}),\eta(t_j)\right) \end{aligned}$$

Recall that the stationary (and unconditional) variance of the univariate OU process is

$$Cov(\eta(s),\eta(t)) = \frac{\sigma^2}{2\theta}e^{-\theta|t-s|}$$

Then, we can write the variance of the mean as

$$\begin{aligned} \sum_{j'=1}^{J}\sum_{j=1}^{J}\left(\frac{1}{J}\right)^2 Cov\left(\eta(t_{j'}),\eta(t_j)\right) &= \frac{1}{J^2}\sum_{j'=1}^{J}\sum_{j=1}^{J}\frac{\sigma^2}{2\theta}e^{-\theta|t_j-t_{j'}|} \\ &= \frac{1}{J^2}\frac{\sigma^2}{2\theta}\sum_{j'=1}^{J}\sum_{j=1}^{J}e^{-\theta|t_j-t_{j'}|} \\ &= \frac{1}{J}\frac{\sigma^2}{2\theta}\left[1+\sum_{d=1}^{J-1}2(1-\frac{d}{J})exp\left\{-\theta\frac{d}{J}\times t\right\}\right] \\ &= \frac{1}{J}\frac{\sigma^2}{2\theta}\left[1+\sum_{d=1}^{J}2(1-\frac{d}{J})exp\left\{-\theta\frac{d}{J}\times t\right\}\right] \end{aligned}$$

To make the grid finer and finer, let $J \to \infty$. Then, we find that

$$\lim_{J\to\infty}\frac{1}{J}\frac{\sigma^2}{2\theta}\left[1+\sum_{d=1}^{J}2(1-\frac{d}{J})exp\left\{-\theta\frac{d}{J}\times t\right\}\right] = \frac{\sigma^2}{2\theta}2\int_{s=0}^{1}(1-s)e^{-\theta ts}ds$$
$$= \frac{\sigma^2}{2\theta}2\left[\frac{1}{(\theta t)^2}e^{-\theta ts}(\theta ts-\theta t+1)\right]_{s=0}^{1}$$
$$= \frac{\sigma^2}{\theta}\left[\frac{e^{-\theta t}}{(\theta t)^2}-\frac{1-\theta t}{(\theta t)^2}\right]$$

We extend this result to the multivariate OU process for p latent factors, where η is now a p-length column vector. Viewing θ and σ each as $p\times p$ matrices, the variance becomes

$$Var\left(\frac{1}{t}\int_0^t\eta(s)ds\right) = V(\theta^\top t)^{-2}\left(e^{-\theta^\top t}-I+\theta^\top t\right)+(\theta t)^{-2}\left(e^{-\theta t}-I+\theta t\right)V := V_{\bar{\eta}(s,t)}$$

where $V := vec^{-1}\left\{(\theta\oplus\theta)^{-1}vec(\sigma\sigma^\top)\right\}$. We now write the distribution for the average OU process, denoted $\bar{\eta}(0,t)$ across some interval 0 to t as

$$\bar{\eta}(0,t)\sim N_p\left(0, V(\theta^\top t)^{-2}\left(e^{-\theta^\top t}-I+\theta^\top t\right)+(\theta t)^{-2}\left(e^{-\theta t}-I+\theta t\right)V\right) \tag{B.2}$$

Because we have assumed a stationary OU process, the distribution above holds for the average value of η across any interval of width t.

We next derive the conditional distribution of the average over an interval from s to t, denoted as $\bar{\eta}(s,t)$, given known values of η at the endpoints of a wider interval, say from t_L to t_R, where $t_L\le s<t\le t_R$. Based on previous results, we know that

$$\begin{bmatrix}\eta(t_L)\\ \bar{\eta}(s,t)\\ \eta(t_R)\end{bmatrix}\sim N\left(\begin{bmatrix}0_p\\0_p\\0_p\end{bmatrix},\begin{bmatrix}V & ? & Ve^{-\theta^\top|t_R-t_L|}\\ ? & V_{\bar{\eta}(s,t)} & ?\\ e^{-\theta|t_L-t_R|}V & ? & V\end{bmatrix}\right)$$

However, we still need to fill in the missing blocks of this covariance matrix.

For now, we return to the univariate setting. We start by calculating the covariance of the latent process at the left endpoint of the interval and the average value:

$$
\begin{aligned}
Cov\left(\eta(t_L), \bar{\eta}(s,t)\right) &= Cov\left(\eta(t_L), \frac{1}{J}\sum_{j=1}^{J}\eta(t_j)\right) \\
&= \frac{1}{J}\left[Cov\left(\eta(t_L),\eta(t_1)\right) + Cov\left(\eta(t_L),\eta(t_2)\right) + ... + Cov\left(\eta(t_L),\eta(t_J)\right)\right] \\
&= \frac{1}{J}\sum_{j=1}^{J} Cov\left(\eta(t_L),\eta(t_j)\right) \\
&= \frac{1}{J}\frac{\sigma^2}{2\theta}\sum_{j=1}^{J} e^{-\theta|t_j - t_L|} \\
&= \frac{1}{J}\frac{\sigma^2}{2\theta}\sum_{d=0}^{J} e^{-\theta|s+\frac{d(t-s)}{J} - t_L|}
\end{aligned}
$$

Let $J \to \infty$, then

$$
\begin{aligned}
\lim_{J\to\infty}\frac{1}{J}\frac{\sigma^2}{2\theta}\sum_{d=0}^{J} e^{-\theta|s+\frac{d}{J}(t-s)-t_L|} &= \frac{\sigma^2}{2\theta}\int_{u=0}^{1} e^{-\theta|s+u(t-s)-t_L|}du \\
&= \frac{\sigma^2}{2\theta}\int_{u=0}^{1} e^{-\theta[s+u(t-s)-t_L]}du \\
&= \frac{\sigma^2}{2\theta}\left[\frac{-1}{\theta(t-s)}e^{-\theta(s+u(t-s)-t_L)}\right]_{u=0}^{1} \\
&= \frac{\sigma^2}{2\theta}\frac{-1}{\theta(t-s)}\left[e^{-\theta(t-t_L)} - e^{-\theta(s-t_L)}\right]
\end{aligned}
$$

If the OU process is multivariate, then we can write the covariance as

$$
Cov\left\{\eta(t_L), \bar{\eta}(s,t)\right\} = \frac{-1}{(t-s)}V(\theta^\top)^{-1}\left[e^{-\theta^\top(t-t_L)} - e^{-\theta^\top(s-t_L)}\right]
$$

$$
Cov\left\{\bar{\eta}(s,t), \eta(t_R)\right\} = \frac{-1}{(t-s)}V(\theta^\top)^{-1}\left[e^{-\theta^\top(t_R-s)} - e^{-\theta^\top(t_R-t)}\right]
$$

Putting all of this together, the joint distribution of the average value of η over the interval from s to t and the value of η at the endpoints t_L and t_R, $t_L \leq s < t \leq t_R$, is

$$\begin{bmatrix} \bar{\eta}(s,t) \\ \eta(t_L) \\ \eta(t_R) \end{bmatrix} \sim N \left(\begin{bmatrix} 0_p \\ 0_p \\ 0_p \end{bmatrix}, \begin{bmatrix} Var\{\bar{\eta}(s,t)\} & Cov\{\bar{\eta}(s,t), \eta(t_L)\} & Cov\{\bar{\eta}(s,t), \eta(t_R)\} \\ Cov\{\eta(t_L), \bar{\eta}(s,t)\} & Var\{\eta(t_L)\} & Cov\{\eta(t_L), \eta(t_R)\} \\ Cov\{\eta(t_R), \bar{\eta}(s,t)\} & Cov\{\eta(t_R), \eta(t_L)\} & Var\{\eta(t_R)\} \end{bmatrix} \right) \tag{B.3}$$

where

$$\begin{aligned} Var\{\bar{\eta}(s,t)\} =& V(\theta^\top(t-s))^{-2} \left(e^{-\theta^\top(t-s)} - I + \theta^\top(t-s) \right) + \\ & (\theta(t-s))^{-2} \left(e^{-\theta(t-s)} - I + \theta(t-s) \right) V \end{aligned}$$

$$Var\{\eta(t_L)\} = V$$

$$Var\{\eta(t_R)\} = V$$

$$Cov\{\eta(t_L), \bar{\eta}(s,t)\} = \tfrac{-1}{(t-s)} V(\theta^\top)^{-1} \left[e^{-\theta^\top(t-t_L)} - e^{-\theta^\top(s-t_L)} \right]$$

$$Cov\left\{\bar{\eta}(s,t), \eta(t_R)\right\} = \tfrac{-1}{(t-s)} V(\theta^\top)^{-1} \left[e^{-\theta^\top(t_R-s)} - e^{-\theta^\top(t_R-t)} \right]$$

$$Cov\{\eta(t_L), \eta(t_R)\} = V e^{-\theta^\top(t_R-t_L)}$$

Finally, it is straightforward to derive the conditional distribution of $\bar{\eta}(s,t)$ given both $\eta(t_L)$ and $\eta(t_R)$ using properties of multivariate normal distributions.

B.4 Justification for Approximate Estimation Algorithm

In this section, we provide a justification of the approximate expectation-maximization (EM) algorithm that motivates the weighted two-stage approach we use to fit our model.

Stage 1: In the first stage of our estimation algorithm, we fit the longitudinal submodel—specifically a dynamic factor model—by directly maximizing the log-likelihood for the observed longitudinal data, X, via a block coordinate descent algorithm. The distribution of the observed longitudinal outcome for individual i, with parameter vector Θ_L, is:

$$p(X_i; \Theta_L) = \int p(X_i | \eta_i, u_i; \Theta_L) p(u_i; \Theta_L) p(\eta_i; \Theta_L) d\eta du$$

Estimation of this submodel is discussed further in Abbott et al. (2023) [1].

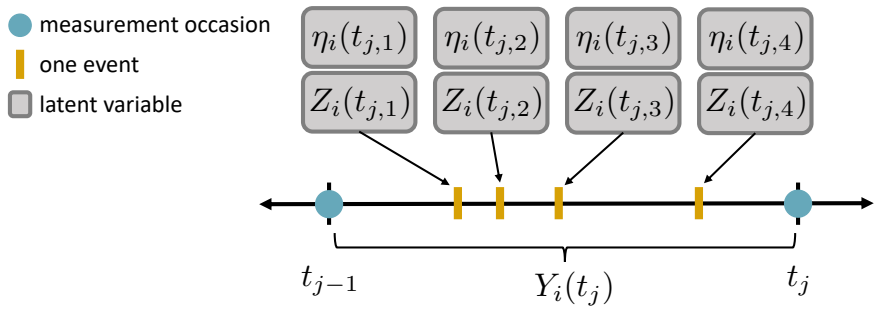


Figure B.2: Illustration of data collected over a single event interval for a single individual. In an ideal world, we would observe each individual event when it occurs, denoted by $Z_i(t_{jl})$, and the value of the latent process at each event, $\eta_i(t_{jl})$. For illustrative purposes, we assume here that $l = 1, 2, 3, 4$ and that $\sum_l Z_i(t_{jl}) = Y_i(t_j) = 4$. In reality, both $Z_i(t_{jl})$ and $\eta_i(t_{jl})$ are latent.

Stage 2: In this stage, we aim to fit the cumulative risk submodel given the longitudinal process—here, the longitudinal process that we consider is entirely latent, rather than the intermittently measured longitudinal outcome. Suppose that we were to observe the complete data; specifically, we observe the time of each individual repeated event, denoted by $Z_i(t_{jl})$, along with the values of the latent process at each event time, $\eta_i(t_{jl})$, for each individual $i = 1, ..., N$. Here we omit the individual-specific index from our time variable t. Because our primary interest is in understanding the association between the latent process and the risk of an event, observing the measured longitudinal outcome, X_i, does not provide any new information beyond what is provided by η_i.

For now, consider a single event interval from t_{j-1} to t_j for a single individual i. We assume for illustrative purpose that four events occur over this interval, as presented in the diagram in Figure B.2.

In reality, we do not get to observe $Z_i(t_{jl}), l = 1, ..., 4$; rather only $Y_i(t_j) = \sum_l Z_i(t_{jl})$. To fit the cumulative risk model, we need a single value of the latent process to associate with each event interval; we take the average of the latent process over each event interval, denoted by $\bar{\eta}_i(t_j)$. Then, we aim to maximize the log-likelihood of our cumulative risk model across all N individuals; the distribution of the cumulative event outcome for individual i, parameterized using vector β, is:

$$p(Y_i(t_j)|\bar{\eta}_i(t_j); \beta) \sim Poisson\left((t_j - t_{j-1})\lambda_i(t_j)\right), \lambda_i(t_j) = f(\bar{\eta}_i(t_j); \beta)$$

where $\bar{\eta}_i$ is the vector of average values of the latent process within each event interval for individual i and $f(\bar{\eta}_i(t_j);\beta)$ is some linear combination of the average values of the p latent factors with parameters β. In our formulation of the cumulative risk model, risk is independent of the observed longitudinal outcome, X, given the latent process, η. To fit the cumulative risk submodel, we need to know the corresponding value of η (or, actually, $\bar{\eta}$) for each event interval. Following the standard approach in an expectation maximization algorithm, we would ideally generate values of the latent process conditional on the number of observed cumulative events, Y. However, sampling from this conditional distribution of $\eta_i|Y_i$ is not straightforward. Instead of directly sampling from $\eta_i|Y_i$, we use weights based on importance sampling to re-weight values of η drawn from the marginal distribution (unconditional on Y). These weights allow us to approximate samples from $p(\eta|Y)$ so that our algorithm simply corresponds to iteratively maximizing a weighted Poisson model.

We derive the importance sampling-based weights here. Our target distribution is

$$p(\bar{\eta}_i(t_j)|Y_i(t_j);\Theta_L,\beta)$$

but we can more easily sample from $p(\bar{\eta}_i(t_j);\Theta_L)$, since this is simply a multivariate normal distribution (see Section B.3 for the exact form). So, we can re-weight the likelihood corresponding to $Y|\bar{\eta}$ (our Poisson regression model) using samples from $p(\bar{\eta}_i(t_j);\Theta_L)$ according to weights,

$$w = \frac{p(\bar{\eta}_i(t_j)|Y_i(t_j);\Theta_L,\beta)}{p(\bar{\eta}_i(t_j);\Theta_L)} \approx p(Y_i(t_j)|\bar{\eta}_i(t_j);\Theta_L,\beta) = p(Y_i(t_j)|\bar{\eta}_i(t_j);\beta)$$

Importantly, uncertainty still exists in the values of $\bar{\eta}_i(t_j)$; we capture this uncertainty through a Monte Carlo approach by drawing multiple samples of the average value of the latent process, indexed by $r = 1, ..., R$, for each event interval. This repeated sampling means that our weights are also indexed by r. To fit the cumulative risk model while capturing this uncertainty, we replicate each observation of the cumulative number of events, $Y_i(t_j)$, R times. Then we fit a weighted Poisson regression model for the outcome $Y_i^{(r)}(t_j)$ with predictors $\bar{\eta}_i^{(r)}(t_j)$, where we normalize the weights to maintain our original sample size.

Putting all of this together, we can fit our cumulative risk model in the second stage of this two-stage approach by iterating between (i) fitting a weighted Poisson regression model for $Y^{(r)}|\bar{\eta}^{(r)}$ with weights $w^{(r)}$ to generate new estimates of β and then (ii) updating the weights based on $Pr(Y_i^{(r)}(t_j)|\bar{\eta}_i^{(r)}(t_j);\hat{\beta}^{(k)})$ where k indexes the iteration number.

B.5 Simulation Study Design

The true parameter values used to generate the data in the simulation study are given here:

Setting 1: The true OU process has higher autocorrelation and higher noise. We expect estimation to be easier in this setting.

$$\theta = \begin{bmatrix} 0.9 & 0.4 \\ 0.5 & 1 \end{bmatrix} \text{ and } \sigma = \begin{bmatrix} 1.19 & 0 \\ 0 & 1.24 \end{bmatrix}$$

The true values of the coefficients in the cumulative risk model are

$$\beta_0 = -2.4, \beta_1 = -0.9, \beta_2 = 1$$

Setting 2: The true OU process has lower autocorrelation and higher noise. We expect estimation to be more difficult in this setting.

$$\theta = \begin{bmatrix} 1.8 & 0.4 \\ 0.2 & 1.5 \end{bmatrix} \text{ and } \sigma = \begin{bmatrix} 1.86 & 0 \\ 0 & 1.71 \end{bmatrix}$$

The true values of the coefficients in the cumulative risk model are

$$\beta_0 = -2.2, \beta_1 = -0.6, \beta_2 = 0.8$$

In both setting 1 and setting 2, we use the same set of true values for the other parameters in the longitudinal submodel. These true parameter matrices are:

$$\Lambda = \begin{bmatrix} 0.82 & 0 \\ 1.04 & 0 \\ 0.88 & 0 \\ 1.16 & 0 \\ 0.81 & 0 \\ 1.27 & 0 \\ 0.98 & 0 \\ 1.15 & 0 \\ 0.92 & 0 \\ 0 & 1.19 \\ 0 & 0.93 \\ 0 & 1.24 \\ 0 & 1.00 \\ 0 & 1.16 \\ 0 & 0.86 \\ 0 & 1.02 \\ 0 & 1.25 \\ 0 & 1.06 \end{bmatrix}$$

$$\Sigma_u = diag\{0.1, 0.2, 0.3, 0.4, 0.2, 0.3, 0.1, 0.6, 0.5, 0.5, 0.05, 0.8, 0.4, 0.1, 0.2, 0.3, 0.4, 0.2\}$$

$$\sigma^2_{\epsilon,k} = 0.1, k = 1, ..., 18$$

Σ_u in covariance matrix for the outcome-specific random intercepts and $\sigma^2_{\epsilon,k}$ denotes the measurement error variance for each of the k observed longitudinal outcomes. The values in the true loadings matrix, Λ, are based on estimates from the smoking cessation data.

B.5.1 Stage 1 Results

For the simulation study, we present the point estimates of parameters in the longitudinal submodel in Figure B.3. Estimated during stage 1, these parameter values are then used in stage 2. If standard error estimates are of interest for stage 1 parameters, a Fisher

CR	Coef.	Setting 1: high correlation			Setting 2: low correlation		
		R = 25	R = 50	R = 100	R = 25	R = 50	R = 100
80%	β_0	**70**	**71**	**71**	83	82	83
	β_1	85	85	85	80	81	82
	β_2	82	82	82	**72**	73	73
95%	β_0	**89**	**89**	**89**	95	94	96
	β_1	96	96	96	94	94	94
	β_2	99	99	99	92	92	92

Table B.1: Coverage rates (CRs; %) for 80% and 95% confidence intervals from the simulation study assessing sensitivity to R, where confidence intervals are calculated from von Hippel standard errors. CRs are averaged across **100 datasets**. We have bolded the CRs that fall outside of the expected range of values based on 80% and 95% binomial proportion confidence intervals.

information-based approach can be used (see Abbott et al. (2023) [1] for more details).

B.5.2 Sensitivity Analysis: Value of R

In the simulation study and data application in the main paper, we selected a value of R with the goal of balancing computational cost and approximation error. Here, we assess if the von Hippel standard errors [120] result in confidence intervals with close to nominal coverage when generating different numbers of average values of the latent process per event interval (i.e., using different values of R). We evaluate the bias of point estimates and coverage of confidence intervals when the true parameters of the latent process correspond to the higher and lower correlation settings. We consider $R = 25, 50, 100$ and present the results in Figure B.4 and Table B.1. Note that $R = 50$ was used in the original simulation study in the main paper.

B.5.3 Computation Time by Stage

We summarize the computation time by each stage in Figure B.5. Stage 1 computation time corresponds to the time required to fit the dynamic factor model. Stage 2 computation time is reported twice: once if only point estimates are desired and once if both point estimates and standard errors are required. Estimation of parameter values, without standard errors, does not require bootstrapping and so computation time is much lower.

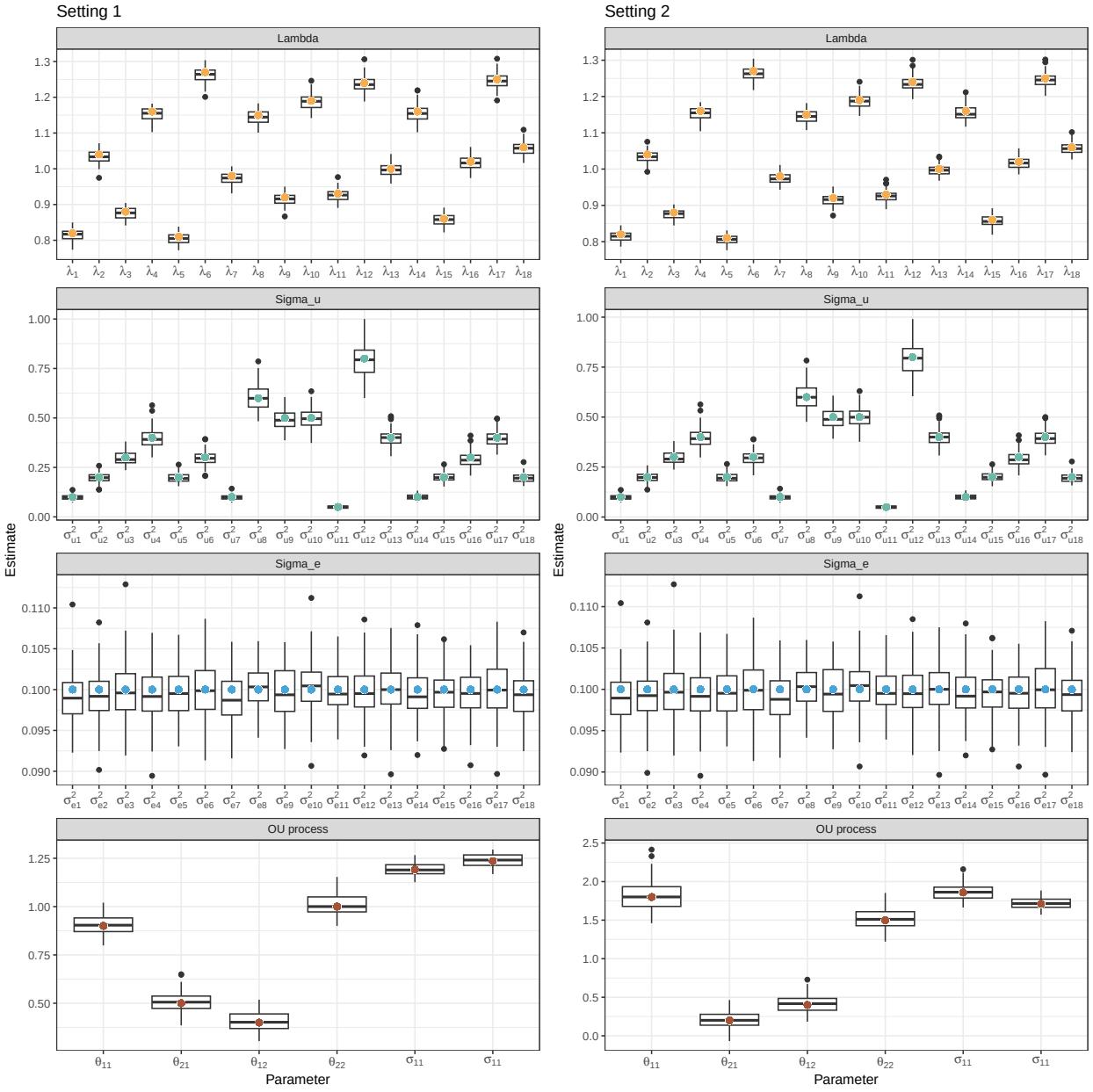


Figure B.3: Stage 1 point estimates from the simulation study. True coefficient values are indicated with colored dots. Point estimates are summarized across 100 replicates using black boxplots.

