Title
Multivariate Spatial Modeling of HIV Risk

Author
Flores, Martiniano Jose

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Multivariate Spatial Modeling of HIV Risk

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Martiniano Jose Flores

2018

ABSTRACT OF THE DISSERTATION

Multivariate Spatial Modeling
of HIV Risk

by

Martiniano Jose Flores
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2018
Professor Robert Erin Weiss, Chair

We analyze data from the Los Angeles LGBT Center, a community-based healthcare organization. When patients visit the clinic, they are given a comprehensive risk-assessment questionnaire. We develop three methods that allow us to identify the risk factors associated with HIV seroconversion and predict who is most likely to become HIV positive.

First, we construct a two-stage multivariate logistic regression model, where stage one models a patient's history of illicit drug use and their history of STIs other than HIV, and stage two models their risk of contracting HIV. Each stage of the model has ZIP code random effects that are correlated over space, and we propose a new statistic which we term the geometric mean ratio (GMR), which measures how much of the variability in the ZIP code random effects for HIV is explained by the stage one random effects. We find that the stage one random effects are negligible in the HIV model and that where a person lives is not predictive of their risk of contracting HIV.

Next, we jointly model a patient's time until HIV seroconversion with their clinic visit frequency through shared frailties. For patients that seroconvert, we do not observe the seroconversion time, only that it occurred within the interval between two visits. We show that if clinic visit frequency is correlated with survival, then the censoring is informative. We examine how the informativeness of the censoring depends on the frailty distributions. We find that patients who visit the clinic more frequently tend to have a higher probability

of contracting HIV, suggesting that patients are accurately assessing that they have a higher risk of disease.

Finally, we take twenty of the measurements from the risk assessment questionnaire and do a factor analysis to construct an overall measure of a patient's propensity for risky behavior. Because patients come to the clinic multiple times, we allow the factors to be correlated within a patient over time, and between patients over space. We then use the factor scores from one visit to predict whether or not a patient will seroconvert by their next visit. We show that this model is equivalent to a larger longitudinal factor model where the factors load onto HIV at one visit, and load onto all other outcomes at another visit. We show that the factor scores are predictive of future risk of HIV.

The dissertation of Martiniano Jose Flores is approved.

Marjan Javanbakht

Thomas R Belin

Sudipto Banerjee

Robert Erin Weiss, Committee Chair

University of California, Los Angeles

2018

In memory of my mother, Sharron

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work was supported in part by: The Ruth L. Kirschstein Institutional NRSA Training Grant 5T32AI737; the Center for HIV Identification, Prevention, and Treatment (CHIPTS) NIMH grant P30MH058107; the UCLA Center for AIDS Research (CFAR) grant 5P30AI028697, Core H; NSF/DMS grant 1513654; NSF/IIS grant 1562303; and NIH/NIEHS grant 1R01ES027027.

A version of the material presented in Chapter 2 has been submitted for publication with contributions from co-authors Drs. Robert E. Weiss, Matthew R. Beymer, and Sudipto Banerjee. Chapters 3 and 4 are currently in preparation for submission. The materials presented in Chapters 2, 3, and 4, were made possible thanks to the generous contribution of electronic health record data from the Los Angeles LGBT Center.

First and foremost, I would like to express my sincerest gratitude to my advisor Dr. Robert Weiss for his support, guidance, and patience throughout both the Master's and PhD programs. His experience, dedication, and attention to detail, have helped me develop as a statistician and academic writer, and for that I am incredibly grateful. I would also like to thank Dr. Matthew R. Beymer at the Los Angeles LGBT Center, without whom this dissertation would not be possible. I would also like to thank Drs. Sudipto Banerjee, Thomas Belin, and Marjan Javanbakht for serving on my dissertation committee for their support and guidance. I would also like to express gratitude to my friend and colleague Sitaram Vangala for being a sounding board for me throughout the dissertation.

I would also like to thank my friends and family for their support. I would especially like to thank my parents, my mother in law, and father in law for always believing in me, and whose love and support have made me into the person I am today. Finally, I would like to thank my wife Kun Li, for always being there to support me, and for always being my guiding light.

VITA

2011 Bachelor of Science, Microbiology. University of California, San Diego. La Jolla, California, USA

2013 Master of Science, Biostatistics. University of California, Los Angeles. Los Angeles, California, USA

PUBLICATIONS

Li L, Nugyen AT, Liang LJ, Lin C, Farmer S, and Flores M (2013). Mental health and family relations among people who inject drugs and their family members in Vietnam. *International Journal of Drug Policy* **24**, 545 – 549.

Reddy D, Njala J, Stocker P, Schooley A, Flores M, Tseng CH, Pfaff C, Jansen P, Mitsuyasu RT, and Hoffman RM (2014). High-risk human papillomavirus in HIV-infected women undergoing cervical cancer screening in Lilongwe, Malawi: a pilot study. *International Journal of STD & AIDS* **26**, 379 – 387.

Calkins KL, Havranek T, Kelley-Quon LI, Cerny L, Flores M, Grogan T, and Shew SB (2017). Low-dose parenteral soybean oil for the prevention of parenteral nutrition-associated liver disease in neonates with gastrointestinal disorders: A multicenter randomized controlled pilot study. *Journal of Parenteral and Enteral Nutrition* **41**, 404 – 411.

Vu JP, Larauche M, Flores M, Luong L, Norris J, Oh S, Liang LJ, Waschecck J, Pisegna JR, and Germano PM (2015). Regulation of Appetite, Body Composition, and Metabolic Hormones by Vasoactive Intestinal Polypeptide (VIP). *Journal of Molecular Neuroscience* **56**, 377 – 387.

Guerrero AD, Flores M, Vangala S and Chung JP (2016). Differences in the Association of Physical Activity and Children's Overweight and Obesity Status Among the Major Racial and Ethnic Groups of U.S. Children. *Health Education and Behavior* **44**, 411 – 420.

Ong ML, Purdy IB, Levit OL, Robinson DT, Grogan T, Flores M, and Calkins KL (2016). Two-year neurodevelopment and growth outcomes for preterm neonates who received low dose intravenous soybean oil. *The Journal of Parenteral and Enteral Nutrition* **42**, 352 – 360.

KL Calkins, DeBarber A, Steiner RD, Flores M, Grogan TR, Henning SM, Reyen L, and Venick RS (2017). Intravenous Fish Oil and Pediatric Intestinal Failure–Associated Liver Disease: Changes in Plasma Phytosterols, Cytokines, and Bile Acids and Erythrocyte Fatty Acids. *The Journal of Parenteral and Enteral Nutrition* **42**, 633 – 641.

Vangala S, Flores M, and Elashoff, DA (2018). The ABCs of Risk Score Calibration (Under Revision).

Flores MJ, Weiss RE, Banerjee S, and Beymer MR (2019). Bayesian tests for equality of generalized variances in multivariate spatial models (Under Revision).

Flores M, Weiss RE, and Beymer MR (2019). Multivariate spatial modeling of time to HIV acquisition and frequency of clinic visits (In preparation).

Ramos AP and Flores M (2019). Democratization's Heterogeneous Effects on Child Mortality: A Longitudinal Analysis of New Data, for 181 countries, 1970 – 2009 (Under Revision).

Flores M, Weiss RE, and Beymer MR (2019). Constructing an HIV risk score using spatiotemporal factor analysis models (In preparation).

Flores M, Weiss RE, and Collins MD (2019). Bayesian hierarchical random effects models for estimating ED50 in sea urchin embryos as a function of covariates (In preparation).

CHAPTER 1

Introduction

We analyze data from the Los Angeles LGBT Center collected between 2008 and 2014. When patients come to the clinic, they are given an 82-item risk assessment questionnaire that requests information about their demographic characteristics, sexual behaviors, recent illicit drug use, recent sexually transmitted infections (STIs), and information about their last two sexual partners. We restrict our analyses to patients living in Los Angeles County. Further, we only consider patients who have at least two visits to the clinic during the study period. We establish patient seronegativity with the first visit, and at least one more visit is needed to learn about patient visit frequency and whether or not they became HIV positive. It is often the case when patients visit the clinic after having contracted an STI that they schedule a follow-up visit within two weeks. Full-risk assessment questionnaires are not given at follow-up visits, and are therefore not included in our analyses. We track patients either until the end of the study period or until they become HIV positive. In total, we have approximately 10,000 patients across 270 ZIP codes in Los Angeles County. In this dissertation, we propose a number of models and methods that will allow us to learn about the characteristics of people who become HIV positive and to determine the extent to which where a person lives influences their probability of contracting HIV. All models are run in a fully Bayesian framework.