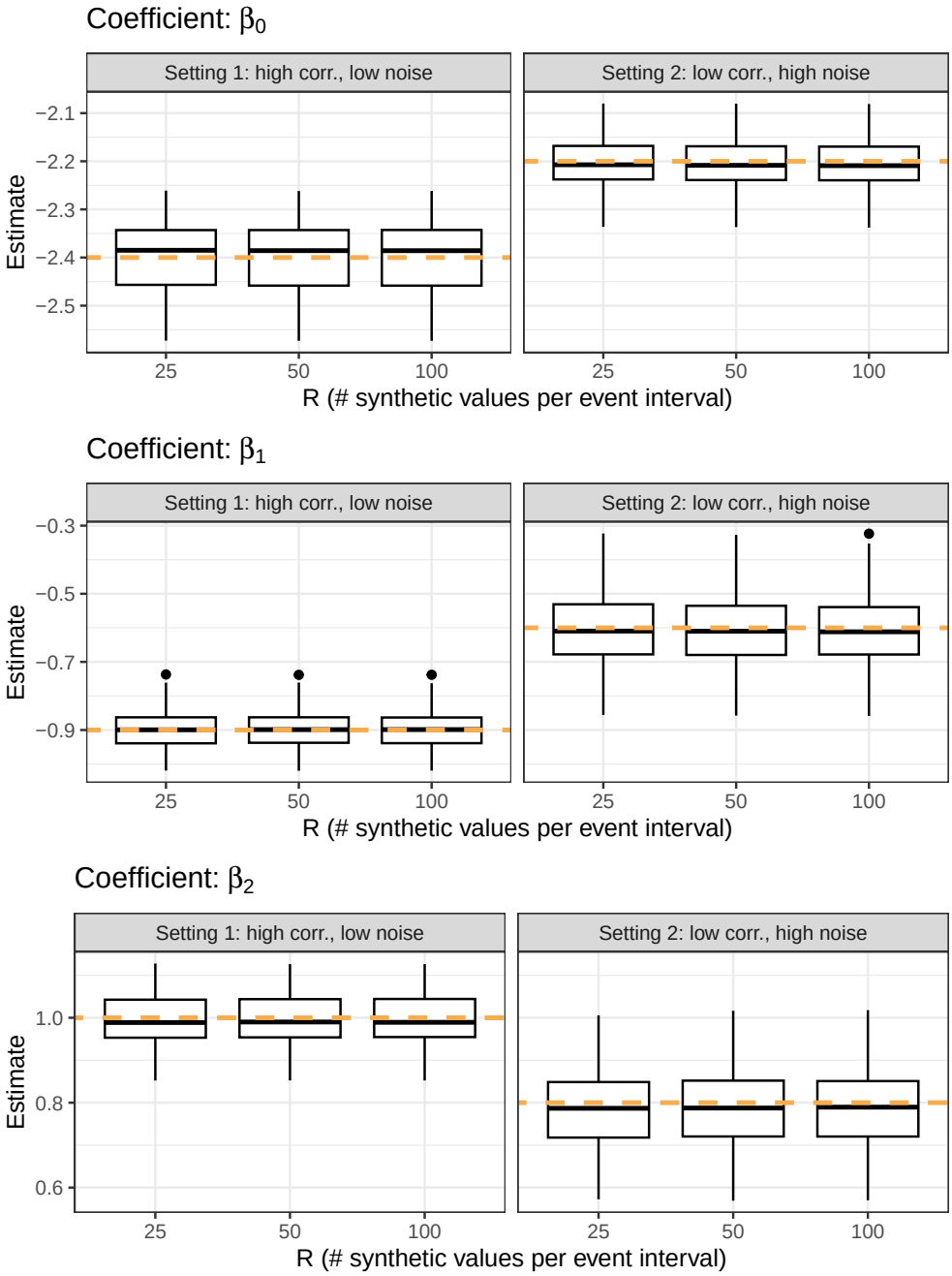


Figure B.4: Point estimates from simulation study assessing sensitivity to R. True coefficient values are indicated with dashed orange horizontal lines. Point estimates are summarized across 100 replicates using black boxplots.

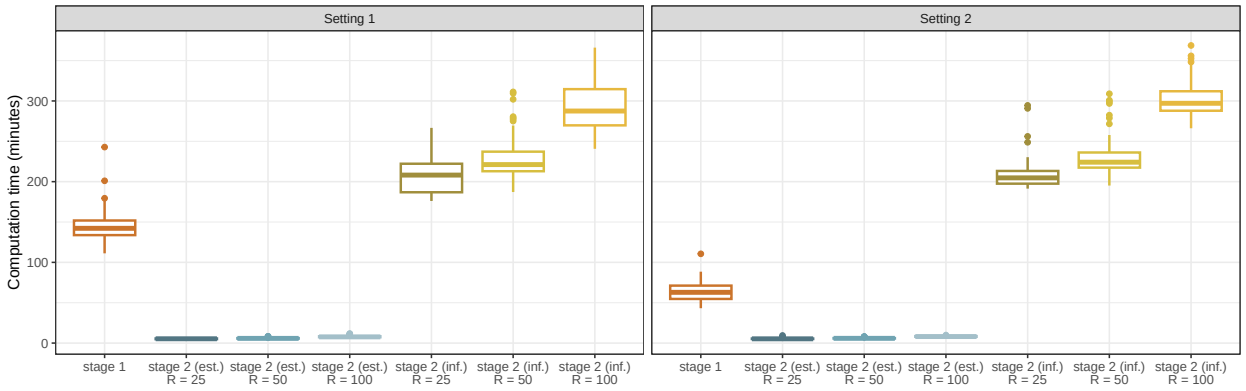


Figure B.5: Computation time (in minutes) for stage 1 and stage 2 in the simulation study, where times are summarized across the 100 replicates using box plots. Times for stage 2 are presented twice: box plots labeled "est." correspond to the time required to calculate point estimates and box plots labeled "inf." correspond to the time required to calculate both point estimates and bootstrap-based standard errors. The computation time of stage 2 depends on the value of R.

B.6 Application to Smoking Cessation Data

In Figure B.6, we plot stage 1 point estimates and 95% confidence intervals estimated for the longitudinal submodel fitted to the larger subset of the smoking cessation data with $N = 218$ individuals. The 95% confidence intervals are based on Fisher information. In Figure B.7, we plot stage 2 point estimates and anti-conservative 95% confidence intervals estimated when we carried out estimation and inference using the non-bootstrapped-based approach with Rubin's Rule [97].

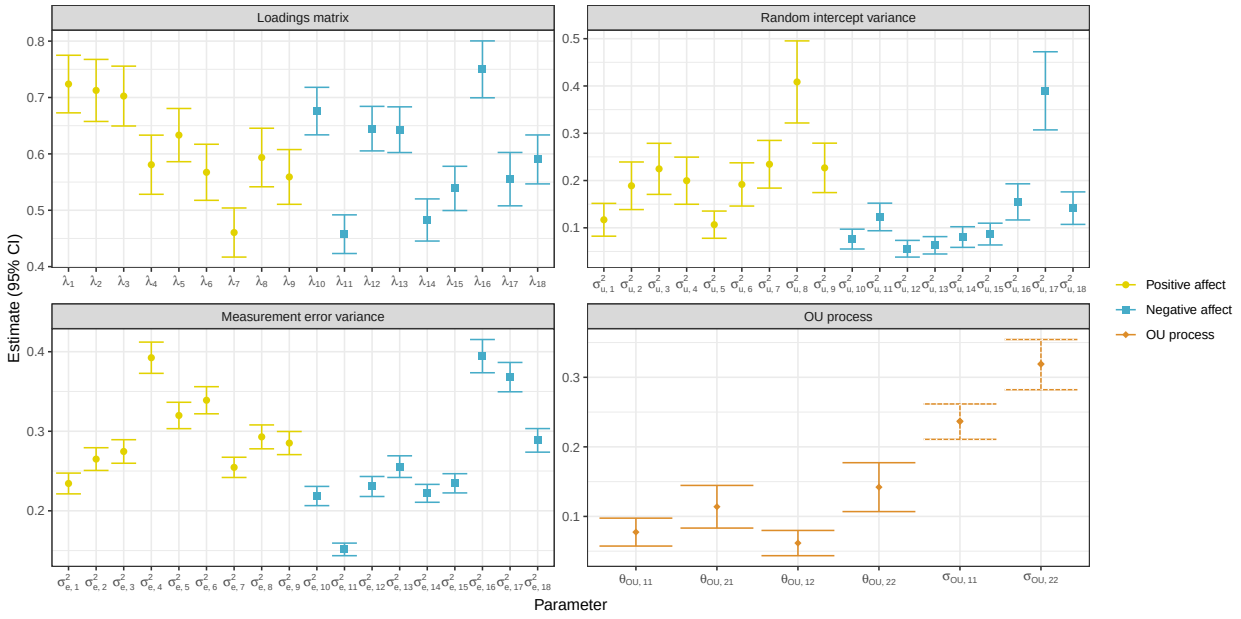


Figure B.6: Point estimates and corresponding 95% confidence intervals (CIs) for parameters in the longitudinal submodel applied to the mHealth smoking cessation study data with $N = 218$ individuals. Intervals for $\sigma_{OU,11}$ and $\sigma_{OU,22}$ are based on a parametric bootstrap. Subscripts for the loading, random intercept, and measurement error parameters correspond to the following emotions: 1 = happy, 2 = joyful, 3 = enthusiastic, 4 = active, 5 = calm, 6 = determined, 7 = grateful, 8 = proud, 9 = attentive, 10 = sad, 11 = scared, 12 = disgusted, 13 = angry, 14 = ashamed, 15 = guilty, 16 = irritable, 17 = lonely, 18 = nervous. This figure has been previously published [1].

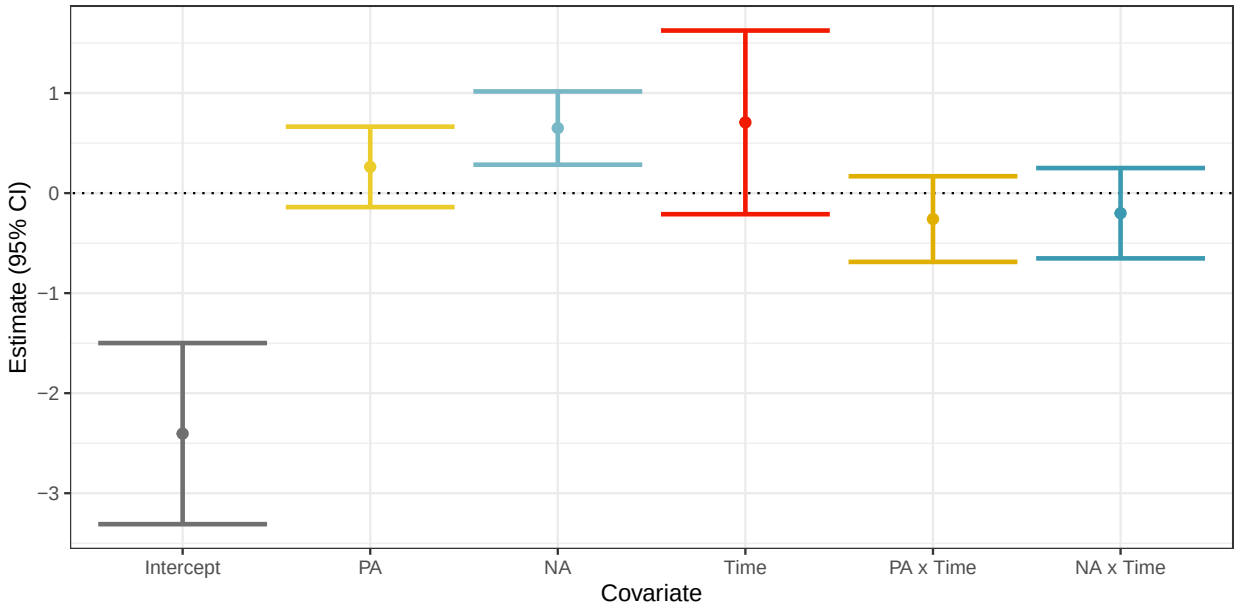


Figure B.7: Point estimates and 95% confidence intervals (CIs) for cumulative risk model parameters applied to the subset of mHealth smoking cessation study data with $N = 214$ individuals. Error bars correspond to 95% CIs calculated from anti-conservative standard error estimates (using Rubin's Rule).

APPENDIX C

Supplementary Material for: A Bayesian Joint Longitudinal-Survival Model with a Latent Stochastic Process for Intensive Longitudinal Data

C.1 Identifiability and Parameter Constraints

C.1.1 Converting Between OU Process Parameterizations

To ensure that our dynamic factor model is identifiable, we model the Ornstein-Uhlenbeck (OU) process on the correlation scale. A p-dimensional OU process is generally parameterized in terms of two p-dimensional OU process, θ and σ, where θ controls the speed of mean reversion and σ controls the volatility of the process. The SDE form of a bivariate OU process, denoted by η_1 and η_2, is:

$$d\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix} = -\underbrace{\begin{bmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\end{bmatrix}}_{:=\boldsymbol{\theta}_{OU}}\begin{bmatrix}\eta_1(t)\\ \eta_2(t)\end{bmatrix}dt + \underbrace{\begin{bmatrix}\sigma_{11} & \sigma_{21}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}}_{:=\boldsymbol{\sigma}_{OU}} d\begin{bmatrix}W_1(t)\\ W_2(t)\end{bmatrix}$$

where $W_1(t)$, $W_2(t)$ are standard Brownian motion. The OU process is most often written in terms of its conditional distribution. Assuming that the initial value of the latent process is $\boldsymbol{\eta}(t_0) \sim N_p(\mathbf{0}, \boldsymbol{V})$, where $\boldsymbol{V} = vec^{-1}\big\{(\boldsymbol{\theta}_{OU} \oplus \boldsymbol{\theta}_{OU})^{-1} vec(\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top})\big\}$, then for $j = 1, ...,$

$$\boldsymbol{\eta}(t_j)|\boldsymbol{\eta}(t_{j-1}) \sim N_p\left(e^{-\boldsymbol{\theta}_{OU}(t_j - t_{j-1})}\boldsymbol{\eta}(t_{j-1}), \boldsymbol{V} - e^{-\boldsymbol{\theta}_{OU}(t_j - t_{j-1})}\boldsymbol{V}e^{-\boldsymbol{\theta}_{OU}^{\top}(t_j - t_{j-1})}\right)$$

If the OU process is modeled on the correlation scale, then $\boldsymbol{V}$ is a correlation matrix. We use ρ to denote the off-diagonal parameter(s) of $\boldsymbol{V}$. For a bivariate OU process, we must

estimate only one off-diagonal parameter in $\boldsymbol{V}$, where

$$\boldsymbol{V} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

When estimating the parameters that describe the OU process, we could do so using the parameterization in the SDE or the parameterization of the conditional distribution. That is, we could estimate either $(\boldsymbol{\theta}_{OU}, \boldsymbol{\sigma}_{OU})$ or $(\boldsymbol{\theta}_{OU}, \rho)$.

While the SDE version of the OU process is a function of $\boldsymbol{\sigma}_{OU}$, the conditional distribution is a function of $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ (that is, $\boldsymbol{\sigma}_{OU}$ never shows up outside of the term $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$). We know that $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ is symmetric and positive definite. Because $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ is symmetric, this means that the conditional distribution only depends on the identifiable parameter $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ (plus additional parameters in $\boldsymbol{\theta}_{OU}$). In the case of the bivariate OU process, $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ contains three identifiable parameters (that is, the upper *or* lower triangle is identifiable).

Although $\boldsymbol{\sigma}_{OU}$ is not immediately identifiable from a known $\boldsymbol{\theta}_{OU}$ and $\boldsymbol{V}$, we can always re-parameterized an arbitrary OU process with a full $\boldsymbol{\sigma}_{OU}$ to have a triangular $\boldsymbol{\sigma}_{OU}$ while maintaining the same correlation structure. This fact results from the stationary covariance formula being a function of $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$:

$$\boldsymbol{V} = vec^{-1}\{(\boldsymbol{\theta}_{OU} \oplus \boldsymbol{\theta}_{OU})^{-1} vec[\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}]\}$$

We can decompose $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ into $\boldsymbol{\sigma}_{OU}$ using the Cholesky decomposition, which says that if matrix A is positive definite, then $\boldsymbol{A}$ can be decomposed into two lower-triangular matrices $\boldsymbol{L}$, where $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{\top}$. Furthermore, since $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ is positive definite, then we know that the Cholesky decomposition of $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$ is unique. To convert an OU process with parameters $\boldsymbol{\theta}_{OU}^{*}, \rho$ to parameters $\boldsymbol{\theta_{OU}}, \boldsymbol{\sigma}_{OU}$, we can simply solve the stationary variance $\boldsymbol{V}$ (containing parameter(s) ρ) for $\boldsymbol{\sigma}_{OU}\boldsymbol{\sigma}_{OU}^{\top}$.

C.1.2 Constraints on θ_{OU}

Tran et al. (2021) [112] discuss constraints on the OU process $\boldsymbol{\theta}_{OU}$ matrix, which ensure that $\boldsymbol{\theta}_{OU}$ corresponds to a mean reverting process but does not restrict it from being non-oscillating. We apply these constraints developed in this prior work, which constrain the real parts of the eigenvalues of bivariate OU process parameter $\boldsymbol{\theta}_{OU}$ to be positive:

$$v_1 = \theta_{OU_{11}} + \theta_{OU_{22}}$$
$$v_2 = \theta_{OU_{11}}\theta_{OU_{22}} - \theta_{OU_{12}}\theta_{OU_{21}}$$

where v_1 and v_2 must be positive. Tran et al. (2021) [112] also discuss eigenvalue constraints for a trivariate OU process.

C.2 Simulation Study Design

C.2.1 Measurement Patterns

In our simulation study, we generate observations of our longitudinal outcomes using four different patterns in measurement occasions:

Measurement pattern 1: The measurements occur frequently and with constant probability. To determine the timing of the measurements, we sample uniformly from a fine grid of possible times spanning 0 to 28 days, where a maximum of 60 and 70 measurement times are drawn in settings 1 and 2, respectively. After censoring of the longitudinal measurements due to the survival outcome, an average of 19.2 and 24.4 measurement occasions are observed per individual in settings 1 and 2, respectively.

Measurement pattern 2: The measurements occur less frequently but still with constant probability. To determine the timing of the measurements, we sample uniformly from a fine grid of possible times spanning 0 to 28 days, where a maximum of 15 and 12 measurement times are drawn in settings 1 and 2, respectively. After censoring of the longitudinal measurements due to the survival outcome, an average of 5.5 and 5.0 measurement occasions are observed per individual in settings 1 and 2, respectively.

Measurement pattern 3: The measurements are distributed according to the measurement times observed in the motivating mHealth study, CARE. To determine the timing of the measurements, we sample (with replacement) individuals from the motivating mHealth dataset and then use their observed measurement times to define the measurement times in the simulated dataset. After censoring of the longitudinal measurements due to the survival outcome, an average of 33.5 and 36.9 measurement occasions are observed per individual in settings 1 and 2, respectively.

Measurement pattern 4: The measurements are clustered together and distributed according to probabilities following a truncated cosine function of time. To determine

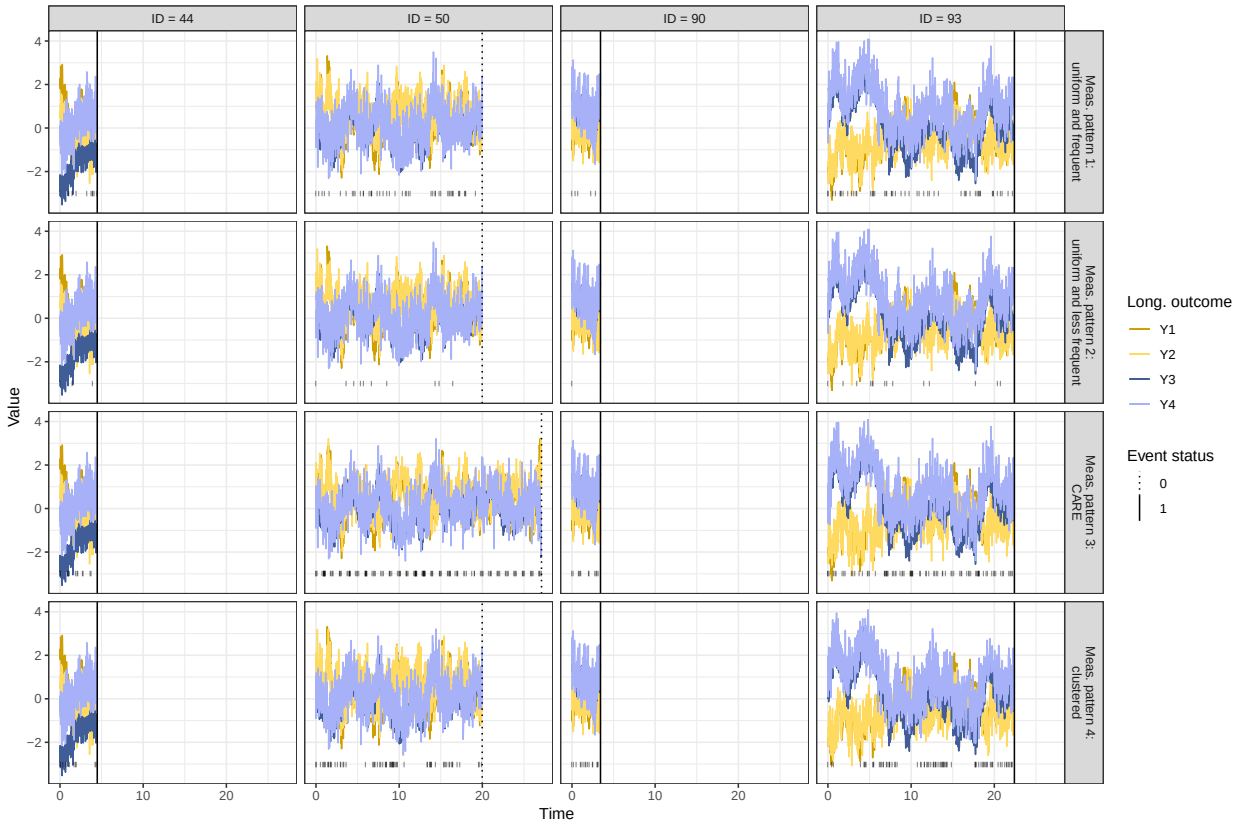


Figure C.1: Each column displays the true values of the four longitudinal outcomes for a different individual in the simulated data under **setting 1**. Each row displays the different patterns in measurement occasions, where the timing of the measurement occasions are indicated by the location of the black tick marks along the x-axis.

the timing of the measurements, we generate measurement probabilities for each point on the fine grid of possible measurement times going from 0 to 28 days. These measurement probabilities are proportional to the absolute value of a cosine function and, in order to induce larger gaps between clusters of measurements, are truncated to 0 when less than 0.4. After censoring of the longitudinal measurements due to the survival outcome, an average of 29.3 and 20.4 measurement occasions are observed per individual in settings 1 and 2, respectively.

We illustrate the different patterns in measurements in Figure C.1 and C.2 for four individuals in datasets simulated under setting 1 and setting 2. Only measurements that occurred prior to the event or censoring time are displayed in this figure.

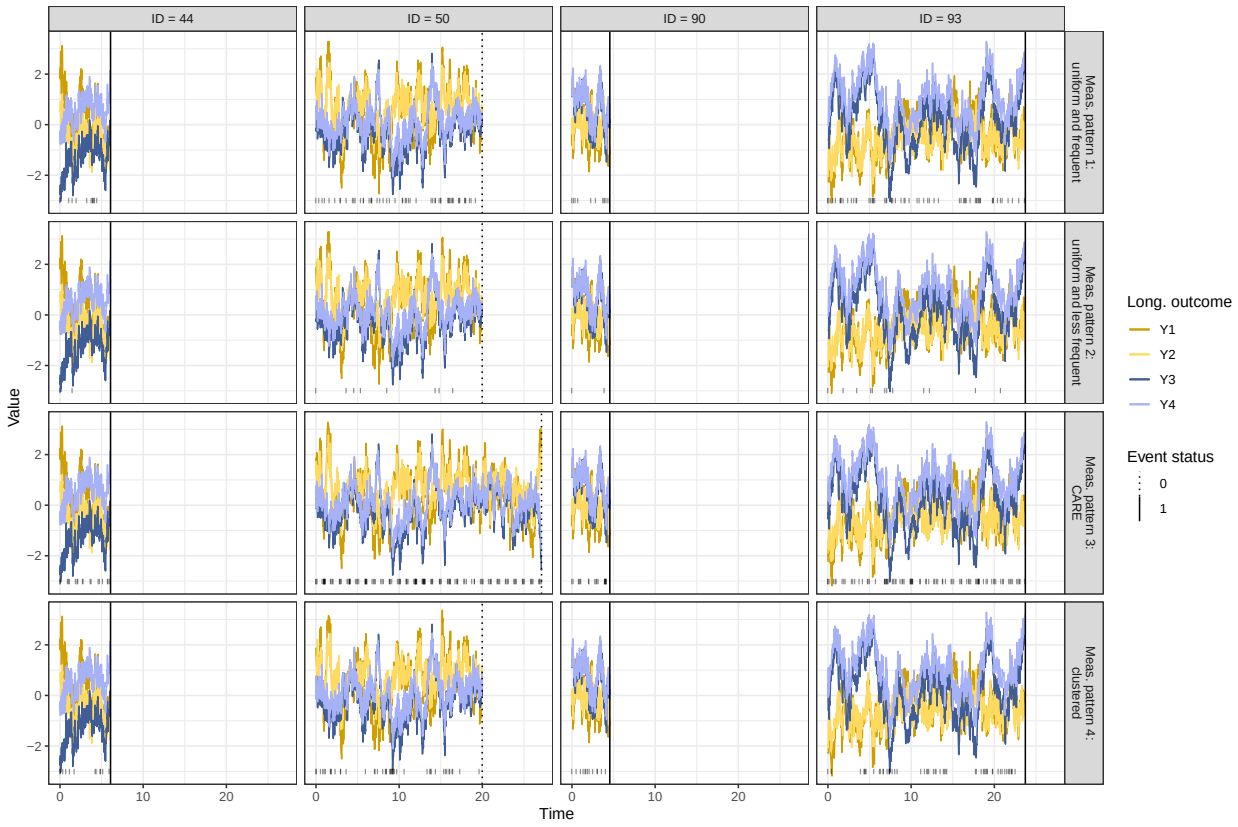


Figure C.2: Each column displays the true values of the four longitudinal outcomes for a different individual in the simulated data under **setting 2**. Each row displays the different patterns in measurement occasions, where the timing of the measurement occasions are indicated by the location of the black tick marks along the x-axis.

C.2.2 True Parameter Values

In setting 1, we assume the following true parameter values:

Measurement submodel:

$$\boldsymbol{\theta}_{OU} = \begin{bmatrix} 1.8 & 0.4 \\ 1.5 & 1.2 \end{bmatrix}, \boldsymbol{\sigma}_{OU} = \begin{bmatrix} 1.76 & 0 \\ 0 & 0.71 \end{bmatrix} \implies \rho = -0.633$$

Structural submodel:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.9 & 0 \\ 0.5 & 0 \\ 0 & 1 \\ 0 & 0.8 \end{bmatrix}$$

$$\sigma_{u_1} = 0.4, \sigma_{u_2} = 0.5, \sigma_{u_3} = 0.8, \sigma_{u_4} = 1$$

$$\sigma_{\epsilon_1} = 0.2, \sigma_{\epsilon_2} = 0.6, \sigma_{\epsilon_3} = 0.3, \sigma_{\epsilon_4} = 0.7$$

Survival submodel:

$$\beta_0 = -2.5, \beta_1 = -0.2, \beta_2 = 0.3$$

In setting 2, we assume the following true parameters values:

Measurement submodel:

$$\boldsymbol{\theta}_{OU} = \begin{bmatrix} 2.4 & 0.4 \\ 0.8 & 2 \end{bmatrix}, \boldsymbol{\sigma}_{OU} = \begin{bmatrix} 2.14 & 0 \\ 0 & 1.89 \end{bmatrix} \implies \rho = -0.273$$

Structural submodel:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.9 & 0 \\ 0.5 & 0 \\ 0 & 1 \\ 0 & 0.8 \end{bmatrix}$$

$$\sigma_{u_1} = 0.4, \sigma_{u_2} = 0.5, \sigma_{u_3} = 0.8, \sigma_{u_4} = 1$$

$$\sigma_{\epsilon_1} = 0.2, \sigma_{\epsilon_2} = 0.3, \sigma_{\epsilon_3} = 0.1, \sigma_{\epsilon_4} = 0.2$$

Survival submodel:

$$\beta_0 = -3, \beta_1 = -0.4, \beta_2 = 0.8$$

In both settings 1 and 2, the true survival submodel is $h_i(t) = exp(\beta_0 + \beta_1\eta_{1i}(t) + \beta_2\eta_{2i}(t))$. The true censoring distribution is $C_i \sim 10 \times Exponential(rate = 0.25)$ for measurement patterns 1, 2, and 4. For measurement pattern 3 in which the timing of measurements is

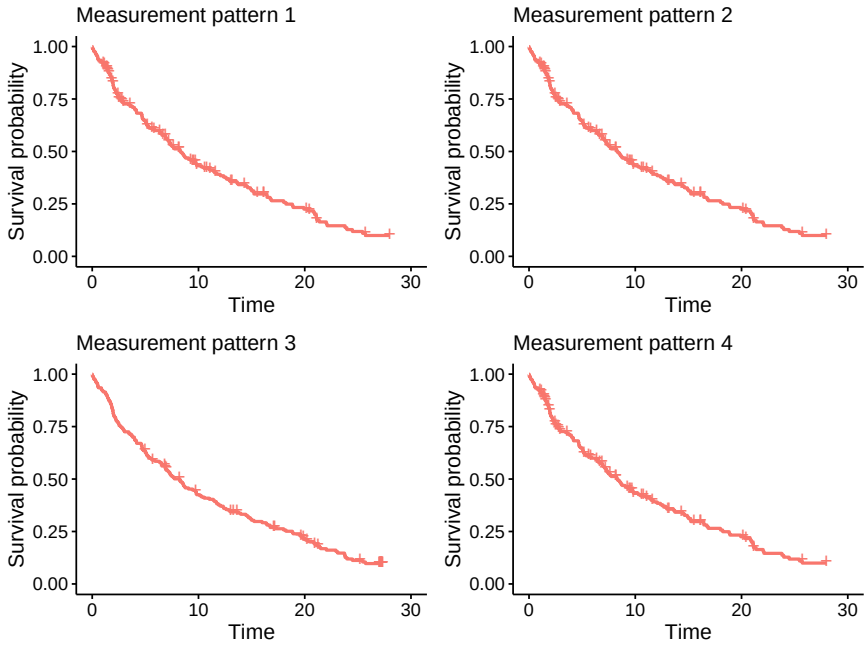


Figure C.3: Kaplan-Meier curves for a single dataset with each measurement pattern (1-4) simulated using the true set of parameters in **setting 1**.

based on those observed in the motivating mHealth study, the timing of an individual's final random EMA is used as the censoring time. Kaplan-Meier curves are plotted in Figures C.3 and C.4 for each combination of setting and measurement pattern.

C.2.3 Initial Parameter Values

When fitting the model in the simulation study (both settings 1 and 2), we specify initial parameter values that are correct in their sign and in their order of magnitude. The specific values are listed below. For parameters not listed below (e.g., η), we rely on the random initial values generated by Stan [18]. In practice, we could take a two-stage approach to initialization.

Measurement submodel:

$$\boldsymbol{\theta}_{OU} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \rho = -0.5$$

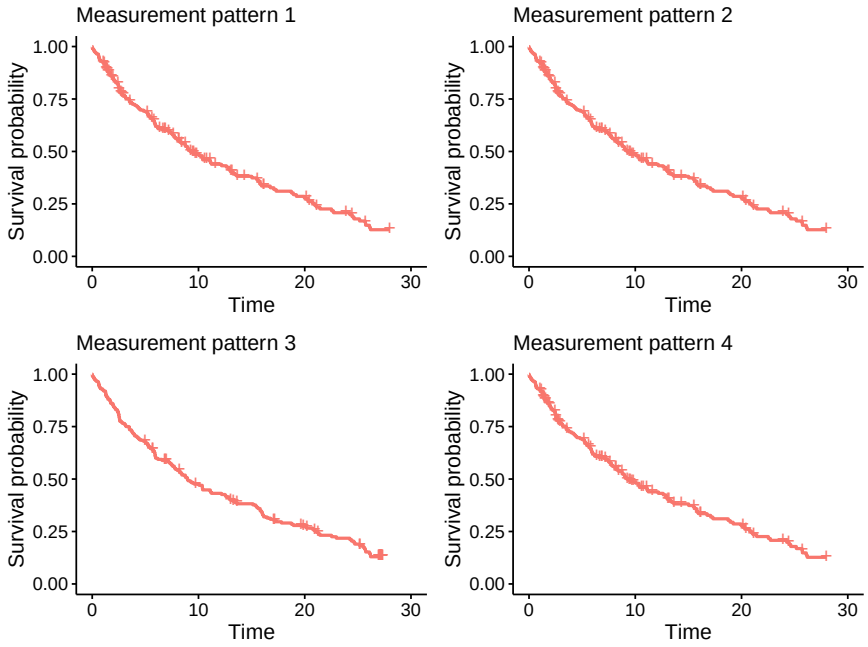


Figure C.4: Kaplan-Meier curves for a single dataset with each measurement pattern (1-4) simulated using the true set of parameters in **setting 2**.

Structural submodel:

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\sigma_{u_k} = 0.1, k = 1, ..., 4$$

$$\sigma_{\epsilon_k} = 0.1, k = 1, ..., 4$$

Survival submodel:

$$\beta_0 = -1, \beta_1 = -1, \beta_2 = 1$$

C.2.4 Prior Distributions

We specify the following prior distributions when fitting the models in our simulation study. These priors are based on those used in Tran et al. (2021) [**?**].

$$\lambda_k \sim \text{half-}N(1, \sigma_\lambda^2); k = 1, ..., 4$$
$$\sigma_\lambda \sim \text{half-Cauchy}(0, 5)$$
$$\theta_{OU_{11}}, \theta_{OU_{21}}, \theta_{OU_{12}}, \theta_{OU_{22}} \sim N(0, 10^2)$$
$$\rho \sim \text{Uniform}(-0.999999, 0.999999)$$
$$\sigma_{u_k} \sim \text{half-}Cauchy(0, 5); k = 1, ..., 9$$
$$\sigma_{u_\epsilon} \sim \text{half-}Cauchy(0, 5); k = 1, ..., 9$$
$$\beta_0, \beta_1, \beta_2 \sim N(0, 5^2)$$

C.3 Investigation of Grid Width

For the simulation study described in the main text, we present complete results in this section for all combinations of true parameter values (settings 1 and 2), measurement patterns (1-4), and grid widths (0.2, 0.4, 1.2, and no grid). Due to the high computation cost, we do not fit models using the finest grid (0.2) for data generated under measurement patterns 3 and 4. Computation times are summarized in Figure C.5.

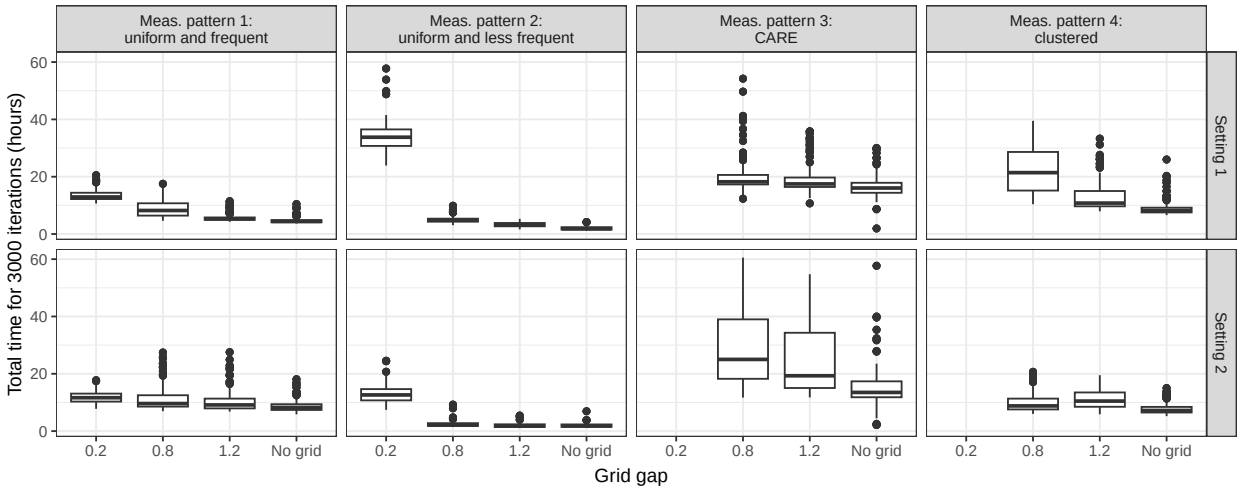


Figure C.5: For data generated under settings 1 and 2 with each of the four measurement patterns, we summarize the time required to run Stan for 3000 iterations.

Posterior medians are summarized by grid width in Figure C.6. In some cases, when we

fit the joint model with no added grid points to data generated under measurement pattern 3, we encounter issues with convergence and the posterior medians take extreme values. Due to the range of the y-axis in Figure C.6, the instances in which posterior medians take extreme values do not appear in this figure. As such, we re-plot the results for this scenario (measurement pattern 3, no added grid points) separately in Figure C.7. For both setting 1 and 2 with measurement pattern 3, we find that adding grid points resolves this issue; the posterior medians are close to the truth when we fit the model using a grid of width 0.8 or 1.2, vs. no grid at all.

Coverage rates are summarized by grid width in Figure C.8.

C.4 Analysis of Data from Smoking Cessation Study

Our analytic sample consists of 238 individuals who did not lapse within the first 12 hours of the post-quit period (i.e., within 12 hours of 4am on the quit day) and who also responded to the emotion-related questions in at least one random EMA after the first 12 hours of the study. We also required that the baseline covariates of pre-quit smoking history and partner status be non-missing.

Definition of the Survival Outcome: In Vinci et al. (2017) [119], the authors analyze the same dataset to assess associations between position emotions and smoking habits after attempted quit. The authors conduct two analyses: (a) they assess the association between pre-quit positive emotions and a binary outcome of lapse on the first day after quit and (b) the association between post-quit positive emotions and the risk of first lapse (as a time-to-event outcome) among individuals who did not lapse on the first day after quit. In the first analysis, the authors used data from all individuals; in the second analysis, the authors restricted their final dataset to the subset of individuals who did not lapse on the first day.

We take the same subsetting approach here in order to avoid uncertainty surrounding the exact time of quit and reduce the number of lapse events that occur within minutes of the assumed quit time. We opt to use 12 hours as our cutoff, rather than 24, since we assume that the day of quit is known even if the exact time is not. A Kaplan-Meier curve for time-to-first-lapse for those who did not lapse within the first 12 hours is given in Figure C.9. Given that individuals who lapse almost immediately contribute limited information when fitting the model, we do not expect our results to be particularly sensitive to our exact definition of the time origin (e.g., 4pm (which we use) vs. 5pm vs. 7pm, for example). Sensitivity analyses could be conducted to better assess the impact of our assumed quit time on the fitted model.

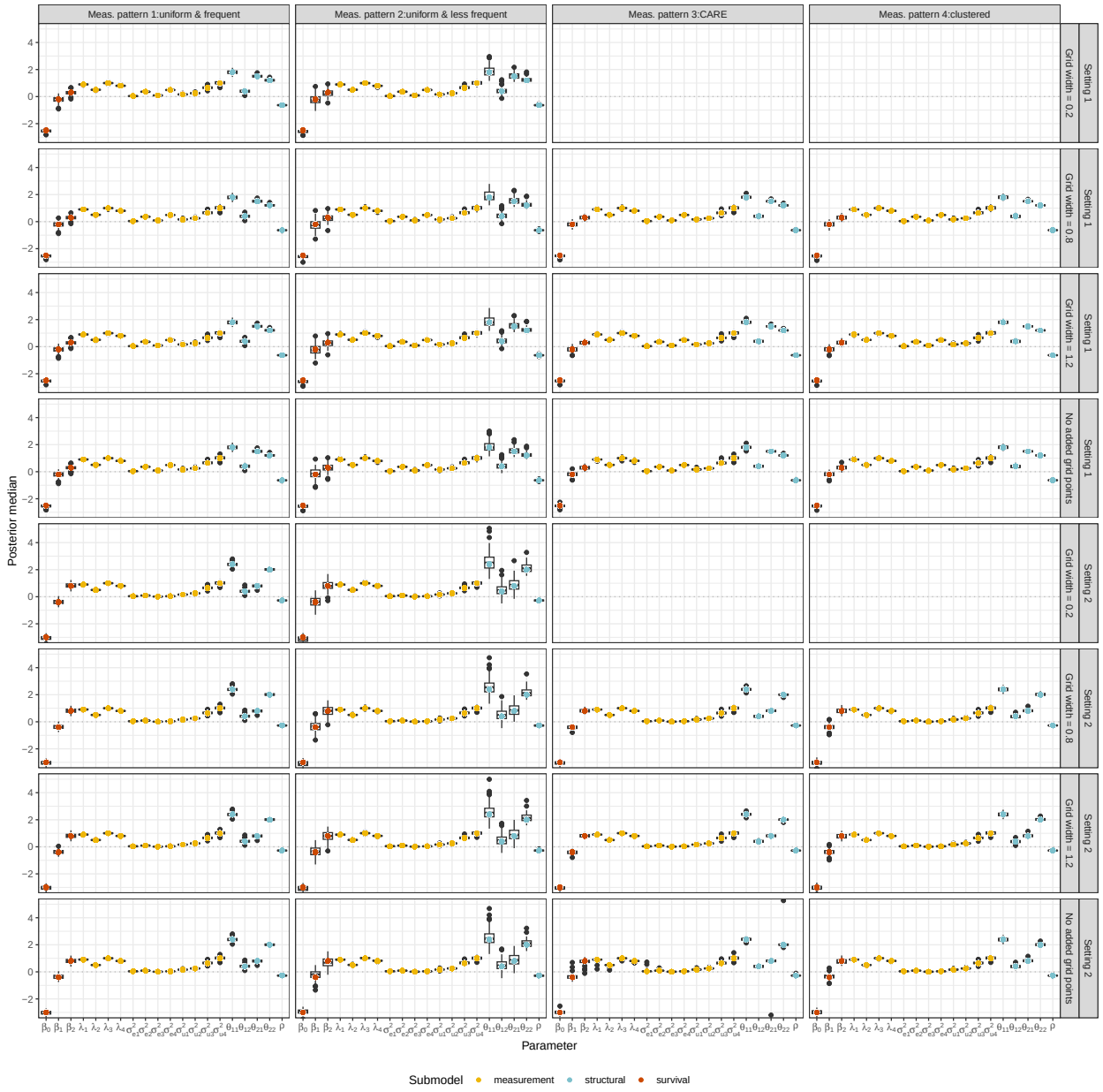


Figure C.6: For data generated under settings 1 and 2 with each of the four measurement patterns, we use box plots to summarize the distribution of the **posterior medians for all parameters** across the 100 simulated datasets. True parameter values are indicated with colored dots. Note that the y-axis range is truncated to span -3 to 5 and so some posterior medians corresponding to measurement pattern 3 and no added grid points are not shown; results for this combination of measurement pattern and grid width are re-plotted separately in Figure C.7.

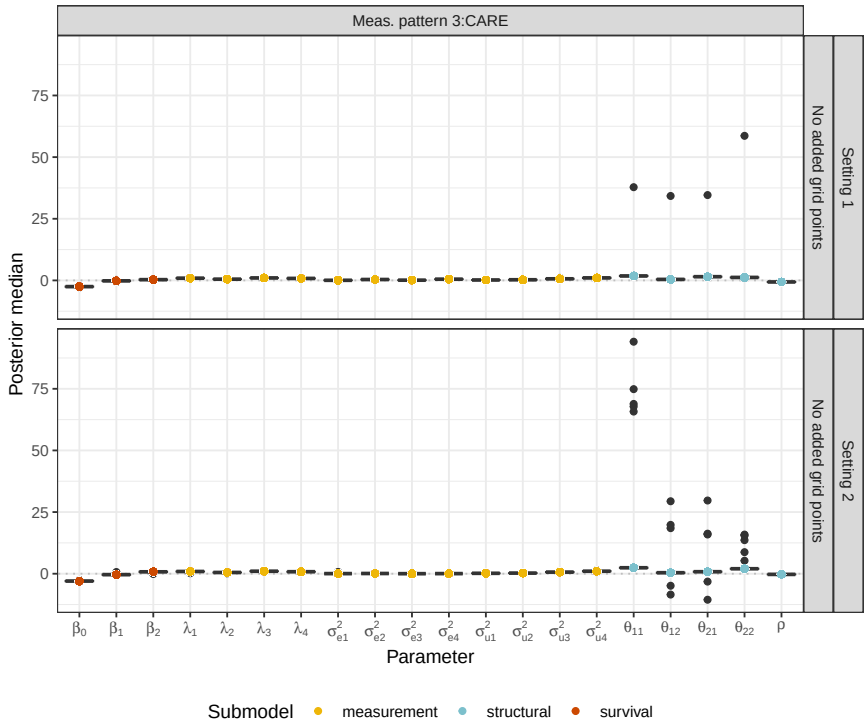


Figure C.7: For data generated under settings 1 and 2 with **measurement pattern 3**, we use box plots to summarize the distribution of the **posterior medians for all parameters** across the 100 simulated datasets when fitting the model **without added grid points**. True parameter values are indicated with colored dots. These plots show the same set of results as in Figure C.6 for measurement pattern 3 with no additional grid points, but here we allow a wider range of values on the y-axis so that all posterior medians are visible in the plot.

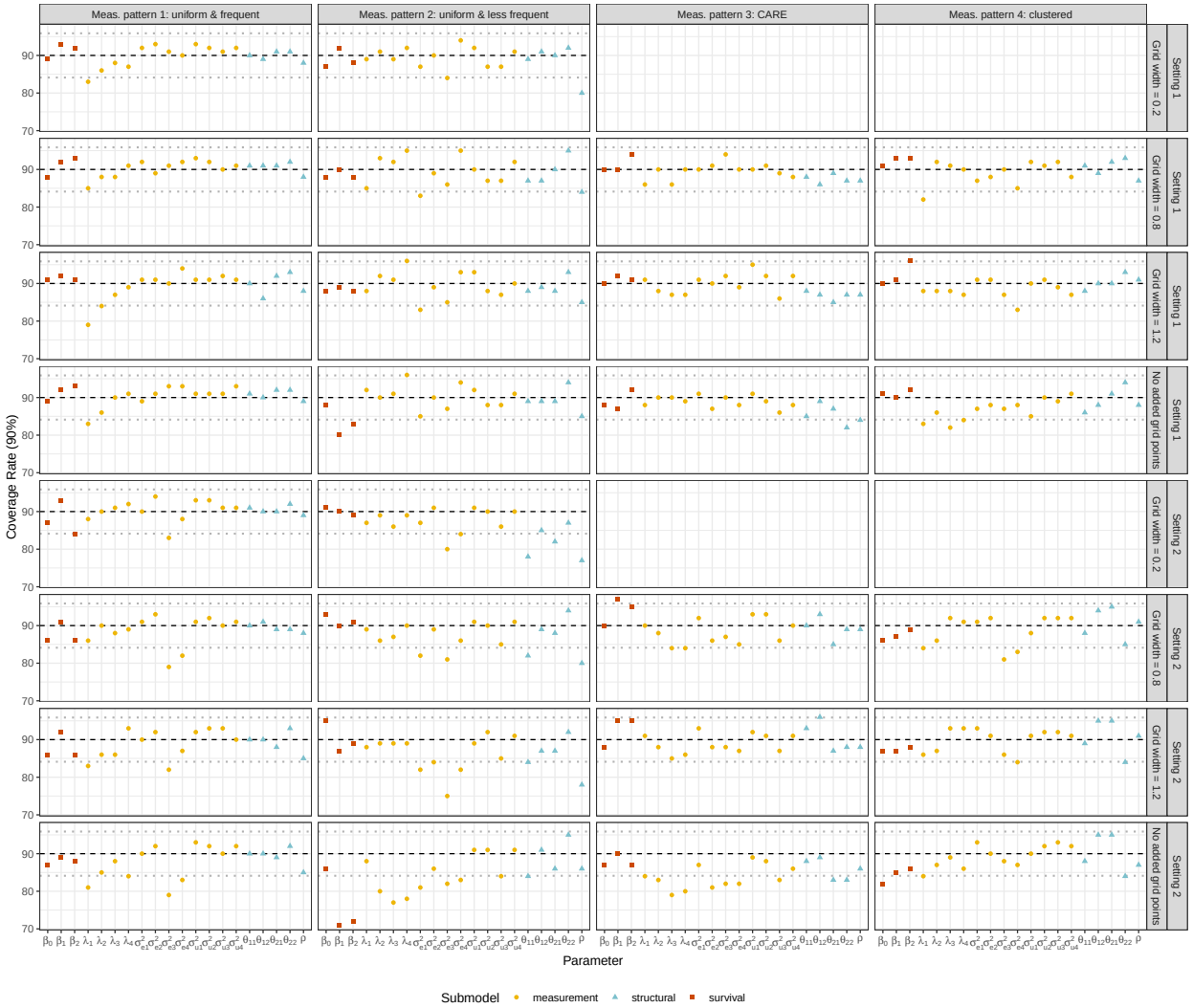


Figure C.8: For data generated under settings 1 and 2 with each of the four measurement patterns, we summarize the **coverage rate of 90% credible intervals** across the 100 simulated datasets with the colored dots. The black dashed line indicates target coverage and the grey dashed lines mark the expected range of values based on a 90% binomial proportion confidence interval.

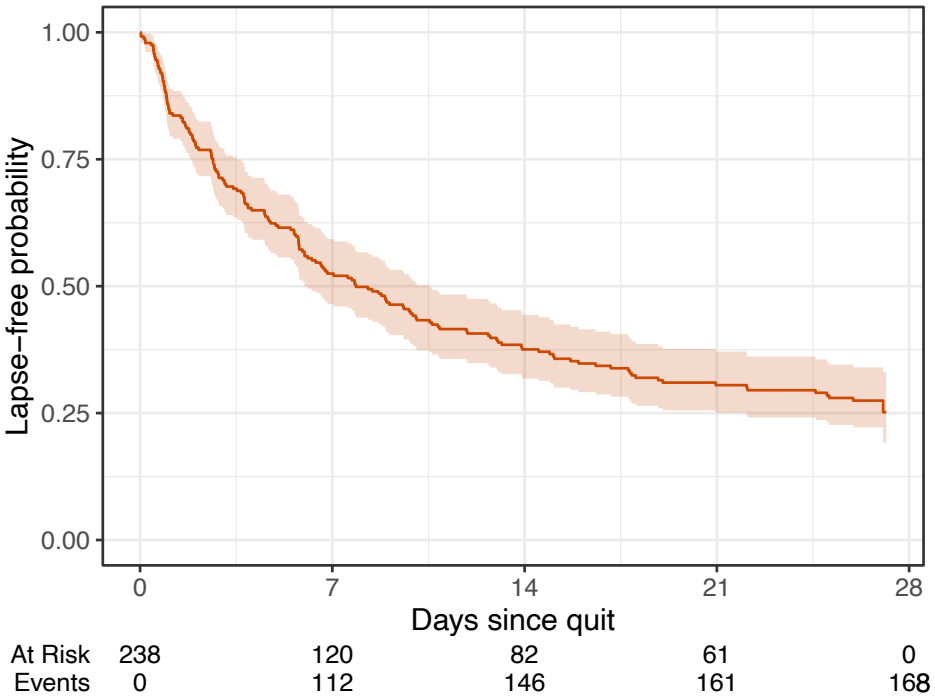


Figure C.9: The Kaplan-Meier curve for the time-to-event outcome of time until first lapse. The time origin corresponds to 4pm on the reported quit day.