1.1 Scientific Contributions

It is known that people with STIs and people who use illicit drugs are at increased risk of HIV (Fleming and Wasserheit, 1999; Buchacz et al., 2005). Presumably such individuals are

engaging in risky behaviors more generally. We want to help the LBGT Center estimate which patients have an increased propensity for risky behaviors, how riskiness is distributed across Los Angeles county, and the role that riskiness plays in their probability of becoming HIV positive.

In Chapter 2, we use patients' STI and illicit drug use as proxies for riskiness in predicting their probability of becoming HIV positive. We allow for patients' risk of becoming HIV positive to be correlated over space to determine whether or not living in neighborhoods of high risk puts them at increased risk of HIV irrespective of their own behaviors. Chapter 2 uses a cross sectional version of the dataset that only considers whether or not patients become HIV positive by the end of the study, but because we have repeated visits on all patients, we have additional information about the timing of their infections.

Therefore in Chapter 3, we use a survival model to predict not only *if* patients become HIV positive, but *when*. We wish to learn whether a patient's frequency of visiting the clinic gives us any information about whether or not they're more likely to become HIV positive. One potential mechanism driving the correlation between clinic visits frequency and HIV risk is that a patient who comes to the clinic more frequently is generally a less risky person, and so clinic visit frequency should be *positively* correlated with seroconversion times. On the other hand, it may be the case that patients who visit the clinic more frequently are doing so because they have just engaged in or are more likely to engage in risky behaviors. This hypothesis predicts that seroconversion times should be *negatively* correlated with clinic visit frequency. We want to learn which of these mechanisms is more likely to be true.

In Chapter 4, we are interested in expanding the definition of what it means to be a risky individual beyond simply whether or not patients *ever* had an STI or *ever* used drugs. For example, it may be the case that injection drugs generally indicate much higher levels of riskiness than drugs such as cocaine or alcohol, and so we want a model that allows for each individual drug to separately tell us about a patients propensity for risky behavior. We also want to learn which sexual behaviors indicate that patients are more risky. Therefore, we use a factor analysis model to take 20 variables and reduce them to a lower dimensional set of factors. We estimate their factor scores at each visit and allow factors within a patient

to evolve over time. This allows us to examine each individual patient's risk trajectory over time and learn how a patient's factor scores at one visit predict HIV by their next visit. Technical details of the three methods described in this section are discussed at greater length in the next section.

1.2 Bayesian tests for equality of generalized variances in multivariate spatial models

In Chapter 2, we analyze a cross-sectional version of the dataset where we take the patients' final records on study and determine whether or not they've ever had an STI or used illicit drugs, and whether or not they became HIV positive by the end of the study. We jointly model history of STIs, history of illicit drug use, and HIV serostatus by the end of the study using a two-stage model. In stage 1, we jointly model history of STIs and illicit drug use with logistic regression models that have correlated ZIP code level random effects (McCulloch, 2008; Matheron, 1982). In stage two, we treat the random effects from stage 1 as covariates in a logistic regression model predicting HIV seroconversion by the end of the study. We incorporate a ZIP code level random effect in the HIV model to capture any of the spatial noise in HIV risk that was not picked up by the stage 1 random effects.

The total ZIP code level random effect in the HIV model is a linear combination of the stage 1 random effects and the residual ZIP code level random effect. To assess the extent to which the variability encapsulated in the covariance matrix of the total random effect (the *marginal covariance matrix*) is explained by the stage 1 random effects, we propose to take the ratio of the determinants of the covariance matrix of the total random effect after conditioning on the stage 1 random effects (the *conditional covariance matrix*) and the marginal covariance matrix. For interpretability, we raise the ratio by the inverse of the dimensionality of the matrix, which gives us a ratio of the geometric means of the eigenvalues of the conditional and marginal covariance matrices. We call this statistic the geometric mean ratio (GMR). Estimation of determinant ratios and related quantities have received some treatment in the literature (Wilks, 1932; SenGupta, 1987; SenGupta, 1987).

Typically, a question of interest is whether or not two determinants are equal. Hypothesis tests for equality of determinants are usually done with likelihood ratio tests. We show that the structure of our model provides a simple test for the equality of two determinants. In addition, we show that the GMR is bounded in the interval (0,1) and has interpretation similar to an intraclass correlation coefficient statistic.

1.3 Multivariate spatial modeling of interval-censored time-to-event data and clinic visit counts

In Chapter 3, we extend the model for HIV seroconversion. In addition to modeling whether or not patients become HIV positive, we model the time from the start of the study until they become HIV positive, and jointly model the seroconversion times with their frequency of clinic visits. For the patients who do seroconvert, we do not observe their actual seroconversion time, only that it occurred between two of their visits. Therefore, the seroconversion times are all interval censored. We model survival times as lognormal and clinic visit frequency by an approximate Poisson process, with correlation between seroconversion times and visit frequency modeled through shared frailty parameters (Cai et al., 2012; Liang et al., 2009; Liu et al., 2008; Sun et al., 2007; Zhang et al., 2007). We show that if clinic visit frequency is correlated with seroconversion time that the censoring is informative, and evaluate how the informativeness of the censoring is influenced by the distributions of the frailty parameters.

1.4 Developing a risk profile for HIV seroconversion using a spatio-temporal factor analysis

In Chapter 4, we extend the modeling approaches from Chapters 2 and 3. Rather than jointly modeling only one or two outcomes with HIV seroconversion, we reduce twenty of the outcomes from the risk assessment questionnaire into a lower dimensional set of factors that model aspects of a patient's propensity for risky behavior. The outcomes in the factor

model are a mix of discrete and continuous, and we treat all outcomes as functions of latent normal random variables (Ansari and Jedidi, 2000; Conti et al., 2014; Hu et al., 2004; Quinn, 2004). We have repeated measures on all patients and ZIP code information on patients at each visit. We model longitudinal correlation among factors within a patient and spatial correlation among factors between patients through Gaussian processes. We then model HIV seroconversion at a given visit with a probit model, treating the factors from the previous visit as covariates. The latent multivariate normal structure of the model allows us to derive a Gibbs sampler for most of the model parameters in the model, reducing the computational burden due to the large number of patients and ZIP codes.

1.5 Outline of the dissertation

The next three chapters are detailed presentations of the statistical methods described in Sections 1.1 – 1.4. Each chapter is a version of a manuscript that is in preparation for submission to peer-reviewed journals. Each chapter uses the same data set, and the spatial modeling is the same across all three chapters, so some of the material is repeated across chapters.

Finally, the dissertation finishes in Chapter 5 with a brief discussion of the conclusions and potential areas for future research. In particular, we developed the GMR in Chapter 2 the context of a linear regression of one multivariate normal random variable on another. We would like to extend the development of the GMR to the case of a general multivariate normal. Further, for the survival model in Chapter 3, we modeled seroconversion times as lognormal. This is mostly done to simplify the calculations, but this is not generally necessary. We would like to consider other parametric survival models and provide a formal test of which model fits the data best. Finally, for the factor model in Chapter 4, we would like to use latent variable models other than Probit to allow more flexibility in how outcomes load onto the factors.

CHAPTER 2

Bayesian Tests for Equality of Generalized Variances in Multivariate Spatial Models

Multivariate outcomes are common in studies of HIV acquisition (Bachireddy et al., 2014; Grinsztejn et al., 2014; Vergeynst et al., 2015). When multivariate outcomes are discrete, or a mix of continuous and discrete outcomes, one of the more common methods for jointly modeling them introduces correlation through random effects (Arminger and Küsters, 1988; Zhu and Weiss, 2013; Grover et al., 2015; Martins et al., 2016). To model multivariate outcomes, we can use a conditional approach, where a random effect is introduced in the model for one outcome and then treated as a covariate for the other outcomes (Wulfsohn and Tsiatis, 1997). This approach can also be used to model multivariate spatial data (Banerjee et al., 2014). We can also introduce separate random effects for each outcome and give the multivariate random effects a joint distribution, usually multivariate normal. This approach defines the marginal distributions and covariances directly. McCulloch (2008) discusses consequences of these approaches in the analysis of discrete outcomes.

With multivariate spatial data, it is difficult to directly model the cross covariances between multivariate spatial random effects in such a way that the resulting covariance matrix for all random effects at all spatial locations is positive definite. One way to model the cross covariances assumes that the random effects for each outcome have the same marginal multivariate normal distribution (Banerjee and Gelfand, 2002). Another approach is the linear model of coregionalization (LMC), which dates back at least to Matheron Matheron (1982), with many modern applications (Gneiting et al., 2010; Orton et al., 2014; Konomi et al., 2015). LMC approaches allow us to have different marginal distributions for the multivariate random effects and define joint distributions for multivariate spatial processes

as linear combinations of independent spatial processes.