C.4.1 Specification of Piecewise Constant Baseline Hazard

Lázaro et al. (2021) [56] describe a regularized version of a piecewise constant baseline hazard where the priors are specified such that the segments are correlated. The baseline hazard, made up of B segments, is

$$h_0(t;\beta_0) = \sum_{i=1}^{B} \beta_{0_b} \boldsymbol{I}(c_{b-1} < t \leq c_b)$$

where $\beta_0 = (\beta_{0_1}, ..., \beta_{0_B})$, $\boldsymbol{I}$ is an indicator function, $c_0 = 0$, and c_B is the time of the final (observed or censored) event time. We assume $B = 10$ and that the segments are of equal length. Then, prior distributions are specified as

$$
\begin{aligned}
p(log(\beta_{0_1})) &\sim N(0, \sigma_\beta^2) \\
p(log(\beta_{0_2})) &\sim N(log(\beta_{0_1}), \sigma_\beta^2) \\
&\vdots \\
p(log(\beta_{0_B})) &\sim N(log(\beta_{0_{B-1}}), \sigma_\beta^2)
\end{aligned}
$$

where $\sigma_\beta^2 \sim$ half-Cauchy$(0, 25)$, as suggested in Gelman (2006) [29].

C.4.2 Prior Distributions

We report prior distributions for the remainder of the parameters here:

$$
\begin{aligned}
&\lambda_k \sim \text{half-}N(1, \sigma_\lambda^2); k = 1, ..., 9 \\
&\sigma_\lambda \sim \text{half-Cauchy}(0, 5) \\
&\theta_{OU_{11}}, \theta_{OU_{21}}, \theta_{OU_{12}}, \theta_{OU_{22}} \sim N(0, 10^2) \\
&\rho \sim \text{Uniform}(-0.999999, 0.999999) \\
&\sigma_{u_k} \sim \text{half-}Cauchy(0, 5); k = 1, ..., 9 \\
&\sigma_{u_\epsilon} \sim \text{half-}Cauchy(0, 5); k = 1, ..., 9 \\
&\beta_1, \beta_2 \sim N(0, 5^2) \\
&\alpha_1, \alpha_2 \sim N(0, 5^2)
\end{aligned}
$$

C.4.3 Model Diagnostic Plots

For the joint model with the piecewise constant baseline hazard, we provide trace plots and posterior densities in Figure C.10 and C.11. We also provide a plot that evaluates the goodness-of-fit of our model in Figure C.12; this plot shows the distribution of the predicted survival probabilities. This approach to assessing goodness-of-fit uses a strategy similar to Cox-Snell residuals: We know that if $T \sim S(t)$ and $F(t) = 1 - S(t)$, then $F(T) \sim Unif(0, 1)$. So, for each set of posterior samples for $\hat{\beta}$ and $\hat{\eta}$ for all $i = 1, ..., N$, we:

1. Calculate $\hat{S}_i(T_i)$ using $\hat{\beta}$ and $\hat{\eta}$ where T_i is the observed event time
2. Fit a Kaplan-Meier curve to $(1 - \hat{S}_i(T_i), \delta_i)$

If our predicted survival probabilities are accurate, they should be approximately uniformly distributed between 0 and 1 and so the Kaplan-Meier curve should follow a diagonal line from $(1, 1)$ to $(1, 0)$.

We compared the model with the piecewise constant baseline hazard to a model with a Weibull baseline hazard function and found that the piecewise constant baseline hazard appeared to result in a model that better fit the data. The goodness-of-fit plot for the fitted Weibull model is given in Figure C.13.

C.4.4 Plotting Correlation Decay in the Latent Factors

When estimating the OU process, we rely on the conditional distribution. To plot the estimated correlation decay in the latent factors across increasing time intervals (as in Figure 4.5), we use the (unconditional) covariance formula. We provide this marginal covariance formula below. Assuming that $\eta(s)$ and $\eta(t)$ are two observation of latent factors from an OU process at times s and t, where $s < t$, then the marginal joint distribution is:

$$\begin{bmatrix} \boldsymbol{\eta}(s) \\ \boldsymbol{\eta}(t) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{V} & \boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t-s)} \\ e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{V} & \boldsymbol{V} \end{bmatrix} \right)$$

To calculate the estimated correlation between the latent factors across increasing time intervals, we plug posterior samples of the structural submodel parameters into the off-diagonal elements of $\boldsymbol{\Psi}$ along with $s = 0$ and increasing values of t. Because our identifiability constraint assumes that we model the OU process on the correlation scale, the covariance matrix above will be the correlation matrix here. In Figure 5 in the main manuscript, the x-axis corresponds to increasing values of t and the y-axis corresponds to different elements of $\boldsymbol{\Psi}$.

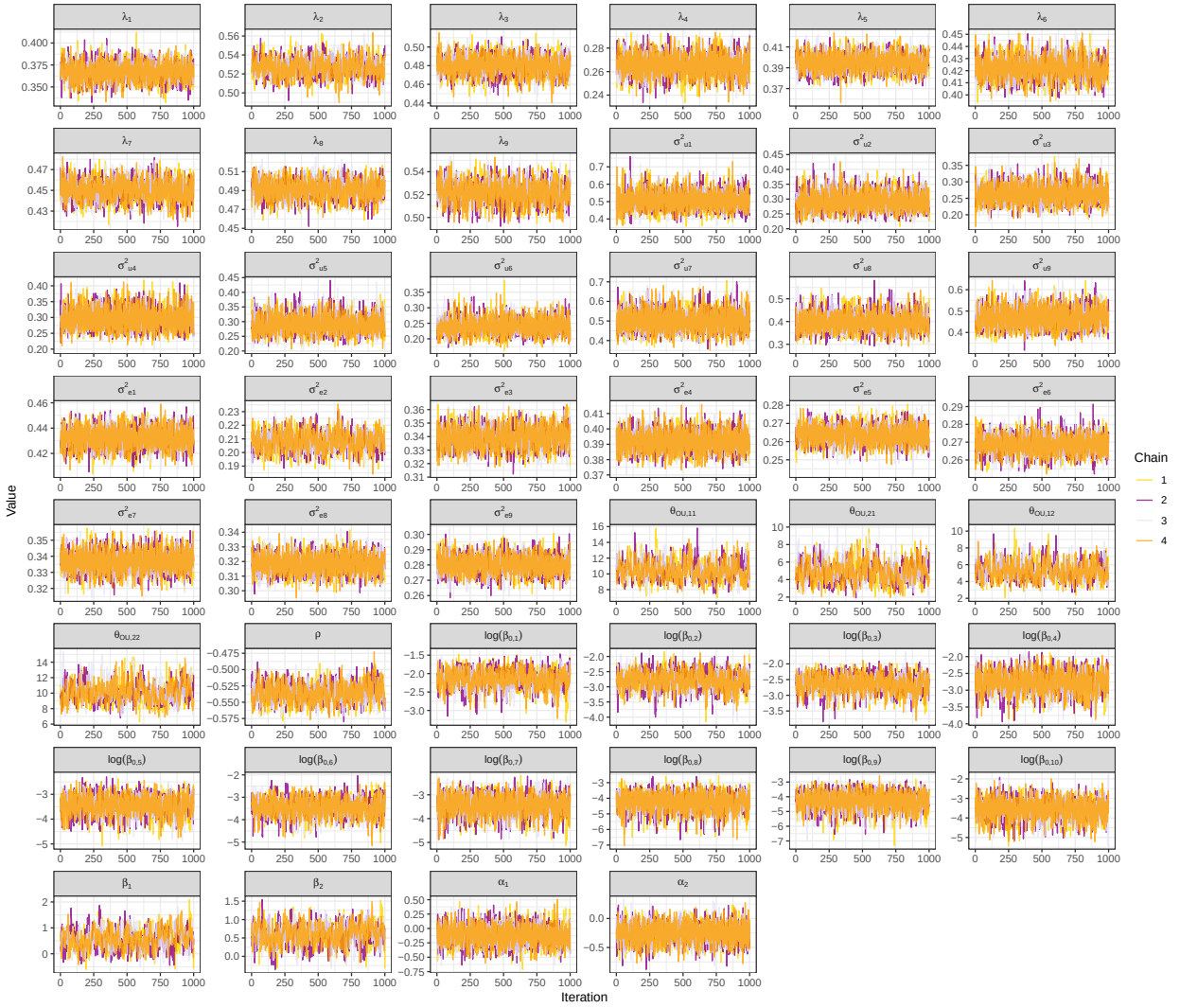


Figure C.10: Trace plot of posterior samples (after burn-in) for the joint model with the **piecewise constant baseline hazard** fit to data from the mHealth smoking cessation study.

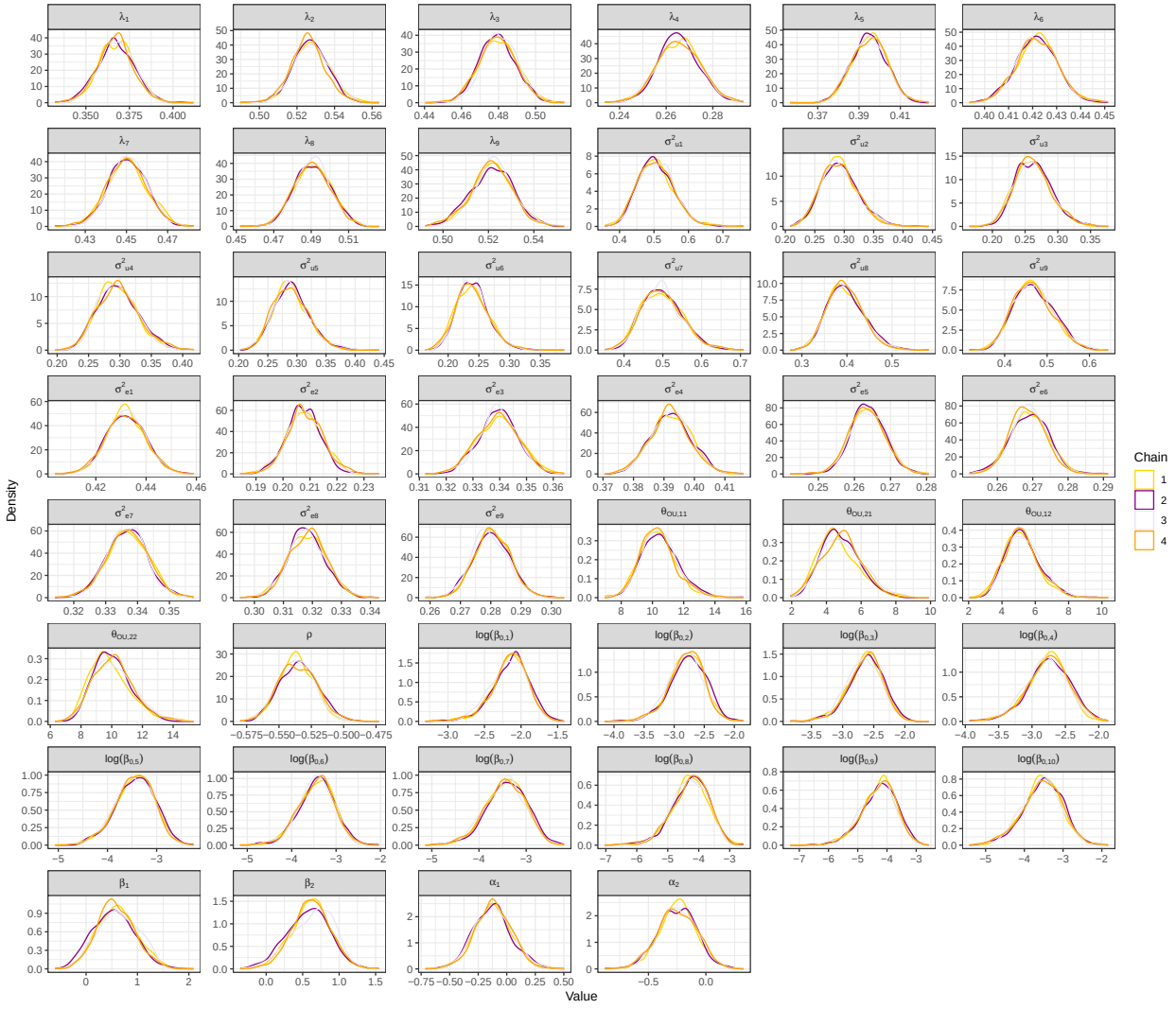


Figure C.11: Posterior densities of parameters for the joint model with the **piecewise constant baseline hazard** applied to data from the mHealth smoking cessation study.

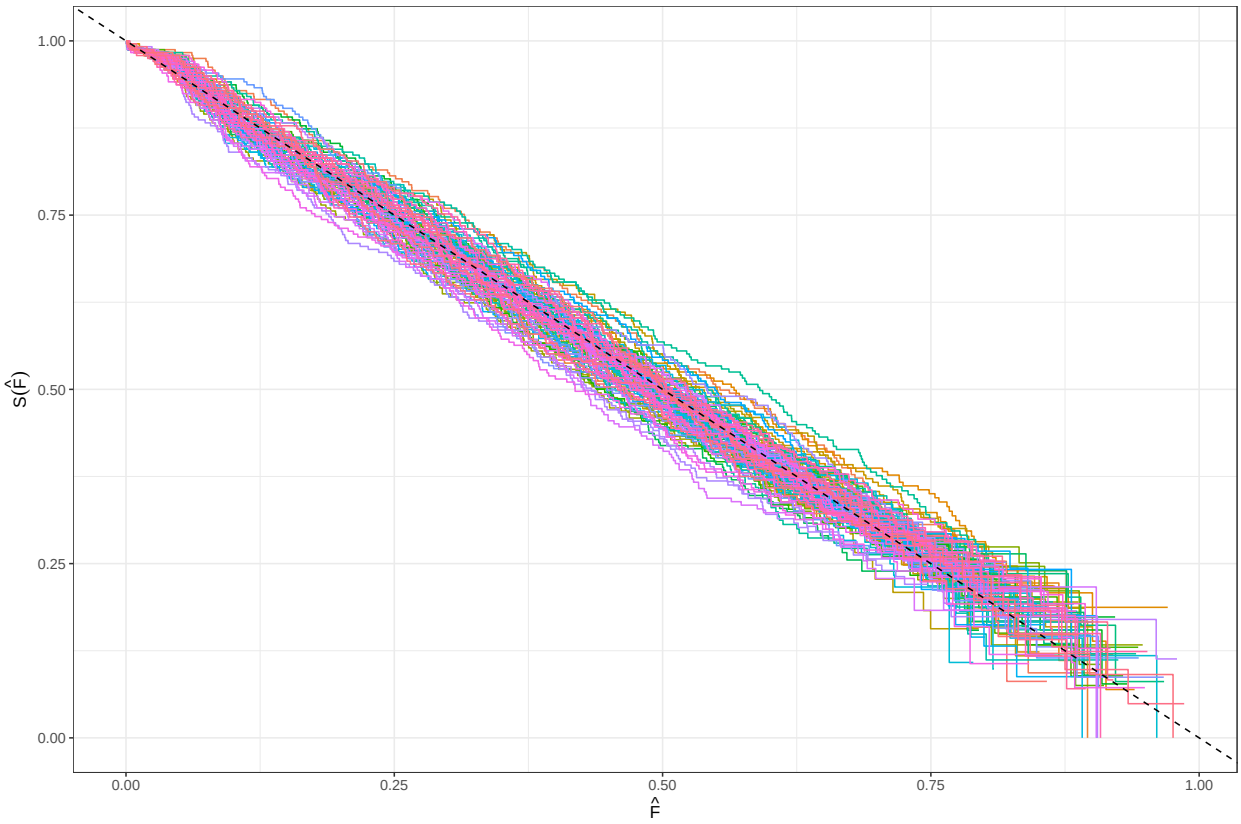


Figure C.12: Goodness of fit for the joint model with the **piecewise constant baseline hazard** in the survival submodel. Each solid line corresponds to the Kaplan-Meier survival curve calculated from a single set of posterior samples; curves are plotted for 100/1000 total posterior samples samples. If the model fits well, then we expect the Kaplan-Meier curves to follow the dashed line going from $(1, 1)$ to $(1, 0)$.

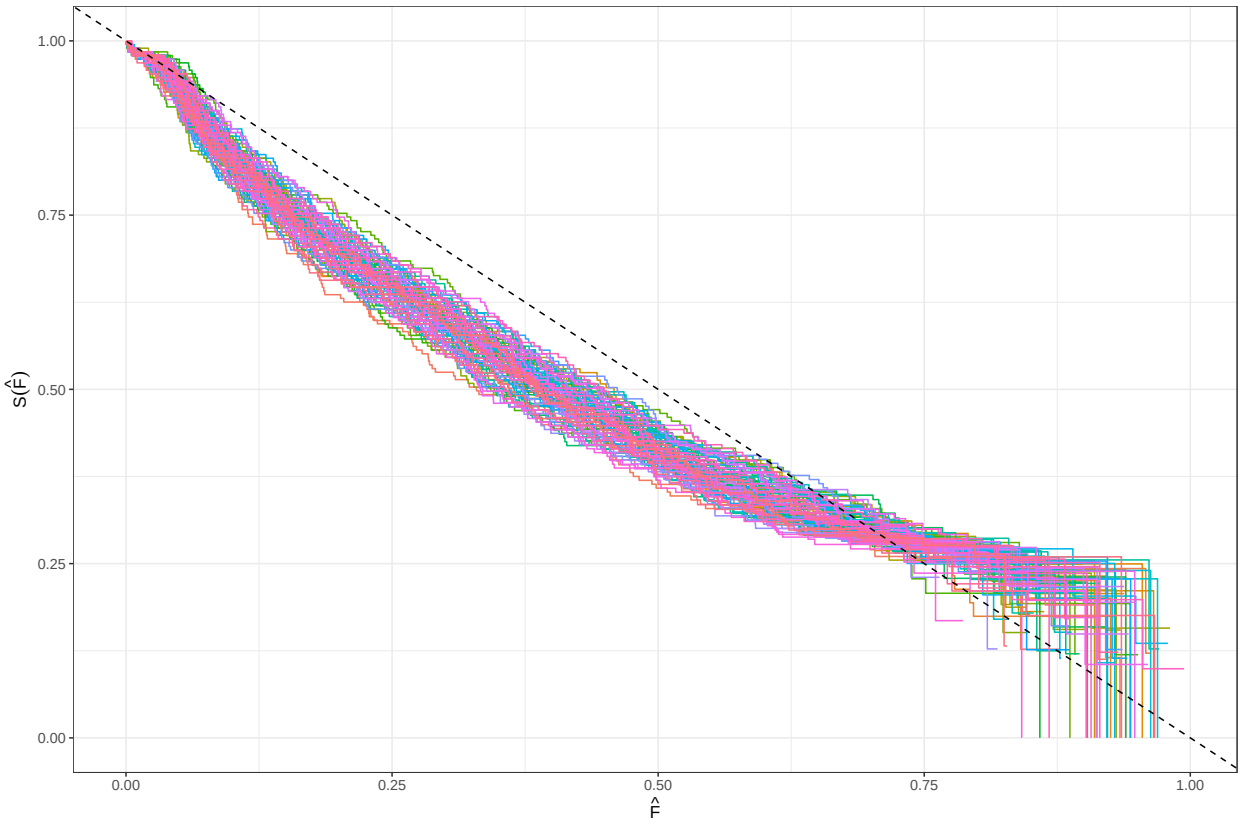


Figure C.13: Goodness of fit for the joint model with the **Weibull baseline hazard** in the survival submodel. Each solid line corresponds to the Kaplan-Meier survival curve calculated from a single set of posterior samples; curves are plotted for 100/1000 total posterior samples samples. If the model fits well, then we expect the Kaplan-Meier curves to follow the dashed line going from $(1, 1)$ to $(1, 0)$.

APPENDIX D

Supplementary Material for: Estimation of Time-Varying Treatment Effects in a Joint Model for Longitudinal and Recurrent Event Outcomes in Mobile Health Data

D.1 Case Study

D.1.1 Defining the Recurrent Event Outcomes

Poly-substance use is defined as engaging in any of the following activities: using marijuana, vaping, or smoking cigarettes. At each EMA, participants are asked to respond to a set of questions in which they report if they used any of these substances since the last EMA, and if so, the approximate time of use. More details on the questions used to assess each type of substance use, and our approach to converting these responses into event times suitable for modeling, are given below. Note that these rules are defined specifically for the purpose of curating an event-time outcome appropriate for modeling, and are not intended to be used to draw general conclusions about comparable instances of substance use (for example, cigarette vs. vaping equivalence).

Marijuana Use If participants report any marijuana use since the prior EMA, they are also asked to provide a single time corresponding to time of use. Each reported time of marijuana use is considered a recurrent event.

Cigarette Smoking If participants report smoking any cigarettes since the prior EMA, then they are prompted to respond to additional questions about how many cigarettes they smoked and approximately when they smoked the cigarettes.

If a participant reports smoking a partial or a single cigarette since the last EMA, then they are also asked to provide a single time corresponding to when they smoked. We consider this time as a recurrent event time.

If a participant reports smoking more than one cigarette since the last EMA, then they are asked to respond to two additional questions. These questions ask (i) when they smoked the first cigarette after the last EMA and (ii) when they smoked their most recent cigarette. These questions result in an interval of time over which a participant has smoked a known number of cigarettes. To convert this interval into a recurrent event time appropriate for modeling, we evenly distribute the reported number of cigarettes across the interval and then consider each time at which a cigarette is assumed to be smoked as a recurrent event time. For example, if two cigarettes were smoked over an interval from A to B, then we would place one event at time A and one event at time B. If three cigarettes were smoked over an interval from A to B, then we would place one event at time A, one event at time B, and one event halfway between times A and B. If more than 10 cigarettes were reported smoked, then participants do not report the exact number but instead select the option of "more than 10 cigarettes" when filling out the EMA. For cases when "more than 10 cigarettes" was selected, we assume that the cigarettes were smoked (and recurrent events occurred) at a rate of approximately 1 event per hour across the reported interval of time.

In some cases, an individual might smoke multiple cigarettes in a row, which could be viewed as a single episode of smoking, rather than multiple separate episodes. To account for cases like these, we define an additional rule: if multiple cigarettes were reported smoked over an interval of less than an hour, then we consolidate these events into a single event and use the midpoint of the interval of time as the corresponding recurrent event time.

Vaping Vaping is assessed in a similar way to cigarette smoking; however, vaping is reported in units of "puffs". Before converting the responses to the vaping-related questions into recurrent event times, we first consider how many puffs constitute a single event. One puff delivers much less nicotine than a single cigarette, and so a single puff is not equivalent to a single cigarette. The nicotine contained in one e-cigarette pod is approximately equivalent to the nicotine contained in a pack of cigarettes (20 cigarettes); it also takes approximately 200 puffs to use up a pod [75]. Using these two facts, we assume that on average 10 puffs are equivalent to smoking one cigarette, and we define a single event of vaping as 0-15 puffs. It follows that 16-25 puffs are 2 events, 26-35 puffs are 3 events, and so on. The conversion of puffs to events is summarized in Table D.1. After converting puffs to events, we next apply some rules to convert the information reported in the EMAs into recurrent event times.

# puffs	count	# events	count
1	18		
2	6		
3	5		
4	1		
5	8	1	47
6	4		
7	2		
9	1		
10	1		
11	1		
20	4	2	5
21	1		
27	2	3	2
40	1		
41	1	4	3
43	1		
52	1	5	1
100+	1	12	1

Table D.1: Number of times that each number of puffs was observed and conversion of puffs to events.

Vaping is assessed using questions that have the same structure as those used to assess cigarette smoking since the last EMA. Participants are asked to report the total number of puffs since the prior EMA, the approximate time of the puff if only a single puff was taken, and the approximate time of the first and most recent puff if multiple puffs were reported. After converting puffs to events, we apply rules similar to those used when converting cigarettes to event times. When more than 1 but fewer than 16 puffs are reported, we convert these puffs into a single event; the approximate time of the event corresponds to the midpoint of the reported interval over which the puffs were taken.

In the illustrative analysis in the main paper, we combine all instances of marijuana use, vaping, and cigarette smoking into a single recurrent event outcome called poly-substance use. We do not distinguish between use of specific substances in our analysis. In Figure D.1, we plot the timing of the recurrent poly-substance use events for each individual in the study.

In Figure D.2, we plot the mean cumulative function (MCF) estimate across days in the study. This plot tells use the expected cumulative number of events per person by each day in the study.

We can also look at the expected cumulative number of events per person for the pre- and post-quit periods separately, as shown in Figure D.3.

D.1.2 Specifying the Hazard Model

In the analysis of the motivating MRT data, our hazard model is

$$h_{ir}(t) = h_0(t) \exp\left[\beta_1 \eta_{1i}(t) + \beta_2 \eta_{2i}(t) + \tilde{\mu}_i(t)\right]$$

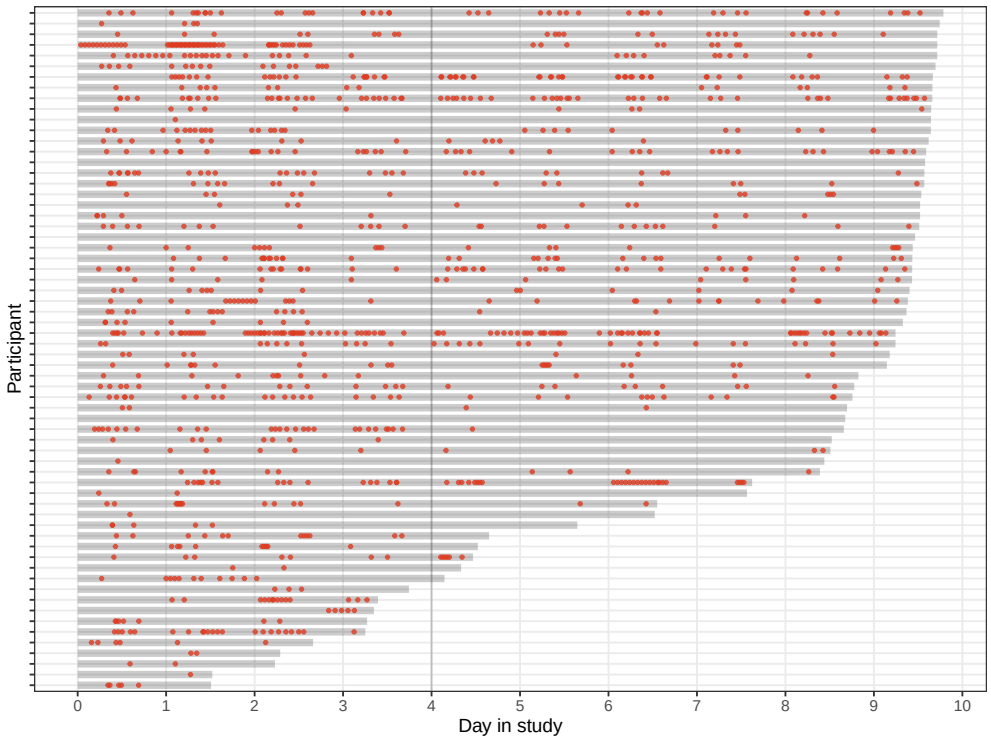


Figure D.1: Timing of recurrent poly-substance use events. Grey bars indicate time periods when individuals are at risk of a recurrent event. The vertical grey line at day 4 indicates the end of the quit day, which we use as the transition from the pre-quit to the post-quit period.

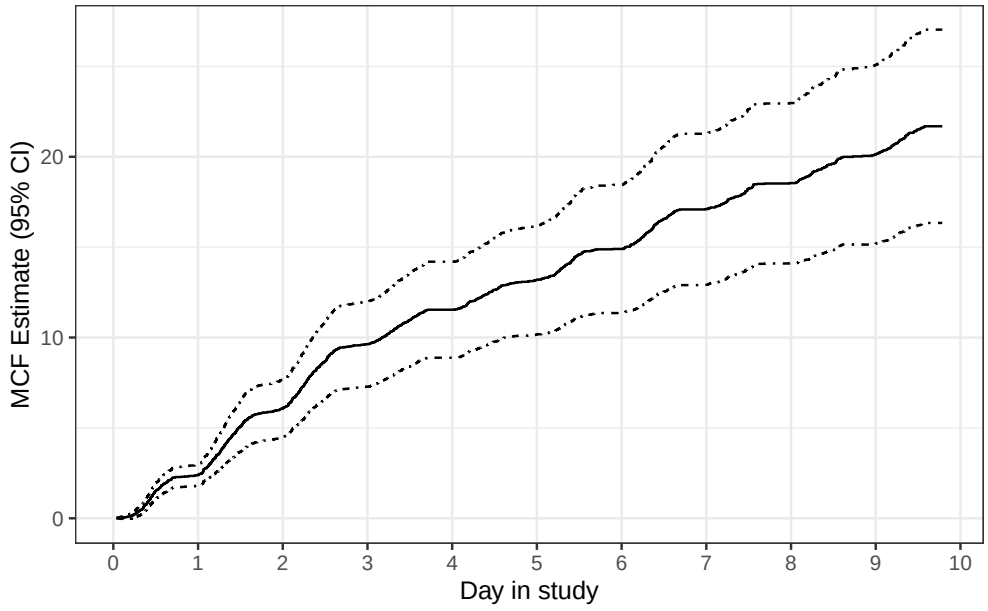


Figure D.2: Mean cumulative function for recurrent poly-substance use.

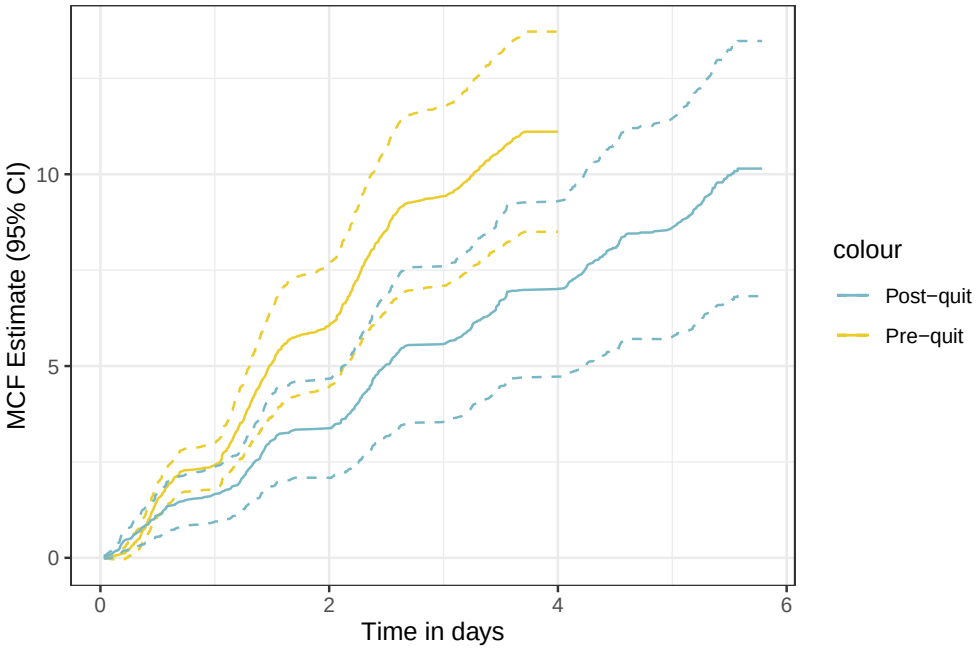


Figure D.3: Mean cumulative function for recurrent poly-substance use by pre- and post-quit periods. For the pre-quit period, time 0 corresponds to time since the start of the study. For the post-quit period, time 0 corresponds to time since the end of the designated quit day (day 4).

The baseline hazard is assumed to be Weibull, with separate parameters for the pre-quit vs. post-quit period. Participants are instructed to quit on day 4 and so we use the end of day 4 as the transition between the pre-quit and post-quit periods. To allow for period-specific parameters, we specify the baseline hazard as

$$h_0(t) = \beta_{0,\text{pre}}\gamma_{\text{pre}}t^{\gamma_{\text{pre}}-1} \times \text{I}(\text{study day } < 4 \text{ days}) + \beta_{0,\text{post}}\gamma_{\text{post}}t^{\gamma_{\text{post}}-1} \times \text{I}(\text{study day } \geq 4 \text{ days})$$

where, because we are using the clock-reset approach, time t corresponds to time since the most recent event in the baseline hazard $h_0(t)$. The latent factors, $\eta_{1i}(t)$ and $\eta_{2i}(t)$, are still functions of time since the start of the study.

In one version of the hazard model, we assume that the treatment effect parameter in the treatment model for the hazard is constant throughout the study (i.e., we have one treatment parameter in the hazard model, $\tilde{\tau}$.) In another version of the hazard model, we allow the treatment model in the hazard to differ across the pre- and post-quit periods by re-defining the treatment function $\tilde{\mu}_i(t)$ as

$$\begin{aligned}\tilde{\mu}_i(t) = &\sum_{t_{ia} \in \mathcal{A}_i(t)} \tilde{\tau}_{\text{pre}} \left(1 - \frac{t - t_{ia}}{\delta_a}\right)_+ \times \text{I}(\text{study day } < 4 \text{ days}) \\ &+ \sum_{t_{ia} \in \mathcal{A}_i(t)} \tilde{\tau}_{\text{post}} \left(1 - \frac{t - t_{ia}}{\delta_a}\right)_+ \times \text{I}(\text{study day } \geq 4 \text{ days})\end{aligned}$$

In $\tilde{\mu}_i(t)$, time t is the time since the start of the study.

D.1.3 Additional Results from Fitting the Joint Longitudinal Recurrent Event Model

We fit four different versions of the joint model, each of which corresponds to a different combination of treatment effect model in the longitudinal submodel (i.e., the additive or drift version of the treatment effect model) and hazard model (i.e., a hazard model that assumes the effect of treatment on the hazard is the same across all 10 days of the study or a hazard model that allows the effect of the treatment on the hazard to differ across the pre- vs. post-quit period. Trace plots for each model are provided in Figures D.4–D.7.

D.1.3.1 Visualizing Decay in Correlation Over Time for the Latent Process

We can plot the decay in auto- and cross-correlation for the bivariate OU process estimated from the joint models. These plots, shown in Figure D.8, are created using the posterior means of $\boldsymbol{\theta}$ and ρ.

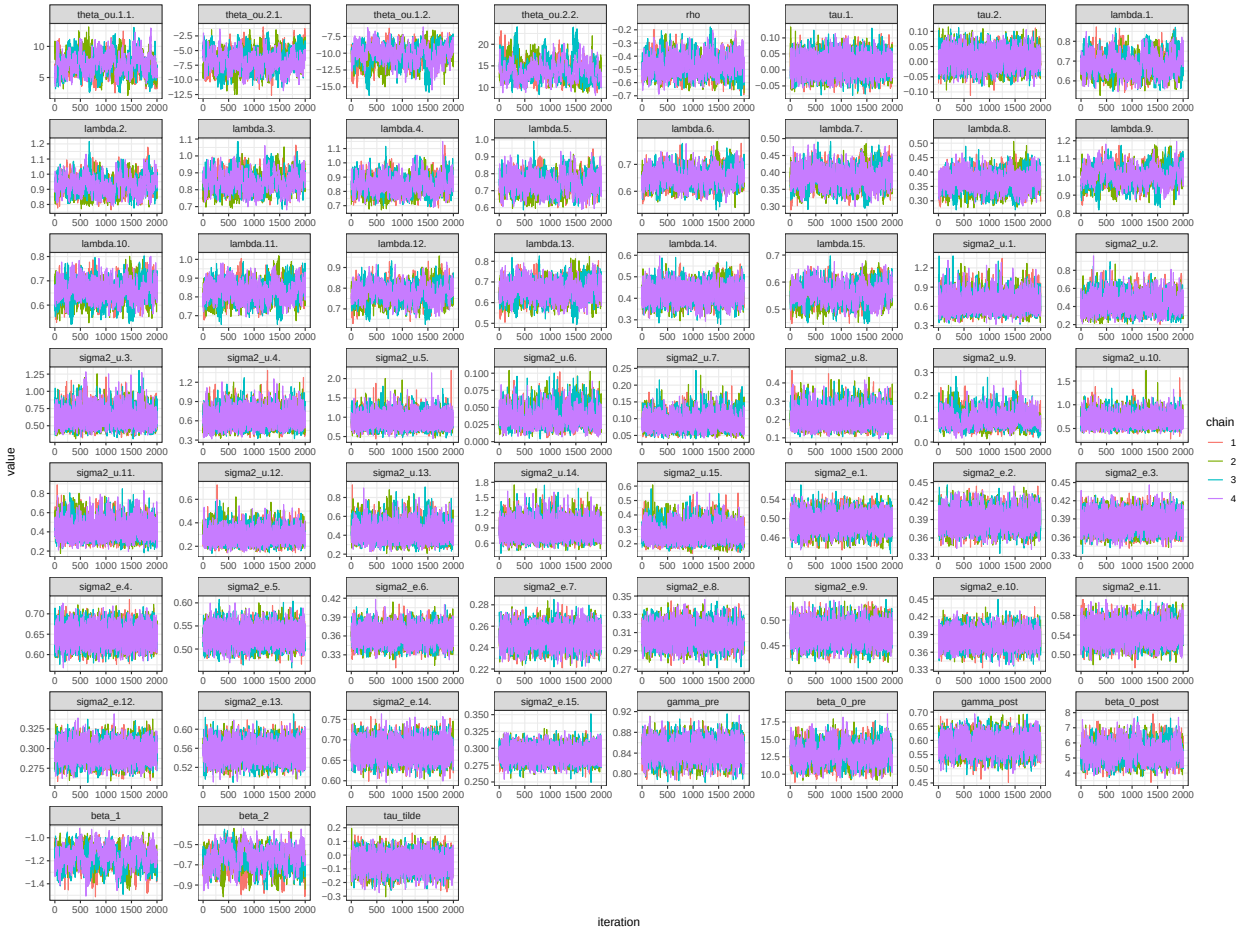


Figure D.4: Trace plot for the joint model with an **additive** model for treatment effect in the longitudinal submodel and with a **single** treatment parameter in the hazard model.

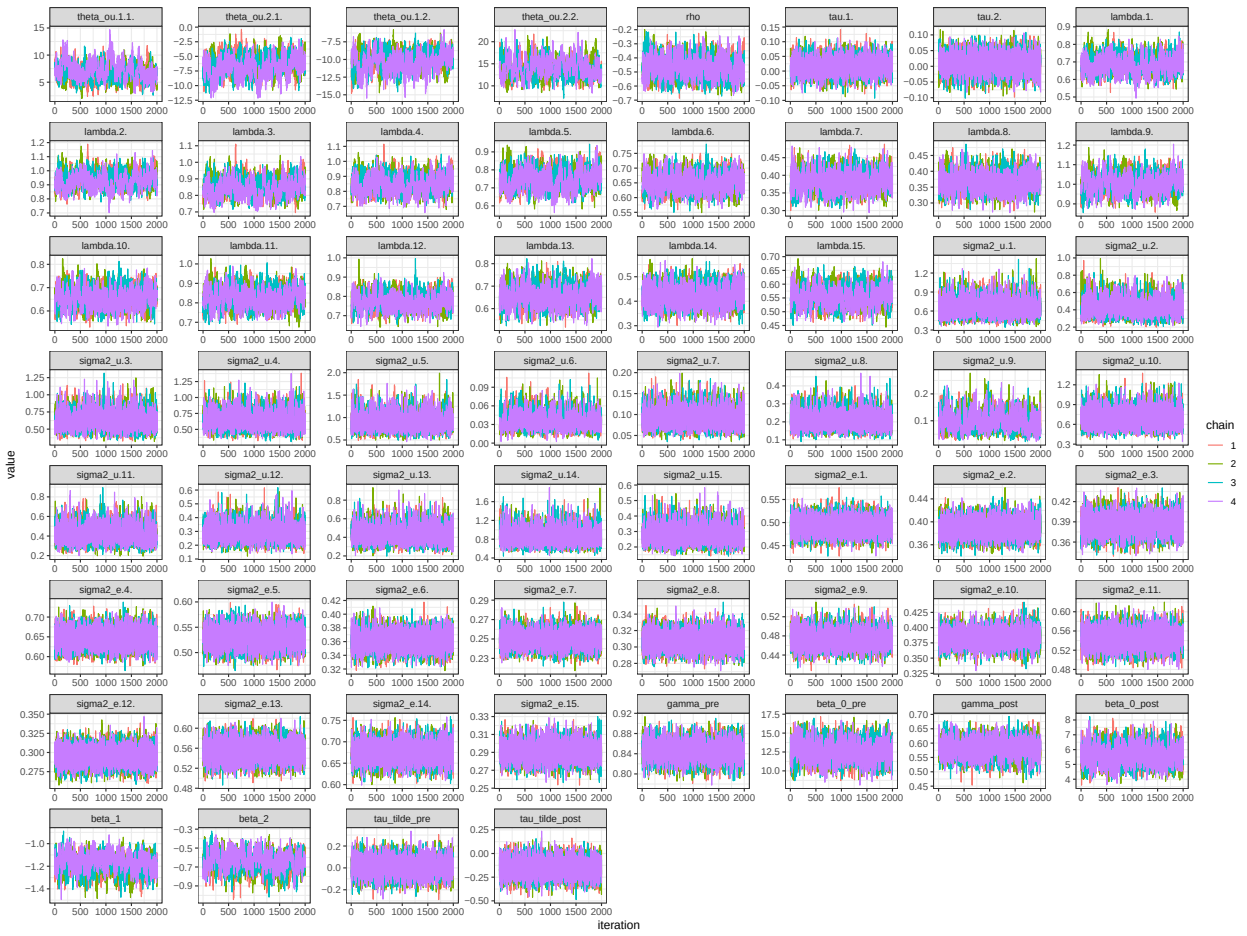


Figure D.5: Trace plot for the joint model with an **additive** model for treatment effect in the longitudinal submodel and with **separate** pre-quit and post-quit treatment parameters in the hazard model.

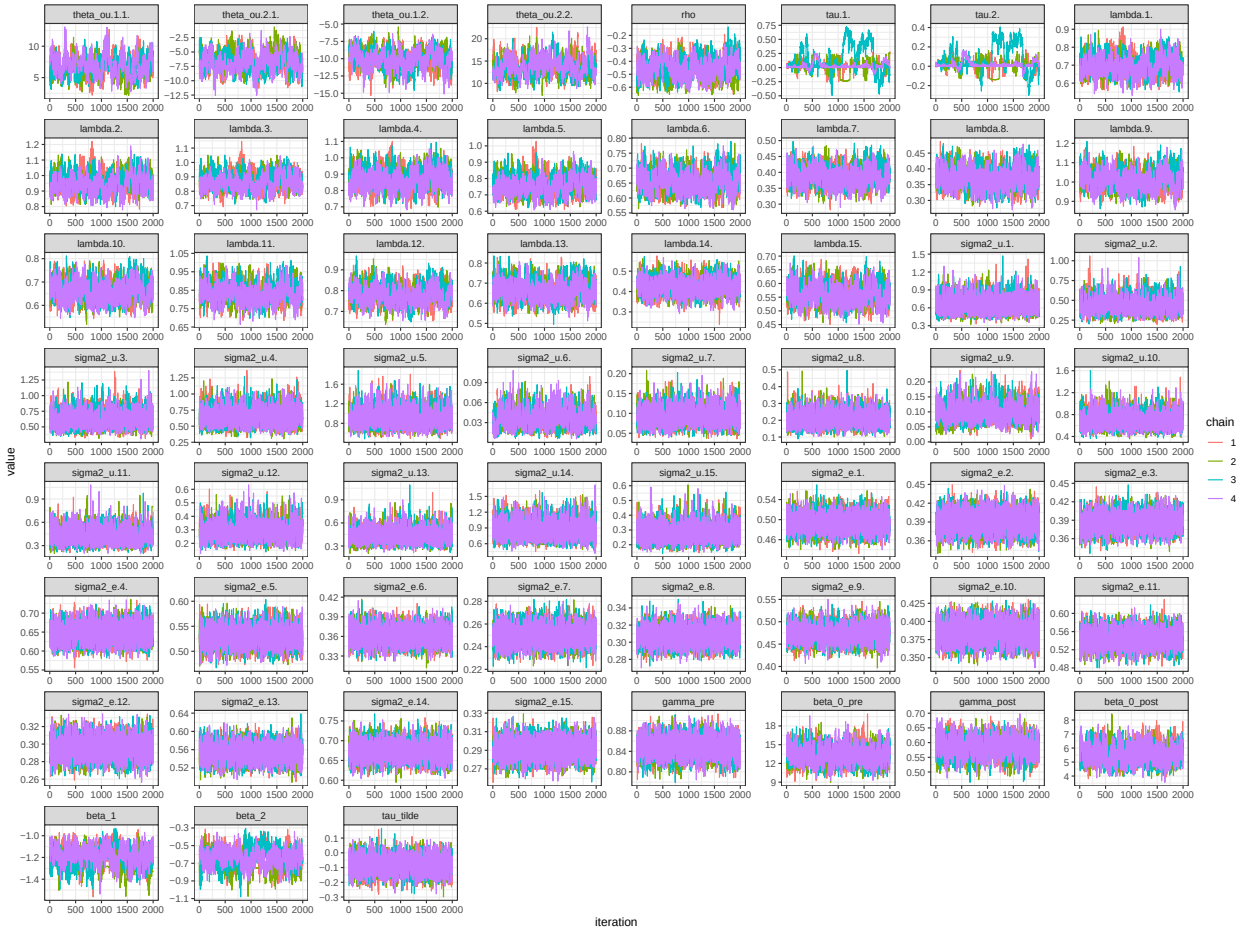
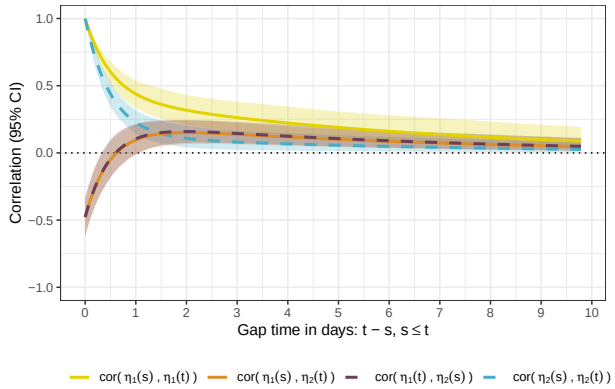
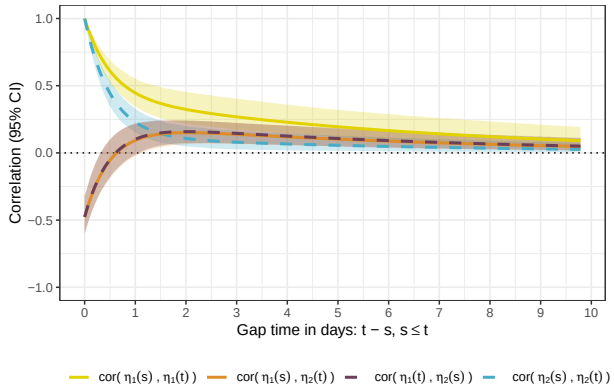


Figure D.6: Trace plot for the joint model with a **drift** model for treatment effect in the longitudinal submodel and with a **single** treatment parameter in the hazard model.

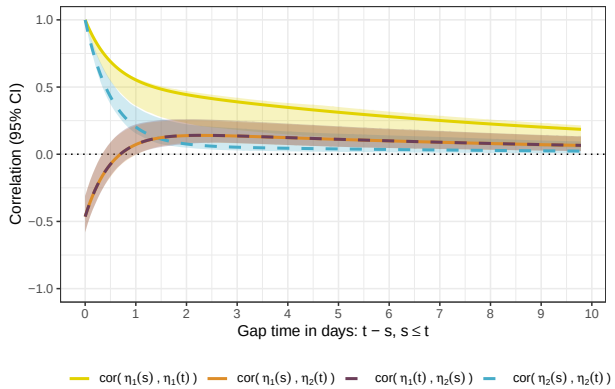
Figure D.7: Trace plot for the joint model with a **drift** model for treatment effect in the longitudinal submodel and with **separate** pre-quit and post-quit treatment parameters in the hazard model.



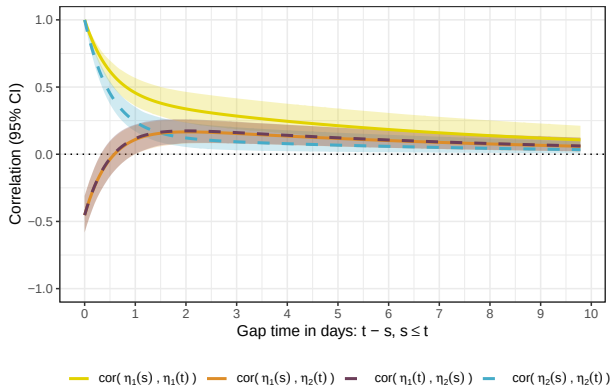
(a) Model assumes an **additive** treatment effect on the latent process and a **single** treatment parameter in the hazard submodel.



(b) Model assumes an **additive** treatment effect on the latent process and **separate** pre- and post-quit treatment parameters in the hazard submodel.



(c) Model assumes a **drift** treatment effect on the latent process and a **single** treatment parameter in the hazard submodel.



(d) Model assumes a **drift** treatment effect on the latent process and **separate** pre- and post-quit treatment parameters in the hazard submodel.

Figure D.8: Decay in cross- and auto-correlation in bivariate OU process from fitted joint models. For the different joint models, each plot shows the estimated decay in autocorrelation and cross-correlation between latent factors that represent positive affect ($\eta_1(t)$) and negative affect ($\eta_2(t)$) across increasing gap times, where time is on the scale of days.