We analyze medical record data on $10,083$ patients collected by the Los Angeles LGBT Center from December 2008 through December 2014. When patients come to the clinic, they are tested for HIV and other sexually transmitted infections and given an 82 item risk assessment questionnaire which requests basic demographic information, history of sexually transmitted infections, and history of drug use. The data are spatially indexed by ZIP code, and we use Census Bureau data to convert ZIP codes to ZIP Code Tabulation Areas (ZCTAs) to get centroid locations. Patients in our study need to come to the clinic at least twice; once to demonstrate that they are HIV negative, and at least once more to determine whether or not they became HIV positive.

We are primarily interested in learning how characteristics such as STIs or drug use are associated with whether or not patients become HIV positive. We are also interested in whether neighborhood characteristics affect HIV risk and how personal characteristics affect HIV risk. We jointly model a patient's history of STIs and drug use using logistic regression models with spatial random effects and an LMC to define the correlation structure between the random effects. We then treat the STI and drug use random effects as covariates in a logistic regression model for HIV and include an additional spatial error term to capture residual spatial variation not accounted for by the STI and drug use random effects.

A key research aim is to understand the spatial dependence of HIV risk among people and how this dependence is affected by STIs and drug use. For longitudinal data, Daniels and Zhao (2003) proposed taking a spectral decomposition of the random effects covariance matrix, using linear and log-linear models to model the resulting parameters as functions of covariates. For random intercept models, Hedeker et al. (2008) proposed log linear models for modeling the within and between subject variances as functions of covariates. They also derived intraclass correlation coefficients (ICCs) but because the ICCs are functions of covariates, the framework does not lend itself to an overall measure of the covariate effects on HIV risk.

To measure the overall variability in an $S \times S$ covariance matrix Σ, Wilks (1932) proposed

the determinant $|\Sigma|$, which he called the Generalized Variance (GV). The GV is sensitive to S, so SenGupta (1987) proposed the standardized GV (SGV) $|\Sigma|^{1/S}$ which is the geometric mean of the eigenvalues of Σ. Inference on the GV and SGV has used likelihood ratio tests in multivariate linear models (Iliopoulos and Kourouklis, 1999; Mathai, 1972; Bhandary, 1996, 2006; Specht, 1975; SenGupta, 1987; Hao and Krishnamoorthy, 2001). Bayesian methods for estimation and inference for GVs and SGVs are scant.

We propose the Geometric Mean Ratio (GMR) which compares the SGVs of the HIV random effect covariance matrix before and after conditioning on the STI and drug use random effects. The GMR is always between zero and one and has interpretation similar to an ICC. We are interested in testing the null hypothesis that GMR $= 1$ versus the alternative hypothesis that GMR < 1. Since the GMR is almost surely less than one, we derive a function of the model parameters that provides an equivalent test of GMR < 1. We also wish to know whether the STI or drug use random effects contribute more to the HIV spatial random effects SGV. GMR statistics can assess the effect of conditioning on one covariate at a time HIV random effects SGV and provide a test of equality of the effects of the STI and drug use random effects on the HIV spatial random effects SGV.

2.1 Notation and Model Formulation

For the i^{th} patient, where $i \in 1, \ldots, N$, let y_{i1} be a binary indicator for having had STIs either before or during the study period, where the STIs we consider are syphilis, gonorrhea, chlamydia, and herpes. Let y_{i2} be a binary indicator for drug use either before or during the study period, where the drugs we consider are methamphetamine, cocaine, ecstasy, and nitrates. Let y_{i3} be an indicator of having become HIV positive by the end of the study. Let x_i be the i^{th} patient's $p \times 1$ vector of covariates and T_i denote a patient's total time on study, defined as the number of years between their first and last visits during the study period. Lastly, let $s \in 1, \ldots, S$ index ZCTAs with S as the total number of ZCTAs in the dataset, and define $s(i) \in \{1, \ldots, S\}$ to be the ZCTA where the i^{th} patient lives.

We construct a multivariate generalized linear model to jointly model a patient's risk of

HIV with their STI and drug use behaviors. To start we jointly model y_{i1} and y_{i2} as Bernoulli random variables with success probabilities π_{i1} and π_{i2} using a bivariate hierarchical logistic regression model with correlated random effects. For $j = 1, 2$,

$$y_{ij}|\pi_{ij} \sim \text{Bern}(\pi_{ij})$$
$$\text{logit}\,(\pi_{ij}) = x_i'\boldsymbol{\alpha}_j + b_{s(i)j},$$

where $\boldsymbol{\alpha}_j$ is a vector of regression coefficients for $y_{ij}, j = 1, 2$, and $(b_{s(i)1}, b_{s(i)2})$ are correlated ZCTA level spatial random effects with prior distributions developed in the next section. We model y_{i3} conditional on STIs and drug use as a Bernoulli random variable with success probability π_{i3},

$$\begin{aligned} y_{i3}|\pi_{i3} &\sim \text{Bern}(\pi_{i3}) \\ \text{logit}\,(\pi_{i3}) &= x_i'\boldsymbol{\alpha}_3 + c_{s(i)} + \lambda_1 y_{i1} + \lambda_2 y_{i2} + \log(\text{T}_i) \end{aligned} \tag{2.1}$$

where λ_1 and λ_2 are regression coefficients, $c_{s(i)} = \beta_1 b_{s(i)1} + \beta_2 b_{s(i)2} + b_{s(i)3}$ is the full HIV spatial random effect, and $b_{s(i)3}$ is a residual spatial random effect that depends on the STI and drug use spatial random effects. Including $\log(\text{T}_i)$ on the right hand side of (2.1) means that we are modeling patient i's odds of HIV in a given year and allows the model to accommodate heterogeneity in patients' times on study.

In general, larger values of $b_{s(i)1}$, $b_{s(i)2}$, and $c_{s(i)}$ indicate that patient i lives in an area with higher rates of STIs, drug use, and HIV, respectively. Sex and Drug risk outcomes y_{i1} and y_{i2} are included in (2.1) to capture the non-spatial effects of the patient's STI and drug use on individual HIV risk. Large positive values of λ_1 and λ_2 indicate that having STIs or using drugs increase HIV risk, while large positive values of β_1 and β_2 indicate that living in a ZCTA with comparatively higher prevalences of STIs or drug use increases HIV risk.

2.2 Random Effect Distributions

Let $r, s \in 1, \ldots, S$, and $\boldsymbol{D}$ be a symmetric $S \times S$ matrix with (r, s) element d_{rs}, where d_{rs} is the distance in miles between the centroids of ZCTAs r and s. Let $\boldsymbol{b}_1 = (b_{11}, \ldots, b_{S1})', \boldsymbol{b}_2 =$

$(b_{12}, \ldots, b_{S2})'$, and $\boldsymbol{b}_3 = (b_{13}, \ldots, b_{S3})'$ be vectors of STI, drug use, and residual HIV spatial random effects.

We define a joint prior for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ as a multivariate Gaussian process using a Linear Model of Coregionalization (LMC). For all s, let $\text{Var}(b_{s1}) = \eta_1^2$, $\text{Var}(b_{s2}) = \eta_2^2$, $\text{Corr}(b_{s1}, b_{s2}) = \rho$. Define

$$\boldsymbol{T} = \text{Var}\begin{pmatrix} b_{s1} \\ b_{s2} \end{pmatrix} = \begin{pmatrix} \eta_1^2 & \rho\eta_1\eta_2 \\ \rho\eta_1\eta_2 & \eta_2^2 \end{pmatrix}$$

as the within ZCTA covariance matrix for b_{s1} and b_{s2} and let $\boldsymbol{A}$ be the Cholesky decomposition of $\boldsymbol{T}$, with $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{T}$ and $\boldsymbol{A}$ lower triangular. We model $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ as functions of independent spatial processes $\boldsymbol{w}_1' = (w_{11}, \ldots, w_{1S})$ and $\boldsymbol{w}_2' = (w_{21}, \ldots, w_{2S})$ with $S \times S$ exponential decay correlation matrices $\boldsymbol{R}_1 = \exp(-\phi_1 \boldsymbol{D})$ and $\boldsymbol{R}_2 = \exp(-\phi_2 \boldsymbol{D})$ respectively, with decay parameters $\phi_1, \phi_2 > 0$. Set

$$\begin{pmatrix} b_{s1} \\ b_{s2} \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} w_{s1} \\ w_{s2} \end{pmatrix},$$

$$\begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} \Bigg| \phi_1, \phi_2 \sim \text{N}\left(\begin{pmatrix} \boldsymbol{0}_{S\times 1} \\ \boldsymbol{0}_{S\times 1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{0}_{S\times S} \\ \boldsymbol{0}_{S\times S} & \boldsymbol{R}_2 \end{pmatrix} \right),$$

and let $\boldsymbol{B}_j = \boldsymbol{a}_j \boldsymbol{a}_j^T$, $j = 1, 2$, where $\boldsymbol{a}_j$ is the j^{th} column of $\boldsymbol{A}$. Then

$$\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \Bigg| \phi_1, \phi_2, \rho \sim \text{N}\left(\boldsymbol{0}_{2S\times 1}, \ \sum_j \left[\boldsymbol{B}_j \otimes \boldsymbol{R}_j \right] \right),$$

where

$$\sum_j \left[\boldsymbol{B}_j \otimes \boldsymbol{R}_j \right] = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_{11} &= \eta_1^2 \boldsymbol{R}_1, \\ \boldsymbol{\Sigma}_{12} &= \rho\eta_1\eta_2 \boldsymbol{R}_1, \\ \boldsymbol{\Sigma}_{22} &= \eta_2^2 \left(\rho^2 \boldsymbol{R}_1 + (1 - \rho^2) \boldsymbol{R}_2 \right). \end{aligned}$$

This specification models the within ZCTA correlation between the random effects as a function of ρ and the across ZCTA correlation for the random effects as functions of ϕ_1 and ϕ_2 and ρ.

We a priori model $\boldsymbol{b}_3$ as a univariate Gaussian Process with a scaled exponential decay covariance matrix $\boldsymbol{\Sigma}_{33} = \eta_3^2 \exp(-\phi_3 \boldsymbol{D})$ with decay parameter $\phi_3 > 0$ and scale parameter $\eta_3^2 > 0$

$$\boldsymbol{b}_3 | \phi_3, \eta_3^2 \sim \mathrm{N}_S(\boldsymbol{0}_{S\times 1}, \boldsymbol{\Sigma}_{33}).$$

2.2.1 Random Effect Variances and Covariances

Let $\boldsymbol{c} = \beta_1 \boldsymbol{b}_1 + \beta_2 \boldsymbol{b}_2 + \boldsymbol{b}_3$ be the $S \times 1$ vector of marginal spatial random effects for HIV, $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)'$, $\boldsymbol{\eta} = (\eta_1^2, \eta_2^2, \eta_3^2)'$, and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. Then

$$\boldsymbol{c} | \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\eta} \sim \mathrm{N}_S(\boldsymbol{0}, \boldsymbol{\Sigma}_{cc}),$$

where

$$\boldsymbol{\Sigma}_{cc} = (\beta_1 \eta_1 + \rho \beta_2 \eta_2)^2 \boldsymbol{R}_1 + \beta_2^2 \eta_2^2 (1 - \rho^2) \boldsymbol{R}_2 + \boldsymbol{\Sigma}_{33}$$

is positive definite and $\mathrm{Var}(c_s) = \beta_1^2 \eta_1^2 + 2\rho\beta_1\beta_2\eta_1\eta_2 + \beta_2^2\eta_2^2 + \eta_3^2$. The joint distribution of $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{c}$ is

$$\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \boldsymbol{c} \end{pmatrix} \Bigg| \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\eta} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{0}_{S\times 1} \\ \boldsymbol{0}_{S\times 1} \\ \boldsymbol{0}_{S\times 1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{c1}^T \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{c2}^T \\ \boldsymbol{\Sigma}_{c1} & \boldsymbol{\Sigma}_{c2} & \boldsymbol{\Sigma}_{cc} \end{pmatrix} \right),$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{c1} &= (\beta_1 + \rho\beta_2) \boldsymbol{R}_1, \\ \boldsymbol{\Sigma}_{c2} &= \rho\beta_1 \boldsymbol{R}_1 + \beta_2 \left[\rho^2 \boldsymbol{R}_1 + (1 - \rho^2) \boldsymbol{R}_2 \right]. \end{aligned}$$

To assess the effect of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ on $|\boldsymbol{\Sigma}_{cc}|$, we also need the following conditional densities,

$$\begin{aligned} \boldsymbol{c} | \boldsymbol{b}_1 &\sim \mathrm{N}\left(\boldsymbol{\Sigma}_{c1} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{b}_1, \boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{c1} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{c1} \right), \\ \boldsymbol{c} | \boldsymbol{b}_2 &\sim \mathrm{N}\left(\boldsymbol{\Sigma}_{c2} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{b}_2, \boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{c2} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{c2} \right), \end{aligned}$$

and

$$\boldsymbol{c}|\boldsymbol{b}_1, \boldsymbol{b}_2 \sim \mathrm{N}\left(\beta_1 \boldsymbol{b}_1 + \beta_2 \boldsymbol{b}_2, \boldsymbol{\Sigma}_{33}\right).$$

2.2.2 Effective Range

For a spatial exponential decay process with parameter ϕ, the correlation between the random effects for ZCTAs r and s is $\exp(-\phi d_{rs})$. A large value of ϕ indicates that the correlation decreases more rapidly across space. Define the effective range of the process as the distance Φ where the correlation has decreased to 0.05. For the STI random effects, $\Phi_1 = -\log(0.05)/\phi_1$. For drug use, the correlation between random effects b_{r2} and b_{s2} is

$$\mathrm{Cor}\left(b_{r2}, b_{s2}\right) = \rho^2 \exp\left(-\phi_1 d_{rs}\right) + \left(1 - \rho^2\right) \exp\left(-\phi_2 d_{rs}\right). \tag{2.2}$$

For HIV risk, the correlation between c_{r3} and c_{s3} is

$$\mathrm{Cor}\left(b_{r3}, b_{s3}\right) = \frac{\sum_{j=1}^{3} v_j \exp\left(-\phi_j d_{rs}\right)}{\sum_{j=1}^{3} v_j}, \tag{2.3}$$

where $v_1 = (\beta_1\eta_1 + \rho\beta_2\eta_2)^2$, $v_2 = \beta_2^2\eta_2^2(1-\rho^2)$, and $v_3 = \eta_3^2$ are the marginal variances of the STI, drug use, and residual random effects. To calculate the effective ranges Φ_2 and Φ_3 for the drug use and HIV random effects, set the left hand side of equations (2.2) and (2.3) equal to 0.05 and solve numerically for d_{rs}.

2.3 Geometric Mean Ratio Statistics

The vector of marginal random effects $\boldsymbol{c}$ for HIV is a function of $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$. To assess the contribution of the STI and drug use random effects $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ to the covariance matrix $\boldsymbol{\Sigma}_{cc}$, we take the ratio of the SGVs of the marginal covariance matrix Σ_{cc} and conditional covariance matrix Σ_{33}. Define the GMR as

$$\mathrm{GMR} = \frac{\left|\mathrm{Var}\left(\boldsymbol{c}|\boldsymbol{b}_1, \boldsymbol{b}_2\right)\right|^{1/S}}{\left|\mathrm{Var}\left(\boldsymbol{c}\right)\right|^{1/S}}. \tag{2.4}$$

The GMR directly measures how much smaller the geometric mean eigenvalue of the HIV spatial covariance matrix becomes after conditioning on $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$. Similarly, the effect of

conditioning on just $\boldsymbol{b}_1$ or $\boldsymbol{b}_2$ can be calculated by replacing the conditional covariance matrix $\text{Var}(\boldsymbol{c}|\boldsymbol{b}_1, \boldsymbol{b}_2)$ in (2.4) with $\text{Var}(\boldsymbol{c}|\boldsymbol{b}_1)$ or $\text{Var}(\boldsymbol{c}|\boldsymbol{b}_2)$, which gives

$$\text{GMR}_1 = \frac{\left|\boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{c1}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{c1}\right|^{1/S}}{\left|\boldsymbol{\Sigma}_{cc}\right|^{1/S}},$$

and

$$\text{GMR}_2 = \frac{\left|\boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{c2}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{c2}\right|^{1/S}}{\left|\boldsymbol{\Sigma}_{cc}\right|^{1/S}},$$

respectively.

The GMR statistics GMR, GMR_1, and GMR_2 have a maximum value of one when the conditional covariance matrices equal to the marginal covariance matrix, meaning that conditioning on $\boldsymbol{b}_1$ or $\boldsymbol{b}_2$ does not reduce the SGV at all. Conversely, the GMR statistics have a minimum value of zero when the conditional covariance matrix has determinant zero, meaning that all of the spatial heterogeneity in HIV risk is explained by STIs and drug use. Thus, the GMR statistics have interpretation similar to an intraclass correlation coefficient or to a regression model R^2.