D.1.3.2 Visualizing Estimates from the Treatment Effect Models

The treatment-related parameters $\boldsymbol{\tau}$ in the longitudinal submodel have different interpretations when modeled as an additive effect vs. as drift and so to compare the estimates of $\boldsymbol{\tau}$ between models that assume different impacts of treatment, we can plot the terms that are added to the conditional mean of the OU process as a result of the treatment models. That is, when treatment is modeled as an additive effect, the conditional expectation of the latent process is:

$$\mathbb{E}\left[\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s)\right] = e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) + \boldsymbol{\mu}_i(t) - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\mu}_i(s).$$

When treatment is modeled in the drift term, the conditional expectation of the latent process is:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s)\right] =& e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) + \\ & \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \left[\left(1 - \frac{u-t_{ia}}{\delta_a}\right)e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\theta}^{-1} + \frac{1}{\delta_a}e^{-\boldsymbol{\theta}(t-u)}\right]\boldsymbol{\tau}\Bigg|_{u=max(t_{ia},s)}^{u=min(t,t_{ia}+\delta_a)}.\end{aligned}$$

To visualize the estimated effect of treatment on the latent process, we plot

$$\boldsymbol{\mu}_i(t) - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\mu}_i(s)$$

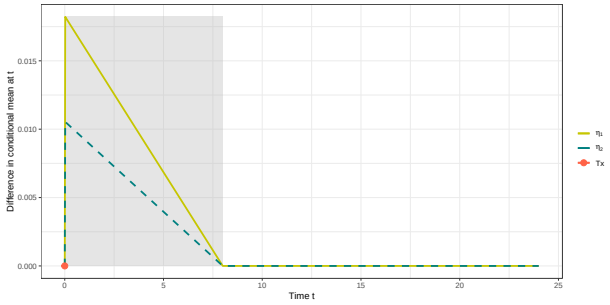
and

$$\sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \left[\left(1 - \frac{u-t_{ia}}{\delta_a}\right)e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\theta}^{-1} + \frac{1}{\delta_a}e^{-\boldsymbol{\theta}(t-u)}\right]\boldsymbol{\tau}\Bigg|_{u=max(t_{ia},s)}^{u=min(t,t_{ia}+\delta_a)}$$

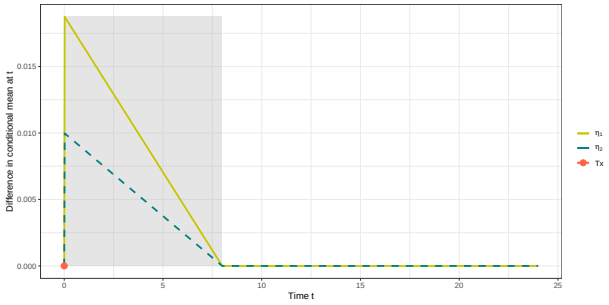
, which correspond to the terms added to the conditional mean of the OU process as a result of our additive and drift treatment effects, respectively. When plotting these terms, we assume that a single treatment was delivered at time $s = 0$. Plots of these treatment-related terms are given in Figure D.9.

D.1.3.3 Variability in the Longitudinal Outcomes

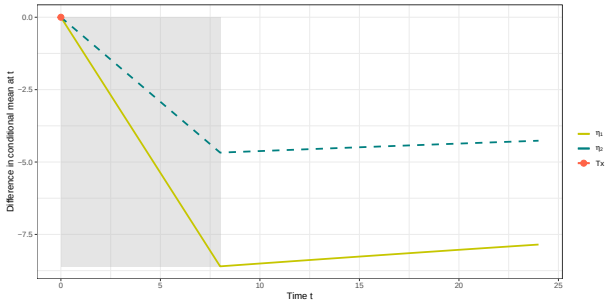
The posterior estimates of parameters from the longitudinal submodel show a fair amount of variability in the random intercept variance and measurement error variance components. We can compare these estimates to the variability that we see in the observed longitudinal outcomes. Table D.2 summarizes the average within-individual variance in the reported emotions and the variance of the within-individual means of reported emotions. The rows are sorted by the rightmost column, corresponding to variance of the within-individual means.



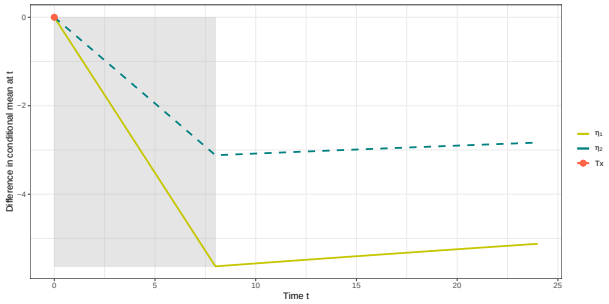
(a) Model assumes an **additive** treatment effect on the latent process and a **single** treatment parameter in the hazard submodel.



(b) Model assumes an **additive** treatment effect on the latent process and **separate** pre- and post-quit treatment parameters in the hazard submodel.



(c) Model assumes a **drift** treatment effect on the latent process and a **single** treatment parameter in the hazard submodel.



(d) Model assumes a **drift** treatment effect on the latent process and **separate** pre- and post-quit treatment parameters in the hazard submodel.

Figure D.9: Estimated treatment-related terms from fitted joint models. For each fitted model, the plots show the estimated value of the difference between the fitted latent process and an OU process with a constant mean of 0, if a single treatment were to be sent at time 0. If the treatment effect is modeled via a time-varying drift term, then this difference is $\int_0^t e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\mu}_i(u)du$. If the treatment effect is modeled via an additive shift, then this term is $\boldsymbol{\mu}_i(t) - e^{-\boldsymbol{\theta}t}\boldsymbol{\mu}_i(0)$

Model index	Emotion	Mean of ind.-specific variances	Var. of ind.-specific means
6	angry	0.537	0.258
7	ashamed	0.290	0.287
8	guilty	0.396	0.439
15	hopeless	0.436	0.542
9	irritable	0.843	0.718
2	happy	0.646	0.796
1	grateful	0.770	0.801
12	sad	0.509	0.853
13	restless	0.749	0.989
14	bored	0.798	1.002
11	anxious	0.769	1.090
3	proud	0.622	1.096
10	lonely	0.536	1.105
4	relaxed	0.859	1.151
5	enthusiastic	0.729	1.215

Table D.2: Empirical variability in measured longitudinal outcomes in the motivating MRT.

D.2 Different Pathways for Treatment Effect

The treatment effect could potentially be incorporated into our joint model in three different ways. As described in the main paper, treatment could (a) directly impact the latent process, (b) alter the risk of recurrent events, or (c) change the measured longitudinal outcomes. These potential pathways for treatment effect are summarized in Figure D.10. We only consider models for (a) and (b).

D.3 Modeling the Impact of Treatment on the Latent Process

If we model the impact of treatment on the latent process by incorporating a time-varying drift term into the OU process (resulting in a process also known as a Hull-White process), then the solution to the SDE involves a term with an integral over the treatment effect function $\boldsymbol{\mu}_i(t)$. As given in the main text, the conditional distribution of $\boldsymbol{\eta}_i(t)$ with time-dependent drift is

$$\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s) \sim N_p\left(e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) + \int_s^t e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\mu}_i(u)du, \boldsymbol{V} - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t-s)}\right) \quad \text{(D.1)}$$

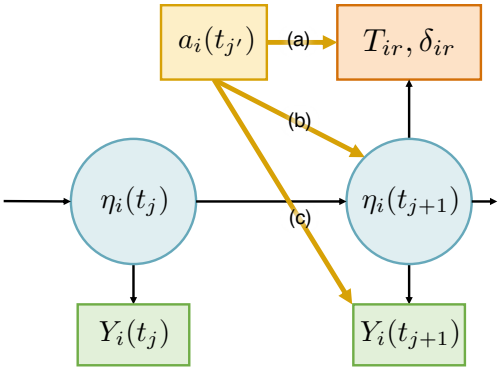


Figure D.10: Diagram illustrating potential mechanisms (a)–(c) for the effect of a single treatment at time $t_{j'}$, $a_i(t_{j'})$, on future values of the latent process $\eta_i(t_{j+1})$, the measured longitudinal outcomes $Y_i(t_{j+1})$, and the recurrent event outcome (T_{ir}, δ_{ir}). In the diagram, circles indicate that variables are latent and squares denote variables that are observed. We omit the random intercept and measurement error terms from this diagram for simplicity; these terms are still included in the model definition.

We let $\mathcal{A}_i(s - \delta_a, t)$ denote the set of times at which treatments were sent to individual i between time $s - \delta_a$ and time t; this set of treatment times corresponds to all treatments that are active at between times s and t. The analytic solution to this integral in Equation D.1 is:

$$\begin{aligned}
\int_s^t e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\mu}_i(u)du &= \int_s^t e^{-\boldsymbol{\theta}(t-u)} \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \boldsymbol{\tau}\left(1 - \frac{u - t_{ia}}{\delta_a}\right)_+ du \\
&= \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \int_s^t e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\tau}\left(1 - \frac{u - t_{ia}}{\delta_a}\right)_+ du \\
&= \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \int_{max(t_{ia},s)}^{min(t,t_{ia}+\delta_a)} e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\tau}\left(1 - \frac{u - t_{ia}}{\delta_a}\right) du \\
&= \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \left[\left(1 - \frac{u - t_{ia}}{\delta_a}\right) e^{-\boldsymbol{\theta}(t-u)}\boldsymbol{\theta}^{-1} + \frac{1}{\delta_a}e^{-\boldsymbol{\theta}(t-u)}\right]\boldsymbol{\tau}\Bigg|_{max(t_{ia},s)}^{min(t,t_{ia}+\delta_a)}
\end{aligned}$$

Plugging this result into Equation D.1, we can re-write the distribution in an analytic form:

$$\boldsymbol{\eta}_i(t)|\boldsymbol{\eta}_i(s) \sim N_p\Bigg(e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{\eta}_i(s) + \sum_{t_{ia}\in\mathcal{A}_i(s-\delta_a,t)} \left[\left(1-\frac{u-t_{ia}}{\delta_a}\right)e^{-\boldsymbol{\theta}(t-u)}\theta^{-1} + \frac{1}{\delta_a}e^{-\boldsymbol{\theta}(t-u)}\right]\boldsymbol{\tau}\Bigg|_{u=max(t_{ia},s)}^{u=min(t,t_{ia}+\delta_a)}, \boldsymbol{V} - e^{-\boldsymbol{\theta}(t-s)}\boldsymbol{V}e^{-\boldsymbol{\theta}^\top(t-s)}\Bigg) \tag{D.2}$$

D.4 Prior Distributions

When fitting our joint model, we base our prior distributions on those used in Tran et al. (2021) [112]. The priors we use are:

$$\begin{aligned}
&\lambda_k \sim \text{half-}N(1,\sigma_\lambda^2); k=1,...,K\\
&\sigma_\lambda \sim \text{half-Cauchy}(0,5)\\
&\theta_{OU_{11}},\theta_{OU_{21}},\theta_{OU_{12}},\theta_{OU_{22}} \sim N(0,10^2)\\
&\rho \sim \text{Uniform}(-0.999999, 0.999999)\\
&\sigma_{u_k} \sim \text{half-}Cauchy(0,5); k=1,...,K\\
&\sigma_{u_\epsilon} \sim \text{half-}Cauchy(0,5); k=1,...,K\\
&\beta_0,\beta_1,\beta_2 \sim N(0,5^2)\\
&\tau_1,\tau_2 \sim N(0,5^2)\\
&\tilde{\tau} \sim N(0,5^2)
\end{aligned}$$

D.5 Simulation Study

True model parameters, which are informed by the longitudinal models fit to data from similar mHealth smoking cessation studies, are given below. Setting 1's longitudinal submodel parameter values are roughly similar to those estimated in the case study in Abbott et al. (2023) [1] and setting 2's longitudinal submodel parameter values are roughly similar to those estimated in the case study in Abbott et al. (2024) [2].

Setting 1: measurement submodel

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.9 & 0 \\ 0.5 & 0 \\ 0 & 1 \\ 0 & 0.8 \end{bmatrix}, \boldsymbol{\Sigma}_u = \begin{bmatrix} 0.16 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0.64 & 0 \\ 0 & 0 & 0 & 1.00 \end{bmatrix}, \boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 0.04 & 0 & 0 & 0 \\ 0 & 0.36 & 0 & 0 \\ 0 & 0 & 0.09 & 0 \\ 0 & 0 & 0 & 0.49 \end{bmatrix}$$

Setting 1: structural submodel

$$\boldsymbol{\tau} = [2, -1]^\top, \boldsymbol{\theta} = \begin{bmatrix} 2.4 & 1.2 \\ 2.9 & 3.6 \end{bmatrix}, \boldsymbol{\sigma} = \begin{bmatrix} 1.78 & 0 \\ 0 & 1.80 \end{bmatrix} \Longrightarrow \rho = -0.68$$

Settings 1: event-time submodel

1. $\beta_0 = -1.8, \beta_1 = -0.5, \beta_2 = 0.5, \tilde{\tau} = -0.8$
2. $\beta_0 = -1.5, \beta_1 = -0.5, \beta_2 = 0.5, \beta_3 = 0.4, \tilde{\tau} = -0.8$

Setting 2: measurement submodel

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.4 & 0 \\ 0.25 & 0 \\ 0 & 0.5 \\ 0 & 0.6 \end{bmatrix}, \boldsymbol{\Sigma}_u = \begin{bmatrix} 0.16 & 0 & 0 & 0 \\ 0 & 0.16 & 0 & 0 \\ 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0.16 \end{bmatrix}, \boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 0.04 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.09 & 0 \\ 0 & 0 & 0 & 0.04 \end{bmatrix}$$

Setting 2: structural submodel

$$\boldsymbol{\tau} = [2, -1]^\top, \boldsymbol{\theta} = \begin{bmatrix} 10.2 & 5.1 \\ 4.9 & 10 \end{bmatrix}, \boldsymbol{\sigma} = \begin{bmatrix} 3.92 & 0 \\ 0 & 3.89 \end{bmatrix} \Longrightarrow \rho = -0.50$$

Settings 2: event-time submodel

1. $\beta_0 = -1.8, \beta_1 = -0.5, \beta_2 = 0.5, \tilde{\tau} = -0.8$
2. $\beta_0 = -1.5, \beta_1 = -0.5, \beta_2 = 0.5, \beta_3 = 0.4, \tilde{\tau} = -0.8$

The true hazard models used to generate the recurrent event outcomes are:

1. $h_{ir}(t) = h_0 \exp\left\{\beta_1 \eta_{1i}(t) + \beta_2 \eta_{2i}(t) + \tilde{\mu}_i(t)\right\}$

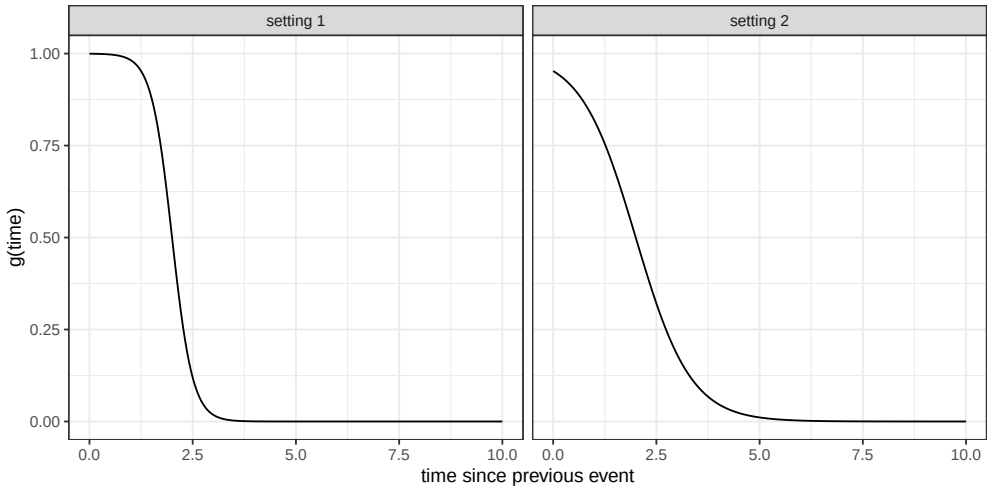


Figure D.11: Curves for the $g(\cdot)$ functions that capture the association between the hazard of an event as a function of time (in days) since the most recent prior event.

2. $h_{ir}(t) = h_0 \exp\left\{\beta_1 \eta_{1i}(t) + \beta_2 \eta_{2i}(t) + \beta_3 g(t - t_{i,r-1}) + \tilde{\mu}_i(t)\right\}$

We generate event outcomes from both of these hazard models using the true parameter values for setting 1 and setting 2. For setting 1, $g(x) = \frac{1}{1+\exp\{4(x-2)\}}$ and for setting 2, $g(x) = \frac{1}{1+\exp\{1.5(x-2)\}}$. In both of these hazard models, we assume that the baseline hazard is constant, $h_0 = \exp(\beta_0)$. In Figure D.11, we plot the curves for the $g(\cdot)$ functions that capture the association between the hazard of an event as a function of time since the most recent prior event.

We also plot the trajectories of the latent bivariate OU process for a subset of individuals in a single simulated dataset, illustrating both the case when the treatment effect is modeled as an additive term and as drift, in Figures D.12-D.15. We just provide these plots when using hazard model 1 to generate the recurrent event outcomes.

D.5.1 Additional Simulations for β_3

In the simulation results shown in the main paper, we see some bias in the posterior medians for parameter β_3 in the recurrent event submodel when assuming hazard model 2. This bias results in a coverage rate that is lower than nominal. The parameter β_3 is the coefficient that captures the association between the hazard of the r^{th} recurrent event and the time since the $(r-1)^{th}$ event. We investigate two factors that might be contributing to this bias: (a) the finite sample size of our simulated datasets and (b) the choice of grid width used when approximating the cumulative hazard function via a midpoint rule. To investigate (a), we generate data with a sample size of $N = 300$ individuals, an increase from the

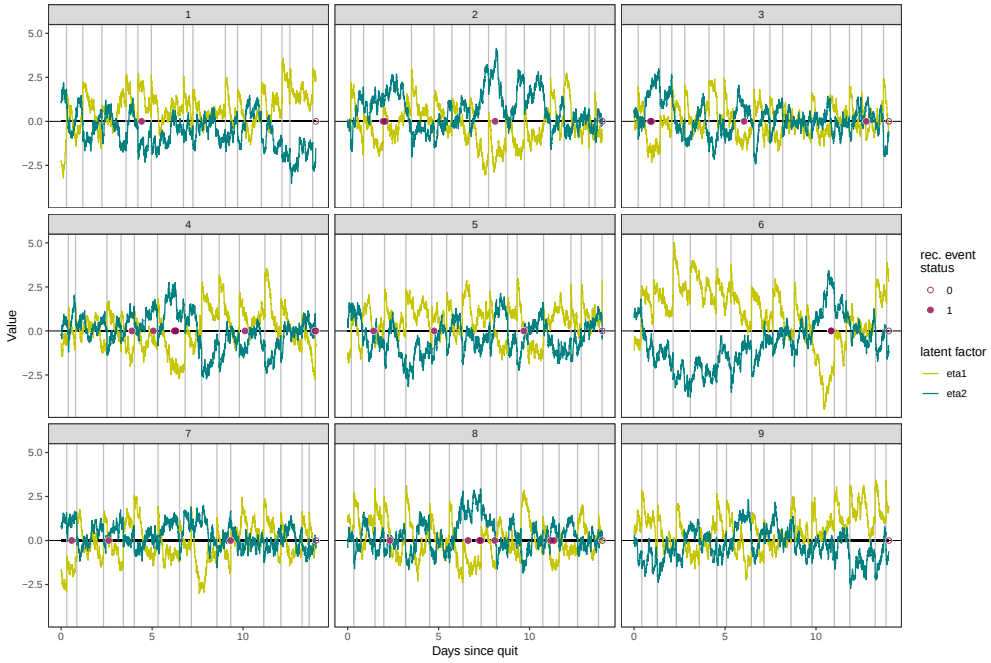


Figure D.12: **Setting 1**: Treatment effect is **additive** and the hazard takes the form of **model 1**. Green lines show the trajectory of the bivariate latent process and vertical grey bars indicate the timing of the treatments, which are sent randomly once per day and are assumed to have an effect that lasts half a day ($\delta_a = 0.5$). Events are shown as solid pink dots and censoring times are indicated with open pink dots.

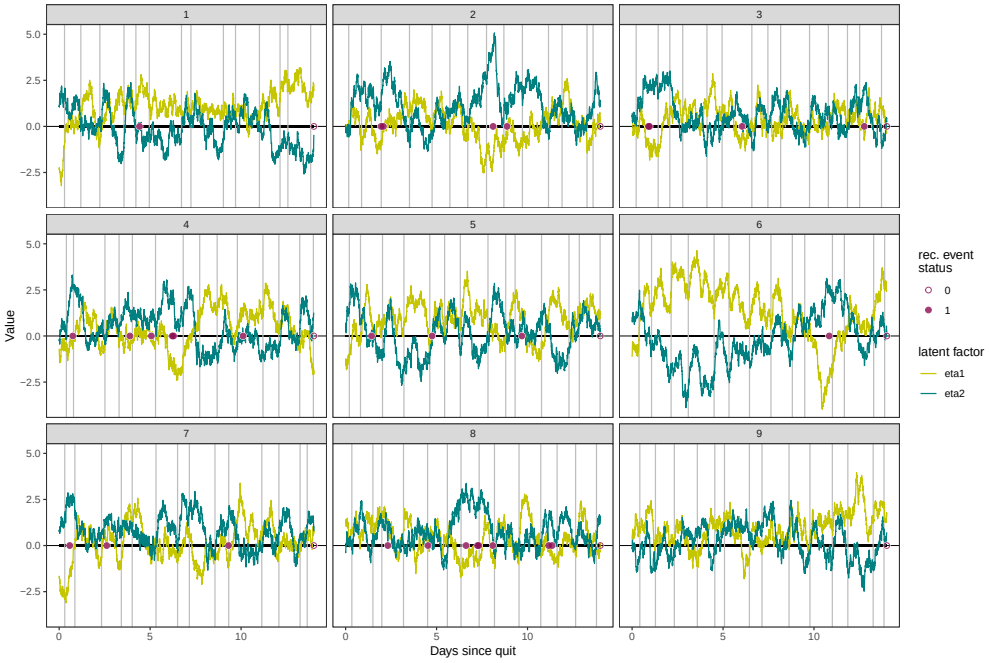


Figure D.13: **Setting 1**: Treatment effect is modeled as **drift** and the hazard takes the form of **model 1**. Green lines show the trajectory of the bivariate latent process and vertical grey bars indicate the timing of the treatments, which are sent randomly once per day and are assumed to have an effect that lasts half a day ($\delta_a = 0.5$). Events are shown as solid pink dots and censoring times are indicated with open pink dots.

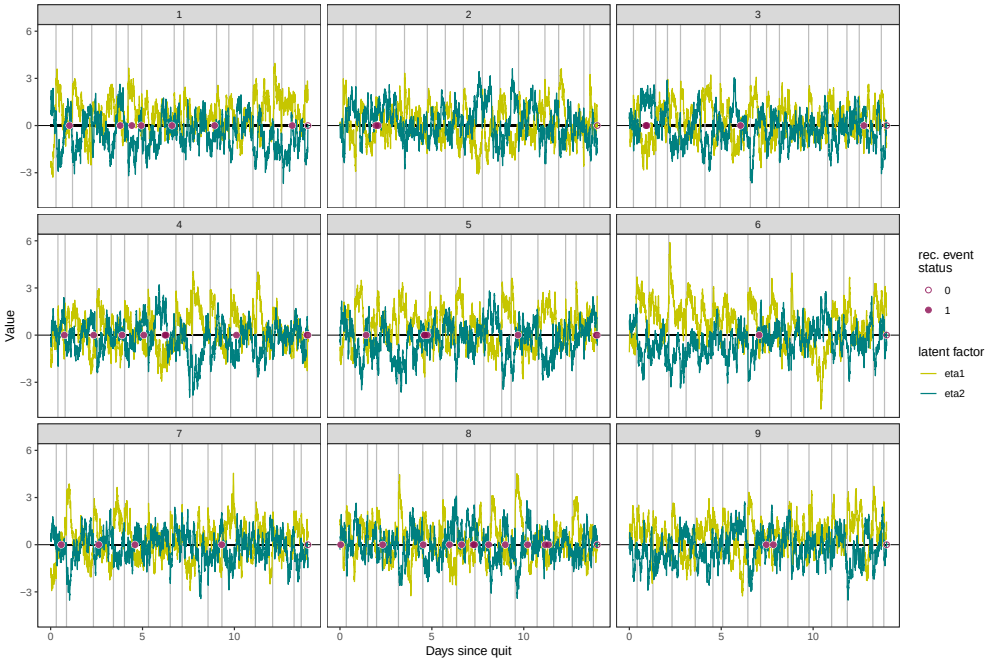


Figure D.14: **Setting 2**: Treatment effect is **additive** and the hazard takes the form of **model 1**. Green lines show the trajectory of the bivariate latent process and vertical grey bars indicate the timing of the treatments, which are sent randomly once per day and are assumed to have an effect that lasts half a day ($\delta_a = 0.5$). Events are shown as solid pink dots and censoring times are indicated with open pink dots.