To determine whether conditioning on $\boldsymbol{b}_1$ or $\boldsymbol{b}_2$ has a greater effect on reducing the SGV of the marginal covariance matrix, we can also look at the ratio GMRR_{12} of GMR_1 to GMR_2,

$$\text{GMRR}_{12} = \frac{\text{GMR}_1}{\text{GMR}_2} = \frac{\left|\boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{c1}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{c1}\right|^{1/S}}{\left|\boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{c2}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{c2}\right|^{1/S}}.$$

When $\text{GMRR}_{12} = 1$, the STIs and drug use random effects have the same effect on the marginal covariance matrix Σ_{cc}. If $\text{GMRR}_{12} > 1$, then the SGV of $\text{Var}(\boldsymbol{c}|\boldsymbol{b}_1)$ is larger than the SGV of $\text{Var}(\boldsymbol{c}|\boldsymbol{b}_2)$, and drug use has a greater effect on the spatial heterogeneity in HIV risk.

2.3.1 Hypothesis Tests for GMR Statistics

We want to test the hypothesis that GMR = 1, meaning that the spatial distribution of STIs and drug use have no effect on spatial heterogeneity in HIV risk. Since $\text{P}(\text{GMR} < 1) = 1$, we cannot do this directly. However GMR = 1 if and only if $\beta_1 = \beta_2 = 0$, so a Bayes Factor

B_{GMR} in favor of the null H_0 $\beta_1 = \beta_2 = 0$ against the alternative, $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ is equivalent to a Bayes Factor in favor of GMR $= 1$ against $H_1 : \text{GMR} < 1$. We calculate the Bayes Factor B_{GMR} with the density ratio first proposed by Dickey and Lientz (1970) and Dickey (1971) and extended by Verdinelli and Wasserman (1995). Let $\boldsymbol{Y}$ be the data and $\omega = (\beta_1, \beta_2)$, and ψ be the collection of all parameters excluding ω. A priori β_1 and β_2 are independent of all other parameters so the Bayes factor B_{GMR} is,

$$B_{\text{GMR}} = \left. \frac{p(\omega|\boldsymbol{Y})}{p(\omega)} \right|_{\omega=\mathbf{0}} \tag{2.5}$$

Thus, B_{GMR} is the ratio of the posterior density over the prior density evaluated at zero. Values of B_{GMR} greater than one indicate data support for $H_0 : \text{GMR} = 1$ against $H_1 : \text{GMR} < 1$.

Unlike the GMR, both GMR_1 and GMR_2 are complex functions of $\boldsymbol{\beta}$ and both are almost surely less than 1, therefore we cannot use $\boldsymbol{\beta}$ to test $\text{GMR}_1 = 1$ versus $\text{GMR}_1 < 1$ or $\text{GMR}_2 = 1$ versus $\text{GMR}_2 < 1$. In contrast, the GMRR is supported on the positive real line, so we can test GMRR $= 1$ versus GMRR $\neq 1$ by directly setting $\omega = \text{GMRR}$ in (2.5) and evaluating the density ratio at $\text{GMRR}_{12} = 1$.

2.4 Prior Distributions

For each of the regression parameters $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, and $\boldsymbol{\lambda}$, we set multivariate normal priors with mean zero and covariance matrix identity. For the variance parameters η_j^2, we set half-normal priors to prevent the variances from becoming too large,

$$\eta_j^2 \sim \text{N}(0,1)\, \mathbf{1}_{\left\{\eta_1^2 > 0\right\}},$$

for $j = 1, 2, 3$.

For ρ, the within-ZCTA correlation between the STI and drug use random effects, we rescale a Beta$(4,4)$ distribution to be in the interval $(-1, 1)$

$$p(\rho) \propto \left(\frac{\rho - 1}{2}\right)^3 \left(1 - \frac{\rho - 1}{2}\right)^3 \mathbf{1}_{\{-1 \leq \rho \leq 1\}}.$$

For the effective range parameters $\Phi_j > 0$ for $j = 1, 2, 3$, the minimum observed distance between any two ZCTAs in our dataset is approximately 0.5 miles. While it is possible that $\Phi_j < 0.5$, it is unlikely that the data could provide evidence for this. The Φ_j are functions of the decay parameters ϕ_j and the within ZCTA correlation ρ. Setting Gamma priors on the ϕ_j,

$$\phi_j \sim \text{Gamma}(2, 2),$$

induces more than 99% of the prior mass to be above 0.5 for all the respective range parameters Φ_j.

We want the GMR to have a prior distribution that is roughly uniform on the unit interval. Setting normal priors for β_1 and β_2,

$$\beta_1 \sim \text{N}(0, 0.85),$$
$$\beta_2 \sim \text{N}(0, 0.85),$$

with our specification induces prior densities for GMR plotted in Figure 2.1. While the prior for the GMR is roughly uniform on the unit interval, this is not true for GMR_1 and GMR_2.

2.5 Results from the Joint Model of STIs, Drug Use, and HIV

We sampled from the posterior distribution using a random walk Metropolis-Hastings sampler programmed in `R` version 3.3.1 using the `Rcpp` package version 0.12.7. We ran four chains with $250,000$ iterations, giving a total effective sample size at least $10,000$ for all parameters including the spatial random effects $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$. Convergence within each chain was checked using Raftery and Lewis statistics (Raftery and Lewis, 1992) and Geweke statistics (Geweke, 1992), and convergence between chains was checked using Gelman statistics (Gelman and Rubin, 1992). No convergence issues were found.

The next section details the results of the data analysis in two parts. First, we examine the posterior distributions for the regression coefficients from the logistic regressions. Then we take a more detailed look at the spatial decay parameters and random effects and at the posterior distribution of the GMR statistics.

2.5.1 Regression Results

We included a patient's baseline age and race as covariates in the model. For age, we fit a cubic B-spline with knots at the quintiles. For HIV, we also included binary indicators of whether a patient had STIs or used drugs. For a patient with median random effects of zero, posterior predicted probabilities for HIV are plotted as a function of age, STIs, and drug use in Figure 2.2 (top). The mean predicted probabilities and 95% pointwise credible bands for having STIs or of using drugs as a function of age are also plotted in Figure 2.2 (bottom). The observed proportions of individuals at each age who seroconverted, had an STI, or used drugs are plotted as points. If the probabilities are not a function of age, then the predicted lines should be flat and age B-spline coefficients would be jointly zero. To test the relationships between age and HIV, STIs, and drug use, we calculate the Bayes Factors supporting no relationship with age.

The risk of HIV seroconversion decreases with age, and the estimated probability decreases by half from 0.02 at age 18 to 0.01 at age 45. In contrast, probability of ever having had and STI increases almost two fold from 0.35 to 0.66 for the same age group. The probability of ever having used drugs peaks at 0.55 at age 30 and decreases to about 0.30 by age 75, indicating that if a patient has not yet used drugs by age 30, they are not likely to use drugs in the future. The log Bayes Factor supporting no age trend for HIV is -1.55, which is weak evidence against the null. For STIs and drug use, the log Bayes Factors are -114 and -38 which are very strong evidence against their respective nulls.

Summaries of the regression coefficient posteriors for HIV seroconversion and history of STIs and drug use are presented in the top part of Table 2.1. Compared to Whites, African Americans and Hispanics have higher odds of having had STIs and of acquiring HIV, and Hispanics have higher odds of drug use.

Consistent with the literature, having STIs (Fleming and Wasserheit, 1999) and using drugs (Buchacz et al., 2005; Plankey et al., 2007; Fisher et al., 2011) are both associated with an increased risk of HIV. Our results further indicate that the prevalence of STIs and drug use where people live does not contribute to risk of HIV acquisition. Thus a patient's own

behaviors are important while whether they live in close proximity with others who engage in risky behaviors is not.

2.5.2 Spatial Parameters

The bottom part of Table 2.1 gives posterior summaries of $\text{Var}(b_{s1})$, $\text{Var}(b_{s2})$, and $\text{Var}(c_s)$, and the associated range parameters Φ_1, Φ_2, Φ_3. The 95% posterior intervals are $(0.93, 5.88)$ for the STI effective range Φ_1 and $(1.05, 7.58)$ for the drug use effective range Φ_2, units in miles. Thus, the spatial correlation for the STI and drug use random effects drops off rapidly with distance. The posterior 95% interval for the HIV range Φ_3 is comparatively larger at $(3.51, 38.6)$.

The 95% posterior interval for the within ZCTA correlation ρ between the STI and drug use random effects is $(-0.23, 0.31)$, indicating that the sign of the random effects correlation for STI and drug use within a given ZIP code is uncertain. The log Bayes Factor B_ρ (2.5) supporting the null hypothesis $\rho = 0$ against the alternative that $\rho \neq 0$ is 0.74, which represents weak evidence in favor of the null.

Prior and posterior densities for the GMR statistics are plotted in Figure 2.3. Conditioning on the STIs and drug use random effects $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ does not significantly reduce the size of the marginal covariance matrix $\boldsymbol{\Sigma}_{cc}$. The log Bayes Factor $\log B_{\text{GMR}} = 3.49$, supporting the null hypothesis that GMR = 1. The posterior distributions for GMR_1 and GMR_2 are also very close to one, and the posterior distribution for GMRR shows that there is no difference in the effect of the STI spatial effects $\boldsymbol{b}_2$ and the drug use random effects $\boldsymbol{b}_2$ on $\boldsymbol{\Sigma}_{cc}(\log B_{\text{GMRR}} = 2.25)$.

2.6 Discussion

A key result of our analyses is that the coefficients associated with the STI and drug use ZIP code level random effects were not different from zero. After controlling for a patient's age and race, this means that a patient's own behaviors are much more important for estimating

their risk of HIV than the behaviors of people who live around them. From a public health perspective, this suggests that targeting high risk neighborhoods is much less important then targeting individually risky people.

A benefit of using the GMR instead of the ratio of determinants is that it is not heavily dependent on the dimensionality S of the covariance matrix, so sampling more spatial points does not lead to a drastically different GMR. Since the GMR statistics are complex functions of model parameters, Bayesian methods provide natural advantages in estimation. We modeled all spatial covariances using exponential decay, but it is possible to calculate the GMR for any situation where we can calculate the determinants for the conditional and marginal covariance matrices, including for CAR models and multivariate longitudinal data.

For any two positive definite matrices A and B, the Minkowski determinant theorem (Marcus and Minc, 1992) implies that $(|A|/|A+B|)^{1/S}+(|B|/|A+B|)^{1/S} \leq 1$. Thus, strictly speaking the GMR is not the fraction of the HIV spatial variance due to the STI and drug use random effects. However, in our data analysis and in preliminary simulations (not presented), $P\left((|A|/|A+B|)^{1/S}+(|B|/|A+B|)^{1/S} > 0.95\right) = 1$ so near equality holds in the inequality and it is not unreasonable to interpret the GMR as the fraction of the HIV spatial variance explained by the STI and drug use random effects. To ensure $(|A|/|A+B|)^{1/S}+(|B|/|A+B|)^{1/S} = 1$, rather than looking at $|\boldsymbol{\Sigma}_{33}\boldsymbol{\Sigma}_{cc}^{-1}|^{1/S}$, we can instead look at $\text{tr}(\boldsymbol{\Sigma}_{33}\boldsymbol{\Sigma}_{cc}^{-1})/S$ (See appendix C). However, the trace $\text{tr}(\boldsymbol{\Sigma}_{33}\boldsymbol{\Sigma}_{cc}^{-1})/S$ is not directly interpretable as a comparison of the eigenvalues of $\boldsymbol{\Sigma}_{33}$ and $\boldsymbol{\Sigma}_{cc}$, which is one of the more desirable properties of the GMR.

Acknowledgments

This work was supported by: The Ruth L. Kirschstein Institutional NRSA Training Grant 5T32AI737; the Center for HIV Identification, Prevention, and Treatment (CHIPTS) NIMH grant P30MH058107; the UCLA Center for AIDS Research (CFAR) grant 5P30AI028697, Core H; NSF/DMS grant 1513654; NSF/IIS grant 1562303; and NIH/NIEHS grant 1R01ES027027. We would like to thank the Los Angeles LGBT Center for providing the data for this analysis.

Tables and Figures

Table 2.1: Summary of the posterior for STI, Drug Use, and HIV logistic regression models.

	Odds Ratio (95% CI)		
Variable	History of STIs	Drug Use	HIV Infection
Race			
White	REF	REF	REF
Black	1.370 (1.035, 1.762)	1.009 (0.687, 1.399)	1.648 (1.145, 2.277)
Hispanic	1.709 (1.462, 1.985)	1.235 (1.006, 1.487)	1.634 (1.305, 2.024)
Other	0.910 (0.699, 1.154)	0.805 (0.587, 1.080)	0.920 (0.620, 1.280)
STI Spatial	–	–	0.975 (0.756, 1.244)
STI Personal	–	–	2.215 (1.690, 2.831)
Drug Use Spatial	–	–	0.960 (0.746, 1.213)
Drug Use Personal	–	–	4.233 (3.199, 5.410)
		Mean (95% CI)	
SpatialVariance	0.168 (0.101, 0.263)	0.266 (0.140, 0.455)	0.945 (0.686, 1.271)
Range	2.760 (0.931, 5.884)	3.324 (1.054, 7.577)	13.71 (3.509, 38.60)

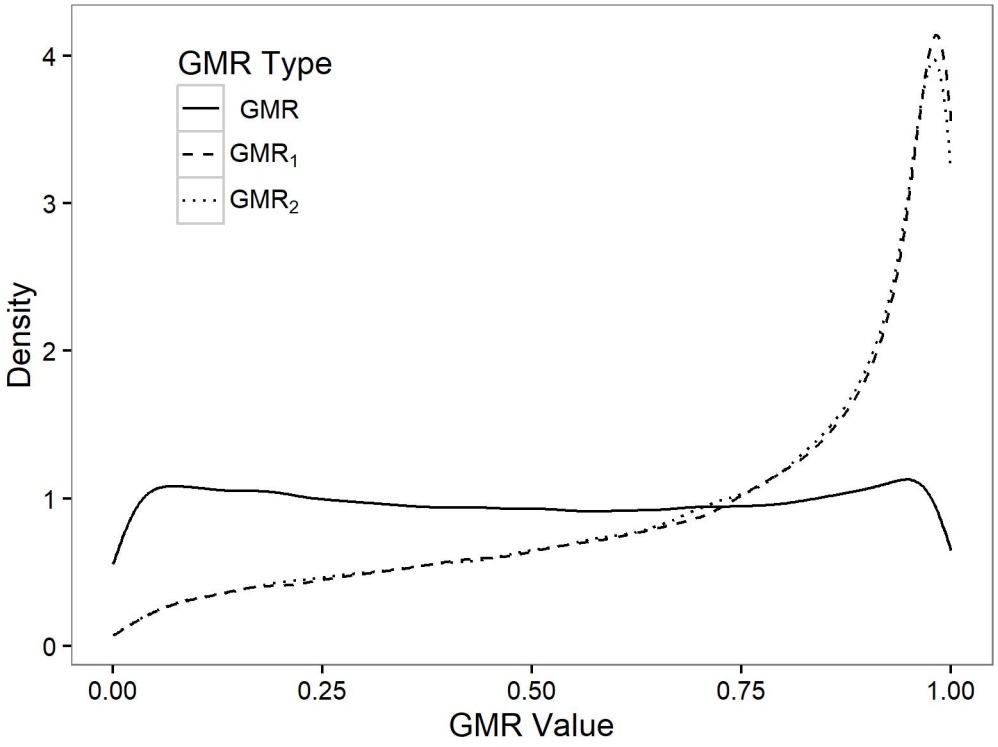

Figure 2.1: Prior distribution for GMR, GMR_1 and GMR_2.