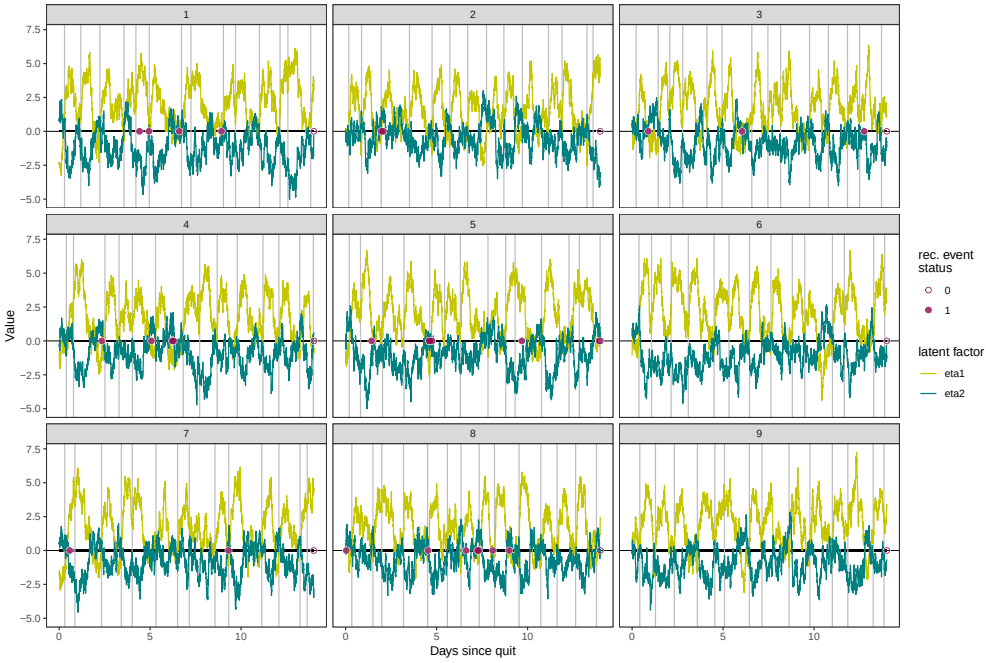


Figure D.15: **Setting 2**: Treatment effect is modeled as **drift** and the hazard takes the form of **model 1**. Green lines show the trajectory of the bivariate latent process and vertical grey bars indicate the timing of the treatments, which are sent randomly once per day and are assumed to have an effect that lasts half a day ($\delta_a = 0.5$). Events are shown as solid pink dots and censoring times are indicated with open pink dots.

$N = 100$ sample size used in the original simulation study. To investigate (b), we increase the density of the grid used in the midpoint approximation of the cumulative hazard function from one point approximately each 12 hours to one point about each 4 hours. We compare the posterior medians and coverage rates for these additional simulations to a subset of the original simulations in Figures D.16 and D.17. We consider 5 replicates.

The results of this small supplemental simulation study suggest that the bias in the posterior median for β_3 is likely related to fitting a complex model to a fairly small dataset. The bias in the posterior median results in low coverage. In setting 2, we see a decrease in the bias for this parameter estimate when we increase the sample size from 100 independent individuals to 300 independent individuals. In setting 1, we see less of a decrease in bias when increasing the sample size. This differential decrease in bias could be partially due to the differences in the amount of temporal correlation across these two settings: in setting 1, the longitudinal latent process has slower decay in correlation over time than in setting 2, and so accurately estimating β_3, the coefficient on another function of time, could be more challenging in setting 1 than in setting 2. If we were to substantially increase the sample size, perhaps to 3000, we would expect to see less bias in both settings.

We also hypothesized that the bias in β_3 could be related to the coarseness of the approximation used to evaluate the cumulative hazard function. As mentioned in the main text, we approximate the cumulative hazard function by replacing the integral with a sum over a fine grid of points; we then evaluate this sum using a midpoint rule. If the grid used in the approximation is too coarse, then it will be poor. However, we found that decreasing the coarseness of the grid by a factor of 3 did not substantially impact the bias in our estimates.

D.6 Model Comparisons

To compare the fit of different models, we consider two different measures of predictive accuracy: DIC and WAIC. DIC and WAIC are both based on estimates of the log pointwise predictive density (lppd). For a general model with data $y_1, ..., y_N$ and parameters Θ, Vehtari et al. (2017) [118] define the lppd as

$$\text{lppd} = \sum_{i=1}^{N} log\left(p(y_i|y)\right) = \sum_{i=1}^{N} log\left(\int p(y_i|\Theta)p(\Theta|y)d\Theta\right)$$

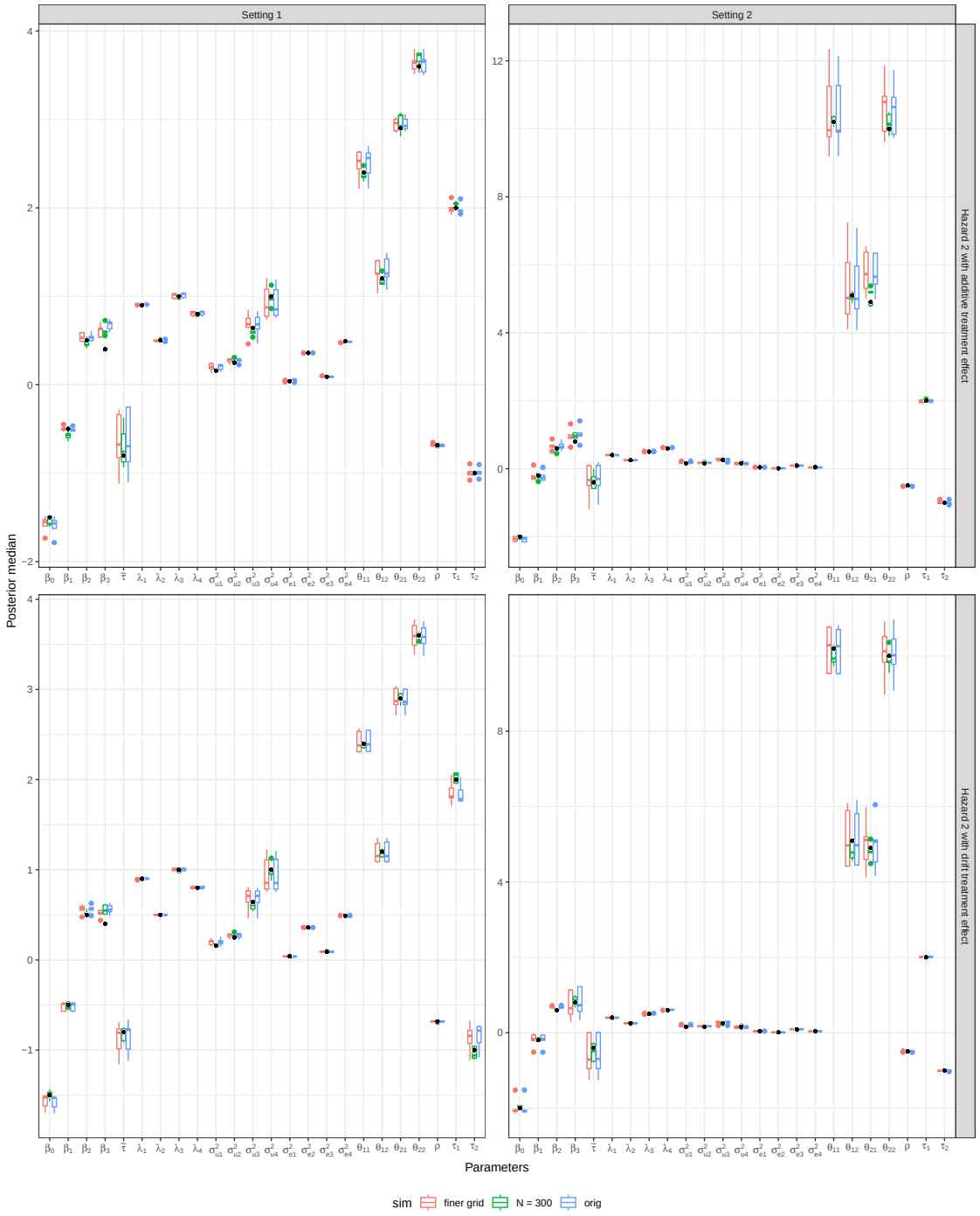


Figure D.16: For data generated under settings 1 and 2 with hazard model 2 and different treatment effect models in the longitudinal submodel, we use box plots to summarize the distribution of the **posterior medians for all parameters** across the 5 simulated datasets. The original simulation design has a sample size of $N = 100$ and assumes a grid width of 12-hour intervals for approximating the cumulative hazard function via a midpoint rule. We modify this original simulation design by either increasing the sample size to $N = 300$ or decreasing the grid width to 4-hour intervals. True parameter values are indicated with colored dots.

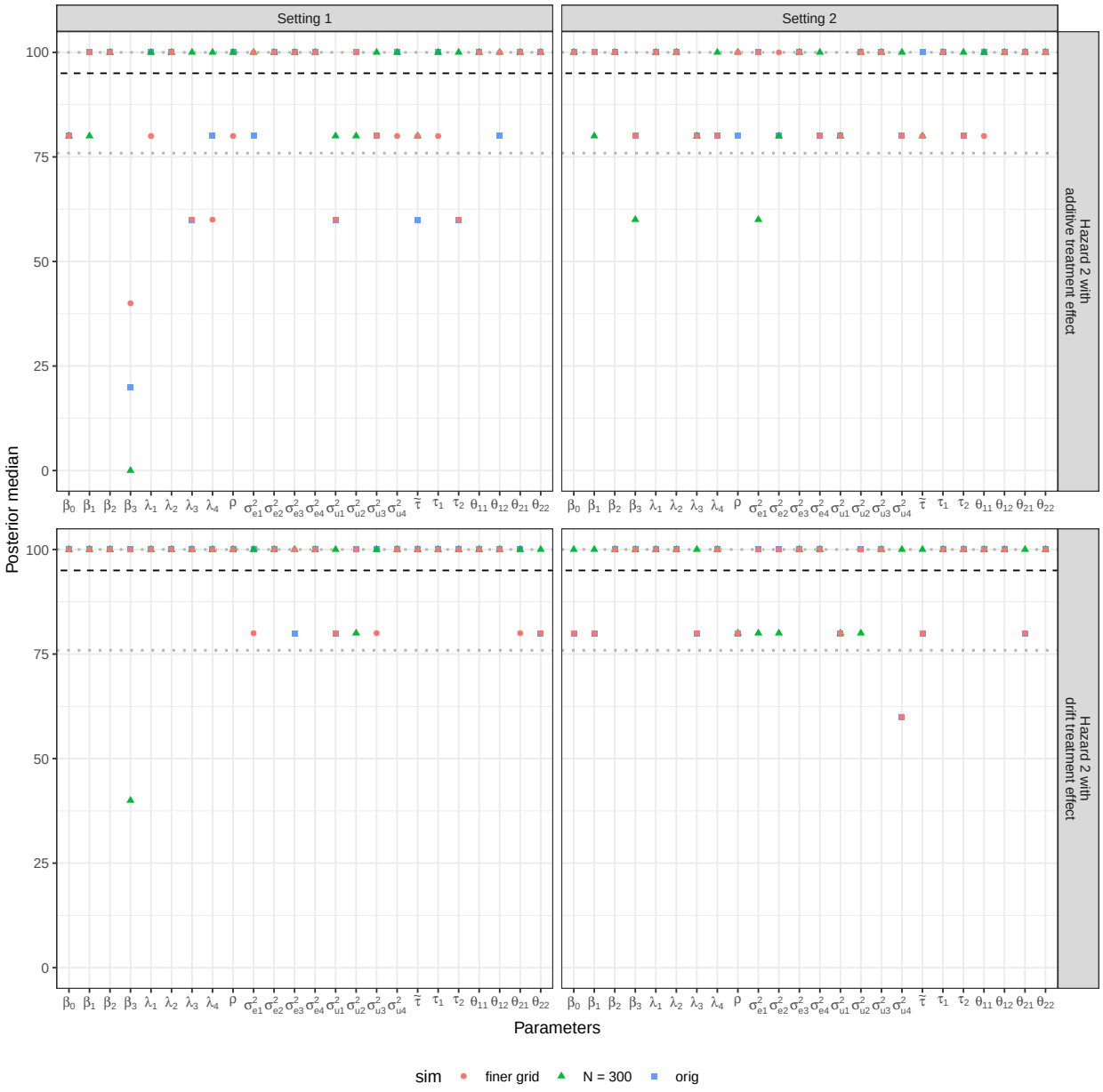


Figure D.17: For data generated under settings 1 and 2 with hazard model 2 and different treatment effect models in the longitudinal submodel, we summarize the **coverage rate of 95% credible intervals** across the 5 simulated datasets with the colored dots. The original simulation design has a sample size of $N = 100$ and assumes a grid width 12-hour intervals for approximating the cumulative hazard function via a midpoint rule. We modify this original simulation design by either increasing the sample size to $N = 300$ or decreasing the grid width to 4-hour intervals. The black horizontal dashed lines indicate target coverage and the dotted grey lines corresponds to the upper and lower bounds of a 95% binomial proportion confidence interval for a probability of 0.95.

The lppd can be computed as

$$\widehat{\text{lppd}} = \sum_{i=1}^{N} log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\Theta^s)\right)$$

where $\Theta^s, s = 1, ..., S$ are draws from the "usual" posterior distribution, $p(\Theta|y)$ is the posterior distribution, and $p(y_i|y)$ is the posterior predictive distribution. In our joint model, the data y_i would consist of the longitudinal outcomes and the recurrent event outcomes.

Continuing with generic notation using y as the data and Θ as the model parameters, Gelman (2014) [30] define DIC and WAIC as:

- DIC:

  $$DIC = -2logp(y|\hat{\Theta}) + 2p_{DIC}$$

  where $\hat{\Theta}$ is the posterior mean and p_{DIC} is the effective number of parameters, with

  $$p_{\text{DIC}} = 2\left(logp(y|\hat{\Theta}) - \frac{1}{S}\sum_{s=1}^{S} logp(y|\Theta^s)\right)$$

  where s indexes posterior samples.

- WAIC:

  $$WAIC = -2\text{lppd} + 2p_{\text{WAIC2}}$$

  where the computed log pointwise predictive density is

  $$\widehat{\text{lppd}} = \sum_{i=1}^{N} log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\Theta^s)\right)$$

  and the effective number of parameters is

  $$p_{\text{WAIC2}} = \sum_{i=1}^{N} V_{s=1}^{S}\left(logp(y_i|\Theta^s)\right)$$

  and $V_{s=1}^{S}(a_s) = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$.

[30] also discusses alternative ways to calculate the effective sample size, but the definitions above use the suggested/standard approaches.

Because our setting involves latent variables, we must carefully consider the form of the likelihood that we would like to use in our calculations of the log pointwise predictive density. That is, is it more appropriate to use a version of the likelihood that conditions

on the latent factors and all other model parameters, or should we use a version of the likelihood in which we marginalize over the latent factors? Working in the context of latent factor models commonly used in psychometric research, Merkle et al. (2019) [60] discuss important differences between these conditional and marginal versions of the likelihood, and how these different definitions target different types of predictive accuracy. They recommend working with the marginal version of the likelihood, which aligns with the recommendations for joint models given in Rizopoulos (2023) [89]. Before we define the computed log pointwise predictive density in our setting, we define some general notation:

- y = data (in our setting, this includes both the longitudinal and recurrent event outcomes)
- Θ = all parameters in our model, excluding the latent process η and its parameters θ_{OU} and σ_{OU}
- ψ = OU process parameters θ_{OU} and σ_{OU}
- f_c = joint model likelihood conditional on the latent factors
- f_m = joint model likelihood marginalized over the latent factors

Then, the lppd can be computed as:

$$\widehat{\text{lppd}} = \sum_{i=1}^{N} log\mathbb{E}_{\Theta,\psi|y}\left\{f_m(y_i|\Theta,\psi)\right\} \tag{D.3}$$

$$= \sum_{i=1}^{N} log\left[\int f_m(y_i|\Theta,\psi)p(\Theta,\psi|y)d\Theta d\psi\right] \tag{D.4}$$

$$= \sum_{i=1}^{N} log\left[\int \mathbb{E}_{\eta|\psi}\left\{f_c(y_i|\eta_i,\Theta)\right\}p(\Theta,\psi|y)d\Theta d\psi\right] \tag{D.5}$$

Focus now on $\mathbb{E}_{\eta|\psi}\left\{f_c(y_i|\eta_i,\Theta)\right\}$, which requires integrating over the latent process:

$$\mathbb{E}_{\eta|\psi}\left\{f_c(y_i|\eta_i,\Theta)\right\} = \int f_c(y_i|\eta_i,\Theta)p(\eta_i|\psi)d\eta_i$$

To directly approximate this integral with a sum over sampled values of η would require having draws of η from the distribution $p(\eta_i|\psi)$. The posterior samples of η that are generated during model fitting are conditional on the data, $p(\eta_i|y)$. We can use the posterior samples and importance sampling to approximate the integral.

$$\mathbb{E}_{\eta|\psi}\left\{f_c(y_i|\eta_i,\Theta)\right\} = \int f_c(y_i|\eta_i,\Theta)p(\eta_i|\psi)d\eta_i \tag{D.6}$$

$$= \int f_c(y_i|\eta_i,\Theta)\frac{p(\eta_i|\psi)}{p(\eta_i|y)}p(\eta_i|y)d\eta_i \tag{D.7}$$

$$\approx \frac{1}{M}\sum_{m=1}^{M} f_c(y_i|\eta_i^m,\Theta)\frac{p(\eta_i^m|\psi)}{p(\eta_i^m|y)} \tag{D.8}$$

where η_i^m is sampled from the unconditional marginal distribution of $\eta_i|y$, as given in Appendix C of Merkle et al. (2019) [60]. This unconditional marginal distribution of $\eta_i|y$ is a normal distribution with a mean and variance that are based on the marginal mean and variance of the usual posterior samples of η_i. If we follow this approach to generate samples of η_i^m, then we can plug Equation D.6 in to Equation D.3:

$$\widehat{\text{lppd}} = \sum_{i=1}^{N} log\left[\int \mathbb{E}\left\{f_c(y_i|\eta_i,\Theta)\right\}p(\Theta,\psi|y)d\Theta d\psi\right] \tag{D.9}$$

$$\approx \sum_{i=1}^{N} log\left[\int\left[\frac{1}{M}\sum_{m=1}^{M} f_c(y_i|\eta_i^m,\Theta)\frac{p(\eta_i^m|\psi)}{p(\eta_i^m|y)}\right]p(\Theta,\psi|y)d\Theta d\psi\right] \tag{D.10}$$

$$\approx \sum_{i=1}^{N} log\left[\frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{M}\sum_{m=1}^{M} f_c(y_i|\eta_i^m,\Theta^s)\frac{p(\eta_i^m|\psi^s)}{p(\eta_i^m|y)}\right]\right] \tag{D.11}$$

where Θ^s, ψ^s are the usual posterior samples. Then, to compute lppd for one individual, we would:

1. sample M values of η_i^m from the unconditional marginal distribution of $\eta_i|y$, which does not depend on Θ or ψ if we use a version of the approach in Merkle et al. (2019) [60]. Note that because we are calculating the empirical covariance matrix for the entire vector η_i, the covariance matrix is large and can be unstable. To avoid non-positive definite covariance matrices, we can add a very small amount to the diagonal of the empirical covariance matrix.

2. for each posterior sample $s = 1, ..., S$ of Θ^s and ψ^s, compute the density across the M sampled values of η_i^m

We repeat steps 1 and 2 for each individual $i = 1, ..., N$ and sum up the values of the marginal log-likelihood to get $\widehat{\text{lppd}}$. We can make this approach more explicit in our definition of DIC

and WAIC:

$$
\begin{aligned}
DIC =& -2logp(y|\hat{\Theta}) + 2p_{DIC} \\
=& -2logp(y|\hat{\Theta}) + 4\left[logp(y|\hat{\Theta}) - \frac{1}{S}\sum_{s=1}^{S} logp(y|\Theta^s)\right] \\
=& -2\sum_{i=1}^{N} log\left[\frac{1}{M}\sum_{m=1}^{M} p(y_i|\hat{\Theta},\eta_i^m)\frac{p(\eta_i^m|\hat{\psi})}{p(\eta_i^m|y)}\right] \\
&+4\left[\sum_{i=1}^{N} log\Big[\frac{1}{M}\sum_{m=1}^{M} p(y_i|\hat{\Theta},\eta_i^m)\frac{p(\eta_i^m|\hat{\psi})}{p(\eta_i^m|y)}\Big]\right. \\
&\left.-\frac{1}{S}\sum_{s=1}^{S}\Big[\sum_{i=1}^{N} log\Big[\frac{1}{M}\sum_{m=1}^{M} p(y_i|\Theta^s,\eta_i^m)\frac{p(\eta_i^m|\psi^s)}{p(\eta_i^m|y)}\Big]\Big]\right]
\end{aligned}
$$

$$
\begin{aligned}
WAIC =& -2\widehat{\text{lppd}} + 2p_{WAIC2} \\
=& -2\sum_{i=1}^{N} log\left[\frac{1}{S}\sum_{i=1}^{S}\Big[\frac{1}{M}\sum_{m=1}^{M} p(y_i|\Theta^s,\eta_i^m)\frac{p(\eta_i^m|\psi^s)}{p(\eta_i^m|y)}\Big]\right] \\
&+2\sum_{i=1}^{N} V_{s=1}^{S}\left(log\Big[\frac{1}{M}\sum_{m=1}^{M} p(y_i|\Theta^s,\eta_i^m)\frac{p(\eta_i^m|\psi^s)}{p(\eta_i^m|y)}\Big]\right)
\end{aligned}
$$

In the definitions above, s indexes posterior samples, $\hat{\Theta}$ and $\hat{\psi}$ are posterior means, m indexes samples of the latent process drawn from the approximate unconditional marginal distribution of $\eta_i|y$, and $V_{s=1}^{S}(\cdot)$ is the sample variance. y contains both our longitudinal and recurrent event data (written as (Y, T, δ) in the main paper). In the main paper, we subsample every 5th iteration of the final 1,000 posterior samples across all chains, resulting in $S = 800$. We use $M = 25$ when calculating DIC and WAIC.

To confirm that our approximate approach to calculate the marginal log-likelihood works reasonably well, we try calculating the marginal log-likelihood using the approximate approach described above and using the exact marginal distribution for just the longitudinal submodel with the additive treatment effect model. For this longitudinal submodel, we can derive the marginal distribution algebraically. We calculate the approximate and exact log-likelihoods at the posterior means for two simulated datasets and compare the results in Figure D.18. If our approximation works well, then we expect the value of the marginal log-likelihood for each individual i to be roughly the same for both the approximate marginal log-likelihood vs. the algebraic/exact marginal log-likelihod evaluated at the posterior mean.

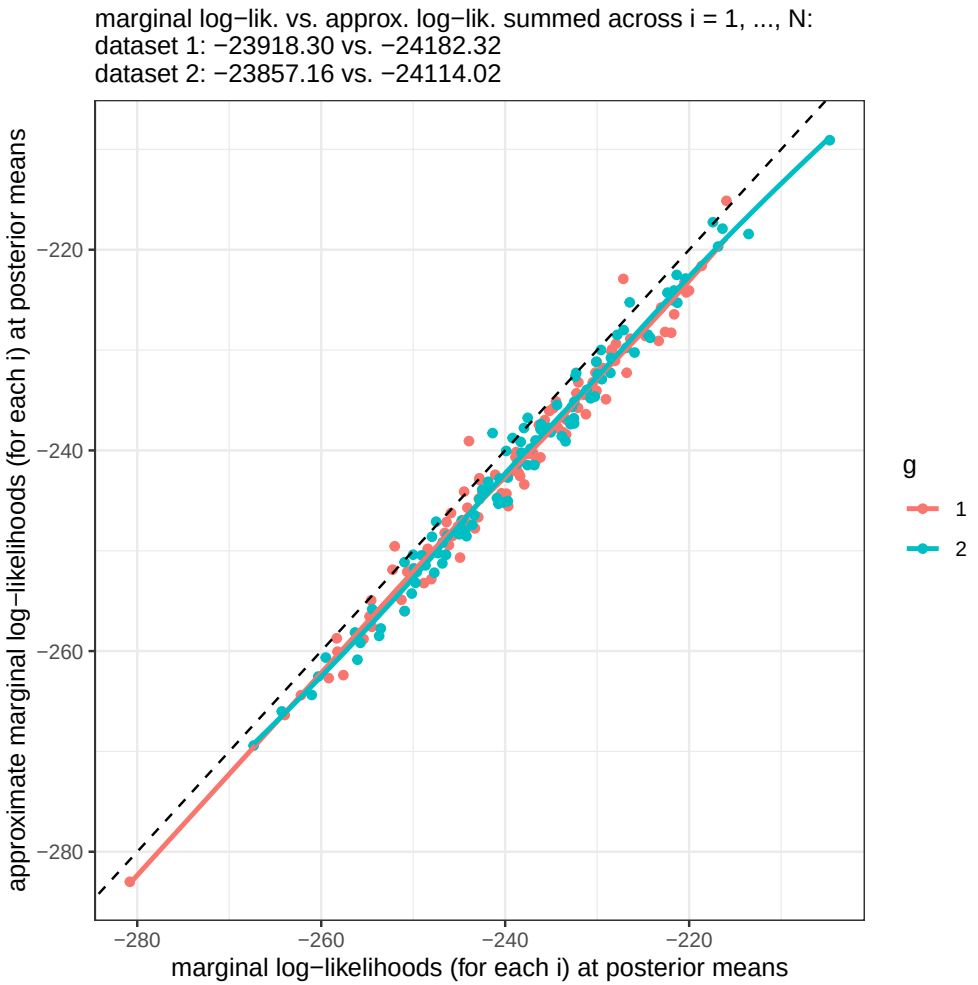


Figure D.18: Comparison of approximate log-likelihood values for each individual evaluated at the posterior mean to the exact log-likelihood for each individual using the true marginal distribution evaluated at the posterior mean. Results from two simulated datasets and two fitted models are shown using red and blue.

Figure D.18 shows this to be true, as indicated by the colored lines and points falling along the 0-1 axis. We do see some bias, but because this bias is consistent, it should not be problematic for the purposes of model selection. We can also repeat the same comparison but for the approximate and exact marginal log-likelihoods evaluated each posterior sample $\boldsymbol{\Theta}^s$, rather than the posterior mean $\hat{\boldsymbol{\Theta}}$ (see Figure D.19).

D.6.1 Case Study Results

In Table D.3, we provide DIC for each of the joint models fit to the case study data. Of the 8000 posterior samples across 4 chains, we subsampled every 5th iteration of the final 1,000

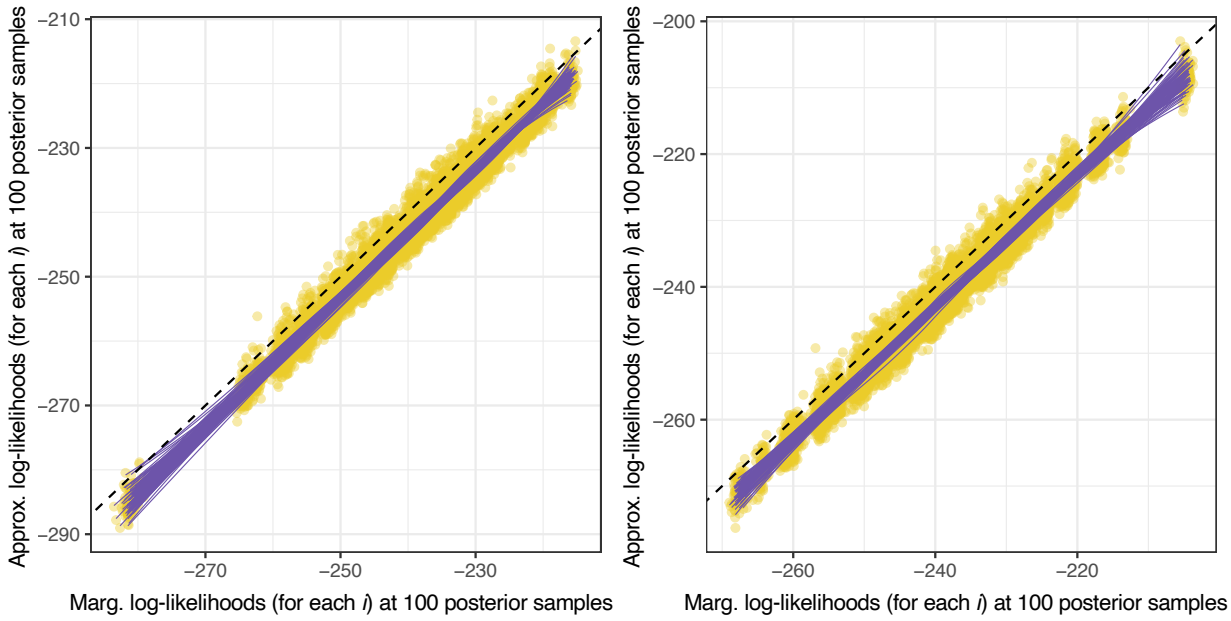


Figure D.19: Comparison of approximate log-likelihood values for each individual evaluated at each posterior sample to the exact log-likelihood for each individual using the true marginal distribution evaluated at each posterior sample s. Note that we subsample 100 of the total 1000 posterior samples. Results from one simulated dataset are on the left, results from the other dataset are on the right. Linear best-fit lines are plotted across individuals for each set of posterior samples s.

DIC	Impact of treatment on latent process	
Hazard model	Additive	Drift
Single treatment parameter	59,894.96	60,037.92
Separate pre- and post-quit treatment parameters	59,819.76	**59,746.39**

Table D.3: DIC for the joint models fit to the motivating MRT data. When approximating the marginal log-likelihood, we subsampled every 5th iteration of the final 1,000 posterior samples to use in our calculations. A lower value of DIC is preferred. The lowest value of DIC is indicated in bold text.

posterior samples (resulting in $S = 800$) to use in our WAIC and DIC calculations. We used $M = 25$ for generating marginal posterior samples of the latent process η.

BIBLIOGRAPHY

[1] M R Abbott, W H Dempsey, I Nahum-Shani, C Y Lam, D W Wetter, and J M G Taylor. A continuous-time dynamic factor model for intensive longitudinal data arising from mobile health studies. *arXiv preprint arXiv:2307.15681*, 2023.

[2] M R Abbott, W H Dempsey, I Nahum-Shani, L N Potter, D W Wetter, C Y Lam, and J M G Taylor. A Bayesian joint longitudinal-survival model with a latent stochastic process for intensive longitudinal data. *arXiv preprint arXiv:2405.00179*, 2024.

[3] T H Abbott. *Interactions between Atmospheric Deep Convection and the Surrounding Environment*. PhD thesis, Massachusetts Institute of Technology, 2021.

[4] O Aguilar and M West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357, 2000.

[5] P S Albert and J H Shih. An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, 4(3):1517–1532, 2010.

[6] T Asparouhov and B Muthén. Continuous-time survival analysis in Mplus, 2018. Accessed June 29, 2023.

[7] J W Bartlett and R A Hughes. Bootstrap inference for multiple imputation under uncongeniality and misspecification. *Statistical Methods in Medical Research*, 29(12):3533–3546, 2020.

[8] K C Berg, R D Crosby, L Cao, C B Peterson, S G Engel, J E Mitchell, and S A Wonderlich. Facets of negative affect prior to and following binge-only, purge-only, and binge/purge events in women with bulimia nervosa. *Journal of Abnormal Psychology*, 122(1):111–118, 2013.

[9] A Boruvka, D Almirall, K Witkiewitz, and S A Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018.

[10] D Brigo and F Mercurio. *Interest Rate Models-Theory and Practice with Smile, Inflation and Credit*, chapter 3. Springer Finance, second edition, 2007.

[11] M Britton, S Haddad, and J L Derrick. Perceived partner responsiveness predicts smoking cessation in single-smoker couples. *Addictive Behaviors*, 88:122–128, 2019.

[12] A F Brouwer, J Jeon, J L Hirschtick, E Jimenez-Mendoza, R Mistry, I V Bondarenko, et al. Transitions between cigarette, ends and dual use in adults in the PATH study (waves 1–4): multistate transition modelling accounting for complex survey design. *Tobacco Control*, 31:424–431, 2022.

[13] E R Brown, J G Ibrahim, and V De Gruttola. A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73, 2005.

[14] M S Businelle, D E Kendzor, L R Reitzel, T J Costello, L Cofta-Woerpel, Y Li, , et al. Mechanisms linking socioeconomic status to smoking cessation: a structural equation modeling approach. *Health Psychology*, 29(3):262–273, 2010.

[15] P-C Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

[16] P-C Bürkner. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411, 2018.

[17] P-C Bürkner. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54, 2021.

[18] B Carpenter, A Gelman, M D Hoffman, D Lee, B Goodrich, M Betancourt, M Brubaker, J Guo, P Li, and A Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32, 2017.

[19] C M Carvalho, J Chang, J E Lucas, J R Nevins, Q Wang, and M West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

[20] S S Chatterjee, M Chakrabarty, D Banerjee, S Grover, S S Chatterjee, and U Dan. Stress, sleep and psychological impact in healthcare workers during the early phase of Covid-19 in India: A factor analysis. *Frontiers in Psychology*, 12, 2021.

[21] K Cui and D B Dunson. Generalized dynamic factor models for mixed-measurement time series. *Journal of Computational and Graphical Statistics*, 23(1):169–191, 2014.

[22] V De Gruttola and X M Tu. Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.

[23] D B Dunson. Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*, 98(463):555–563, 2003.

[24] A F Elmi, K L Grantz, and P S Albert. An approximate joint model for multiple paired longitudinal outcomes and time-to-event data. *Biometrics*, 74(3):1112–1119, 2018.

[25] D H Epstein, M Tyburski, I M Craig, K A Phillips, M L Jobes, M Vahabzadeh, M Mezghanni, J L Lin, C D M Furr-Holden, and K L Preston. Real-time tracking of neighborhood surroundings and mood in urban drug misusers: application of a new

method to study behavior in its geographical context. *Drug and Alcohol Dependence*, 134:22–29, 2014.

[26] D H Epstein, M Tyburski, W J Kowalczyk, A J Burgess-Hull, K A Phillips, B L Curtis, and K L Preston. Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. *npj Digital Medicine*, 3(26), 2020.

[27] T Fernández, N Rivera, and Y W Teh. Gaussian processes for survival analysis. *arXiv preprint arXiv:1611.00817*, 2016.

[28] M C Fiore, C R Jaén, T B Baker, W C Bailey, N L Benowitz, S J Curry, et al. Treating tobacco use and dependence: 2008 update, 2008. In: Panel TUaDG, editor. Rockville, MD.

[29] A Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[30] A Gelman, J B Carlin, and H S Stern. *Bayesian Data Analysis*, chapter 7. Chapman & Hall/CRC texts in statistical science, third edition, 2014.

[31] A Gelman, J Hwang, and A Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24:997–1016, 2014.

[32] P Gilbert, K McEwan, R Mitra, L Franks, A Richter, and H Rockliff. Feeling safe and content: A specific affect regulation system? Relationship to depression, anxiety, stress, and self-criticism. *The Journal of Positive Psychology*, 3(3):182–191, 2008.

[33] N Grant, J Wardle, and A Steptoe. The relationship between life satisfaction and health behavior: a cross-cultural analysis of young adults. *International Journal of Behavioral Medicine*, 16(3):259–68, 2009.

[34] R G Gunter, E H Szeto, S Suh, Y Kim, S H Jeong, and A J Waters. Associations between affect, craving, and smoking in Korean smokers: An ecological momentary assessment study. *Addictive Behaviors Reports*, 12, 2020.

[35] B He and S Luo. Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease. *Statistical Methods in Medical Research*, 25(4):1346–1358, 2016.

[36] R Henderson, P Diggle, and A Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

[37] G L Hickey, P Philipson, A Jorgensen, and R Kolamunnage-Dona. joineRML: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Medical Research Methodology*, 18(1), 2018.

[38] K Hovsepian, M al'Absi, E Ertin, T Kamarck, M Nakajima, and S Kumar. cstress: Towards a gold standard for continuous stress assessment in the mobile environment. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 493–504, 2015.

[39] L T Hoyt, P L Chase-Lansdale, T W McDade, and E K Adam. Positive youth, healthy adults: does positive well-being in adolescence predict better perceived health and fewer risky health behaviors in young adulthood? *Journal of Adolescent Health*, 50(1):66–73, 2012.

[40] J G Ibrahim, H Chu, and L M Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 26(16):2796–2801, 2010.

[41] C Jackson. flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, 70(8):1–33, 2016.

[42] V S Jeganathan, J R Golbus, K Gupta, E Luff, W Dempsey, T Boyden, M Rubenfire, B Mukherjee, P Klasnja, S Kheterpal, and B K Nallamothu. Virtual AppLication-supported Environment To INcrease Exercise (VALENTINE) during cardiac rehabilitation study: Rationale and design. *American Heart Journal*, 248:53–62, 2022.

[43] K Kang, D Pan, and X Song. A joint model for multivariate longitudinal and survival data to discover the conversion to Alzheimer's disease. *Statistics in Medicine*, 41(2):356–373, 2022.

[44] K Kang and X Song. Consistent estimation of a joint model for multivariate longitudinal and survival data with latent variables. *Journal of Multivariate Analysis*, 187:104827, 2022.

[45] P Klasnja, E B Hekler, S Shiffman, A Boruvka, D Almirall, A Tewari, and S A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, (0):1220–8, 2015.

[46] P Klasnja, S Smith, N J Seewald, A Lee, K Hall, B Luers, E B Hekler, and S A Murphy. Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

[47] R J Larsen. Toward a science of mood regulation. *Psychological Inquiry*, 11(3):129–141, 2000.

[48] R Lawton, M Conner, and D Parker. Beyond cognition: predicting health risk behaviors from instrumental and affective beliefs. *Health Psychology*, 26(3):259–67, 2007.

[49] K Li and S Luo. Dynamic prediction of Alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in Medicine*, 38(24):4804—-4818, 2019.

[50] N Li, Y Liu, S Li, R M Elashoff, and G Li. A flexible joint model for multiple longitudinal biomarkers and a time-to-event outcome: With applications to dynamic prediction using highly correlated biomarkers. *Biometrical Journal*, 63(8), 2021.

[51] P Liao, P Klasnja, A Tewari, and S A Murphy. Sample size calculations for micro-randomized trials in mHealth. *Statistics in Medicine*, 35(12):1944–1971, 2016.

[52] M Liu, J Sun, J D Herazo-Maya, N Kaminski, and H Zhao. Joint models for time-to-event data and longitudinal biomarkers of high dimension. *Statistics in Biosciences*, 11(3):614–629, 2019.

[53] M Liu, J Sun, J D Herazo-Maya, N Kaminski, and H Zhao. Joint models for time-to-event data and longitudinal biomarkers of high dimension. *Statistics in Biosciences*, 11(3):614–629, 2019.

[54] S Liuand, L Ou, and E Ferrer. Dynamic mixture modeling with dynr. *Multivariate Behavioral Research*, 56(6):941–955, 2021.

[55] T Lumley. survey: analysis of complex survey samples, 2020. R package version 4.0.

[56] E Lázaro, C Armero, and D Alvares. Bayesian regularization for flexible baseline hazard functions in Cox survival models. *Biometrical Journal*, 63, 2021.

[57] K Mauff, E Steyerberg, I Kardys, E Boersma, and D Rizopoulos. Joint models with multiple longitudinal outcomes and a time-to-event outcome: A corrected two-stage approach. *Statistics and Computing*, 30:999–1014, 2020.

[58] S McCurdy, A Molinaro, and L Pachter. Factor analysis for survival time prediction with informative censoring and diverse covariates. *Statistics in Medicine*, 38(20):3719–3732, 2019.

[59] M D McManus, J T, Siegel, and J Nakamura. The predictive power of low-arousal positive affect. *Motivation and Emotion*, 43:130–144, 2019.

[60] E C Merkle, D Furr, and S Rabe-Hesketh. Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84(3):802–829, 2019.

[61] A Moreno, Z Wu, J R Yap, C Lam, D Wetter, I Nahum-Shani, W Dempsey, and J M Rehg. A robust functional EM algorithm for incomplete panel count data. *Advances in Neural Information Processing Systems*, 33:19828–19838, 2020.

[62] J Murray and P Philipson. A fast approximate EM algorithm for joint models of survival and multivariate longitudinal data. *Computational Statistics & Data Analysis*, 170, 2022.

[63] J Z Musoro, R B Geskus, and A H Zwinderman. A joint model for repeated events of different types and multiple longitudinal outcomes with application to a follow-up study of patients after kidney transplant. *Biometrical Journal*, 57(2):185–200, 2015.

[64] B Muthén. Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29:81–117, 2002.

[65] L K Muthén and B Muthén. Categorical latent variable modeling using Mplus: longitudinal data in Mplus short courses topic 6, 2009. Accessed June 29, 2023.

[66] L K Muthén and B O Muthén. Example 6.22: Continuous-time survival analysis using a parametric proportional hazards model with a factor influencing survival in Mplus user's guide. 8th ed., Los Angeles, CA, 1998-2017. Accessed June 29, 2023.

[67] I Nahum-Shani, E B Hekler, and D Spruijt-Metz. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology*, 34S(0):1209–1219, 2015.

[68] I Nahum-Shani, L N Potter, C Y Lam, J Yap, A Moreno, R Stoffel, Z Wu, N Wan, W Dempsey, S Kumar, E Ertin, S A Murphy, J M Rehg, and D W Wetter. The mobile assistance for regulating smoking (MARS) micro-randomized trial design protocol. *Contemporary Clinical Trials*, 2021.

[69] I Nahum-Shani, S N Smith, B J Spring, L M Collins, K Witkiewitz, A Tewari, and S A Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

[70] I Nahum-Shani, SN Smith, BJ, LM Collins, K Witkiewitz, A Tewari, and SA Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

[71] Z Oravecz, F Tuerlinckx, and J Vandekerckhove. A hierarchical Ornstein-Uhlenbeck model for continuous repeated measurement data. *Psychometrika*, 74:395–418, 2009.

[72] Z Oravecz, F Tuerlinckx, and J Vandekerckhove. Bayesian data analysis with the bivariate hierarchical Ornstein-Uhlenbeck process model. *Multivariate Behavioral Research*, 51(1):106–119, 2016.

[73] L N Potter, C R Schlechter, I Nahum-Shani, C Y Lam, P M Cinciripini, and D W Wetter. Socio-economic status moderates within-person associations of risk factors and smoking lapse in daily life. *Addiction*, 118(5):925–934, 2023.

[74] L N Potter, J Yap, W Dempsey, D W Wetter, and I Nahum-Shani. Integrating intensive longitudinal data (ILD) to inform the development of dynamic theories of behavior change and intervention design: a case study of scientific and practical considerations. *Prevention Science*, 24:1659–1671, 2023.

[75] J J Prochaska, E A Vogel, and N Benowitz. Nicotine delivery and cigarette equivalents from vaping a juulpod. *Tobacco Control*, 31:e88–e93, 2022.

[76] C Proust, H Jacqmin-Gadda, J M G Taylor, J Ganiayre, and D Commenges. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*, 62(4):1014–1024, 2006.

[77] C Proust-Lima, H Amieva, and H Jacqmin-Gadda. Analysis of multivariate mixed longitudinal data: A flexible latent process approach. *British Journal of Mathematical and Statistical Psychology*, 66:470–487, 2013.

[78] C Proust-Lima, J-F Dartigues, and H Jacqmin-Gadda. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Statistics in Medicine*, 35:382–398, 2016.

[79] T Qian, A E Walton, L M Collins, P Klasnja, S T Lanza, I Nahum-Shani, M Rabbi, M A Russell, M A Walton, H Yoo, and S A Murphy. The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychol Methods*, 27(5):874–894, 2022.

[80] Klasnja P Almirall D Murphy SA Qian T, Yoo H. Estimating time-varying causal excursion effect in mobile health with binary outcomes. *Biometrika*, 108(3):507–527, 2021.

[81] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[82] M Rabbi, M Philyaw Kotov, R Cunningham, E E Bonar, I Nahum-Shani, P Klasnja, M Walton, and S Murphy. Toward increasing engagement in substance use data collection: Development of the substance abuse research assistant app and protocol for a microrandomized trial using adolescents and emerging adults. *JMIR Research Protocols*, 7(7), 2018.

[83] S L Rathbun, X Song, B Neustifter, and S Shiffman. Survival analysis with time-varying covariates measured at random times by design. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 62(3):419–434, 2013.

[84] J W Reich, A J Zautra, and M Davis. Dimensions of affect relationships: models and their integrative implications. *Review of General Psychology*, 7(1):66–83, 2003.

[85] R Reisenzein. Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology*, 67(3):525–539, 1994.

[86] N A Remington, L R Fabrigar, and P S Visser. Reexamining the circumplex model of affect. *Journal of Personality and Social Psychology*, 79(2):286–300, 2000.

[87] D Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

[88] D Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*, pages 62–74. CRC Press, 2012.

[89] D Rizopoulos. Joint modeling of longitudinal and time-to-event data with applications in R. https://www.drizopoulos.com/courses/EMC/ESP72.pdf, 2023. Accessed: 2024-06-17.

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

[91] D Rizopoulos, G Papageorgiou, and P Miranda Afonso. JMbayes2: Extended joint models for longitudinal and time-to-event data. 2024.

[92] D Rizopoulos, J M G Taylor, G Papageorgiou, and T M Morgan. Using joint models for longitudinal and time-to-event data to investigate the causal effect of salvage therapy after prostatectomy. *Statistical Methods in Medical Research*, 33(5):894–908, 2024.

[93] D Rizopoulos, G Verbeke, and E Lesaffre. Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(3):637–654, 2009.

[94] D Rizopoulos, G Verbeke, and G Molenberghs. Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, 66(1):20–29, 2010.

[95] M D Robinson, R L Irvin, M R Persich, and S Krishnakumar. Bipolar or independent? relations between positive and negative affect vary by emotional intelligence. *Affective Science*, 1(4):225–236, 2020.

[96] J Roy and X Lin. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, 56(4):1047–1054, 2000.

[97] D B Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

[98] D Rustand, J van Niekerk, E Teixeira Krainski, H Rue, and C Proust-Lima. Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested Laplace approximations. *Biostatistics*, 2023.

[99] N Saleheen, A A Ali, S M Hossain, H Sarker, S Chatterjee, B Marlin, E Ertin, M al'Absi, and S Kumar. puffMarker: A multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, page 999–1010, 2015.

[100] M Schmitt and G S Blum. *State/Trait Interactions*, pages 5206–5209. Springer International Publishing, Cham, 2020.

[101] J Shi, Z Wu, and W Dempsey. Assessing time-varying causal effect moderation in the presence of cluster-level treatment effect heterogeneity and interference. *Biometrika*, 110(3):645–662, 2023.

[102] M Signorelli, P Spitali, C Al-Khalili Szigyarto, the MARK-MD Consortium, and R Tsonaka. Penalized regression calibration: A method for the prediction of survival outcomes using complex longitudinal and high-dimensional data. *Statistics in Medicine*, 40(27):6178–6196, 2021.

[103] X Song, M Davidian, and A Tsiatis. An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics*, 3(4):511–528, 2002.

[104] A Stennett, N M Krebs, J Liao, J P Richie, and J E Muscat. Ecological momentary assessment of smoking behaviors in native and converted intermittent smokers. *American Journal on Addictions*, 27(2):131–138, 2018.

[105] A A Stone and S Shiffman. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, 16(3):199–202, 1994.

[106] J P Sy, J M G Taylor, and W G Cumberland. A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics*, 53(2):542–555, 1997.

[107] BO Taddé, H Jacqmin-Gadda, J-F Dartigues, D Commenges, and C Proust-Lima. Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease. *Biometrics*, 76(3):886–899, 2020.

[108] A M Tang, N S Tang, and D Yu. Bayesian semiparametric joint model of multivariate longitudinal and survival data with dependent censoring. *Lifetime Data Analysis*, 29(4):888–918, 2023.

[109] J M G Taylor, W G Cumberland, and J P Sy. A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, 89(427):727–736, 1994.

[110] J M G Taylor, Y Park, D P Ankerst, C Proust-Lima, S Williams, L Kestin, K Bae, T Pickles, and H Sandler. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213, 2013.

[111] J M G Taylor, J Shen, E H Kennedy, L Wang, and D E Schaubel. Comparison of methods for estimating the effect of salvage therapy in prostate cancer when treatment is given by indication. *Statistics in Medicine*, (2):257–74, 2014.

[112] T D Tran, E Lesaffre, G Verbeke, and J Duyck. Latent Ornstein-Uhlenbeck models for Bayesian analysis of multivariate longitudinal categorical responses. *Biometrics*, 77(2):689–701, 2021.

[113] TD Tran, E Lesaffre, G Verbeke, and J Duyck. Modeling local dependence in latent vector autoregressive models. *Biostatistics*, 22:148–163, 2019.

[114] A A Tsiatis and M Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834, 2004.

[115] A A Tsiatis, V Degruttola, and M S Wulfsohn. 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, 1995.

[116] J Tu and J Sun. Gaussian variational approximate inference for joint models of longitudinal biomarkers and a survival outcome. *Statistics in Medicine*, 42(3):316–330, 2023.

[117] P Vatiwutipong and N Phewchean. Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process. *Advances in Difference Equations*, (276), 2019.

[118] A Vehtari, A Gelman, and J Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27:1413–1432, 2017.

[119] C Vinci, L Li, C Wu, C Y Lam, L Guo, V Correa-Fernandez, et al. The assocation of positive emotion and first smoking lapse: an ecological momentary assessment study. *Health Psychology*, 36(11):1038–1046, 2017.

[120] P T von Hippel and J W Bartlett. Maximum likelihood multiple imputation: Faster imputations and consistent standard errors without posterior draws. *Statistical Science*, 36(1):400–420, 2021.

[121] X Wang, J O Berger, and D S Burdick. Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, 7(1):126—153, 2013.

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

[123] S Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[124] D Watson, L A Clark, and A Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[125] D Watson and A Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219–35, 1985.

[126] K Y Wong, D Zeng, and D Y Lin. Semiparametric latent-class models for multivariate longitudinal and survival data. *The Annals of Statistics*, 50(1):487–510, 2022.

[127] M S Wulfsohn and A A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1), 1997.

[128] W Ye, X Lin, and J M G Taylor. Semiparametric modeling of longitudinal measurements and time-to-event data—a two-stage regression calibration approach. *Biometrics*, 64:1238–1246, 2008.

[129] M Yu, N J Law, J M G Taylor, and H M Sandler. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14:835–862, 2004.

[130] X Yue and R A Kontar. Joint models for event prediction from time series and survival data. *Technometrics*, 63(4):477—486, 2020.

[131] D Zhang, M H Chen, J G Ibrahim, M E Boye, and W Shen. Bayesian model assessment in joint modeling of longitudinal and survival data with applications to cancer clinical trials. *Journal of Computational and Graphical Statistics*, 26(1):121–133, 2017.