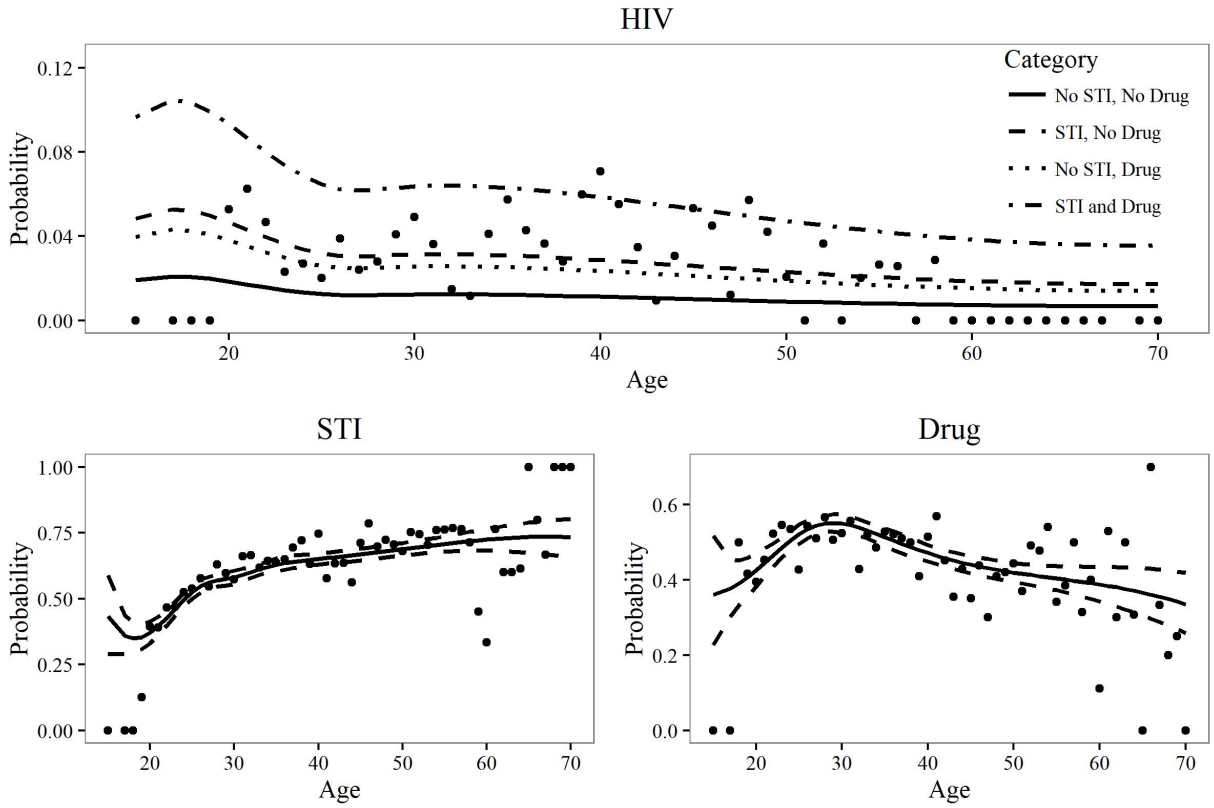

Figure 2.2: Posterior mean predicted probabilities for HIV by STI and Drug use (Top). Posterior predicted mean probability and 95% pointwise bands for having STIs (Bottom Left), and having used drugs (Bottom Right) as a function of age. For each age, the proportions of individuals who experienced the events are plotted as points.

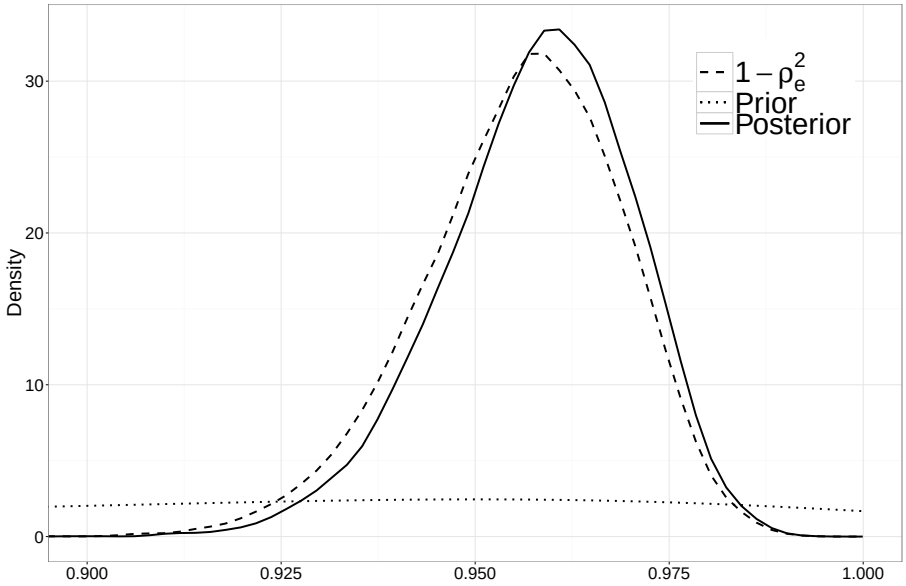

Figure 2.3: Plots of the prior and posterior for the GMR statistics. Log Bayes Factor in favor of GMR = 1 is 3.49, indicating good posterior evidence for GMR = 1. Log Bayes Factor in favor of GMRR = 1 is 2.25, indicating that the data support GMRR = 1.

Appendix A - Properties of Total Random Effect c for HIV

We derive the marginal distribution for $\boldsymbol{c} = \beta_1 \boldsymbol{b}_1 + \beta_2 \boldsymbol{b}_2 + \boldsymbol{b}_3$ and the conditional distribution for $\boldsymbol{c}|\boldsymbol{b}_1, \boldsymbol{b}_2$, where $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ are the spatial random effects for STIs, drug use, and HIV, respectively. We start with the following result.

Let $X \sim \mathrm{N}_{2S}(\mathbf{0}, \Sigma_x)$ and let $Y \sim \mathrm{N}_S(\mathbf{0}, \Sigma_y)$ be independent of X. Let $Z = AX + BY$, where $A_{S\times 2S}$ and $B_{S\times S}$ are known matrices. Then

$$Z \sim \mathrm{N}\left(\mathbf{0}, A\boldsymbol{\Sigma}_x A' + B\boldsymbol{\Sigma}_y B'\right)$$

and

$$Z|X \sim \mathrm{N}\left(AX, B\boldsymbol{\Sigma}_y B'\right)$$

Proof The marginal distribution of Z is follows from the properties of multivariate normal

distributions. For the conditional distribution, we have

$$\begin{aligned}\mathrm{Cov}(Z, X) &= \mathrm{Cov}(AX + BY, X) \\ &= A\boldsymbol{\Sigma}_x\end{aligned}$$

which implies that

$$\begin{aligned}Z|X &\sim \mathrm{N}\left(\bar{\mu}, \bar{\boldsymbol{\Sigma}}\right) \\ \bar{\mu} &= A\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_x^{-1}X \\ &= AX \\ \bar{\boldsymbol{\Sigma}} &= A\boldsymbol{\Sigma}_x A' + B\boldsymbol{\Sigma}_y B' - A\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_x A' \\ &= B\boldsymbol{\Sigma}_y B'.\end{aligned}$$

∎

Since $\boldsymbol{b}_3$ is independent of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$,

$$\begin{aligned}\boldsymbol{c} &= \beta_1\boldsymbol{b}_1 + \beta_2\boldsymbol{b}_2 + \boldsymbol{b}_3 \\ &= \begin{pmatrix}\beta_1\boldsymbol{I}_S & \beta_2\boldsymbol{I}_S & \boldsymbol{I}_S\end{pmatrix}\begin{pmatrix}\boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \boldsymbol{b}_3\end{pmatrix}, \\ \mathrm{Var}\left(\boldsymbol{c}\right) &= \left(\beta_1\eta_1 + \rho\eta_2\beta_2\right)^2\boldsymbol{R}_1 + \\ &\qquad \beta_2^2\eta_2^2(1-\rho^2)\boldsymbol{R}_2 + \eta_3^2\boldsymbol{\Sigma}_{33} \\ \mathrm{E}\left(\boldsymbol{c}|\boldsymbol{b}_1, \boldsymbol{b}_2\right) &= \beta_1\boldsymbol{b}_1 + \beta_2\boldsymbol{b}_2 \\ \mathrm{Var}\left(\boldsymbol{c}|\boldsymbol{b}_1, \boldsymbol{b}_2\right) &= \boldsymbol{\Sigma}_{33}\end{aligned}$$

Appendix B - Properties of the Geometric Mean Ratio (GMR)

From (2.4), we have

$$\begin{aligned}\text{GMR} &= \frac{|\boldsymbol{\Sigma}_{33}|^{1/S}}{|\boldsymbol{\Sigma}_{cc}|^{1/S}} \\ &= \left|\boldsymbol{\Sigma}_{33}\boldsymbol{\Sigma}_{cc}^{-1}\right|^{1/S} \\ &= \frac{|A|^{1/S}}{|A+B|^{1/S}},\end{aligned}$$

where $A = \boldsymbol{\Sigma}_{33}$ and $B = \left(\beta_1\eta_1 + \rho\beta_2\eta_2\right)^2 \boldsymbol{R}_1 + \beta_2^2\eta_2^2(1-\rho^2)\boldsymbol{R}_2$. Further, since A and B are positive definite,

$$0 \leq \frac{|A|}{|A+B|} \leq \frac{|A|}{|A|+|B|} \leq 1,$$

so the GMR is between 0 and 1. Further it approaches zero as $|A|$ approaches zero, which occurs as η^2 or ϕ_3 approach zero. The GMR is equal to one when $B = 0$, which happens if and only if β_1 and β_2 are both equal to zero, as we prove below.

Theorem 1 *GMR* $= 1$ *if and only if* $\beta_1 = \beta_2 = 0$.

Proof Suppose $\beta_1 = \beta_2 = 0$. Then

$$\begin{aligned}\text{GMR} &= \frac{|\text{Var}\left(\boldsymbol{c}|\boldsymbol{b}_1, \boldsymbol{b}_2\right)|^{1/S}}{|\text{Var}\left(\boldsymbol{c}\right)|^{1/S}}, \\ &= \frac{|\boldsymbol{\Sigma}_{33}|^{1/S}}{\left|\left(\beta_1\eta_1 + \rho\beta_2\eta_2\right)^2 \boldsymbol{R}_1 + \beta_2^2\eta_2^2(1-\rho^2)\boldsymbol{R}_2 + \boldsymbol{\Sigma}_{33}\right|^{1/S}}, \\ &= 1.\end{aligned}$$

Now, suppose that GMR $= 1$. Then

$$|\boldsymbol{\Sigma}_{33}|^{1/S} = \left|\left(\beta_1\eta_1 + \rho\beta_2\eta_2\right)^2 \boldsymbol{R}_1 + \beta_2^2\eta_2^2(1-\rho^2)\boldsymbol{R}_2 + \boldsymbol{\Sigma}_{33}\right|^{1/S}$$

By the Minkowski determinant theorem, for two non-negative definite $S \times S$ Hermitian matrices A and B it holds that

$$|A+B|^{1/S} \geq |A|^{1/S} + |B|^{1/S}.$$

Repeated applications of the theorem yield

$$\begin{aligned}
|\boldsymbol{\Sigma}_{33}|^{1/S} &= \left|(\beta_1 + \rho\beta_2)^2 \boldsymbol{R}_1 + \beta_2^2(1-\rho^2)\boldsymbol{R}_2 + \boldsymbol{\Sigma}_{33}\right|^{1/S}, \\
&\geq \left|(\beta_1 + \rho\beta_2)^2 \boldsymbol{R}_1 + \beta_2^2(1-\rho^2)\boldsymbol{R}_2\right|^{1/S} + |\boldsymbol{\Sigma}_{33}|^{1/S}, \\
&\geq \left|(\beta_1 + \rho\beta_2)^2 \boldsymbol{R}_1\right|^{1/S} + \left|\beta_2^2(1-\rho^2)\boldsymbol{R}_2\right|^{1/S} + |\boldsymbol{\Sigma}_{33}|^{1/S}, \\
&= (\beta_1 + \rho\beta_2)^2 |\boldsymbol{R}_1|^{1/S} + \beta_2^2(1-\rho^2) |\boldsymbol{R}_2|^{1/S} + |\boldsymbol{\Sigma}_{33}|^{1/S}.
\end{aligned}$$

This implies that $(\beta_1 + \rho\beta_2)^2 |\boldsymbol{R}_1|^{1/S} + \beta_2^2(1-\rho^2) |\boldsymbol{R}_2|^{1/S} = 0$, and since $\boldsymbol{R}_1$, $\boldsymbol{R}_2$, and $\boldsymbol{\Sigma}_{33}$ are positive definite, this implies that $\beta_1 = \beta_2 = 0$. ∎

As a corollary to the Minkowski determinant theorem,

$$\frac{|A|^{1/S}}{|A+B|^{1/S}} + \frac{|B|^{1/S}}{|A+B|^{1/S}} \leq 1.$$

This implies that strictly speaking, the GMR is not the proportion of the geometric mean of the eigenvalues of $\boldsymbol{\Sigma}_{cc}$ coming from $\boldsymbol{\Sigma}_{33}$. As an alternative to the GMR, we could take the arithmetic mean eigenvalue of $\boldsymbol{\Sigma}_{33}\boldsymbol{\Sigma}_{cc}^{-1}$, which is just $\text{tr}(A(A+B)^{-1})/S$, Unlike with the GMR,

$$\frac{\text{tr}\left(A\,(A+B)^{-1}\right)}{s} + \frac{\text{tr}\left(B\,(A+B)^{-1}\right)}{s} = 1,$$

So this is a true proportion of the trace that is due to the stage one random effects. However, since the trace of a product is not a product of the traces, this is not directly interpretable in terms of the eigenvalues of $\boldsymbol{\Sigma}_{33}$ and $\boldsymbol{\Sigma}_{cc}$ in the way that the GMR is.

CHAPTER 3

Multivariate spatial modeling of interval-censored time-to-event data and clinic visit counts

We analyze electronic health records of Los Angeles County residents from the Los Angeles LGBT Center. We wish to jointly model a patient's risk of HIV with their frequency of clinic visits. Patients must have at least two visits to be included in the analysis. Their first visit establishes seronegativity at the start of the study and at least one more visit is needed to assess serostatus later in the study and to allow us to estimate visit frequency. All survival times for seroconverters in our data are interval censored, meaning we only know that they became HIV positive between two clinic visits. A common assumption in analyses of interval-censored data is that the censoring is non-informative, meaning that aside from knowing the event time lies in the interval, the interval conveys no additional information about the distribution of survival (Gómez et al., 2009; Oller et al., 2004; Zhang and Sun, 2010). If the censoring is actually informative, it is important to model the informativeness appropriately to avoid biased estimates of survival (Campigotto and Weller, 2014).

Modeling informative censoring is typically done by specifying a joint distribution for the event time and censoring time. For right censored data, one option is to assume that the marginal distribution of the censoring time is of the same family as the conditional distribution of the censoring time given survival, however with the latter depending on the parameters in the survival density (Bompotas et al., 2017; Siannis et al., 2005). Zhang et al. (2005) use proportional hazards models for the survival and censoring times, modeling correlation with random effects. Similar approaches can be used to model informative censoring for recurrent events (Zeng et al., 2014) and in interval censored data (Sinha et al., 1999; Zhao et al., 2015).

To jointly model survival times with clinic visit frequency, one method would be to treat a patient's number of visits per year as a fixed covariate in the survival model (Verity et al., 1995). A drawback of this framework is that it prevents us from making inferences on visit frequency. It also ignores the likely substantial measurement error in visit frequency. If we instead treat the survival outcome as a terminal event and the individual visits as recurrent events, then survival and clinic visits can be modeled using correlated frailties. This can be done using fully parametric models (Belot et al., 2014; Cowling et al., 2006; Crowther, 2017; Ma and Krings, 2008) or semi-parametric proportional hazards models (Huang and Liu, 2007; Król et al., 2016) or by writing the survival likelihood as a Poisson likelihood and jointly modeling the survival and count outcomes as correlated Poisson random variables (Aitkin and Clayton, 1980; Chib and Winkelmann, 2001; Sunethra and Sooriyarachchi, 2016).

We model HIV seroconversion times as lognormal random variables and the number of clinic visits by approximating a Poisson process that has been truncated at zero. We model correlation between clinic visits and seroconversion with correlated random effects. In general, seroconversion times have a long right tail, and patients who have a longer time on study will have a larger total number of clinic visits, so we include the patient's time on study as an offset. The model gives rise to informative censoring when survival times are correlated with clinic visit frequency.

Suppose survival times are negatively correlated with clinic visit frequency. Clinic visit frequency is inversely proportional to the average length of time between visits, so patients with shorter intervals will tend to have shorter survival times, suggesting that their unknown seroconversion time is likely earlier in any given interval. In contrast, a person with longer average times between visits will likely have seroconverted later in a given interval. Thus, the correlation between the survival and clinic visit random effects is an important parameter in our model because it controls both the marginal correlation between survival and clinic visits as well as the informativeness of the censoring.

Our work builds on the existing literature in some key ways. First, we use a fully Bayesian model, which aids in estimation of the model parameters. Second, we believe that the inter-visit times follow a Poisson process, which is equivalent to modeling all of the inter-visit

times as exponentially distributed. Because one visit is required for patients to enter the study, we treat the first visit as fixed. Further, patients must have at least two total visits, so a patient's total number of visits will then follow a zero-truncated Poisson distribution, which to our knowledge has not received attention in the literature for jointly modeling the visit process with survival. Finally, the existing literature has explored in detail the effect that informative visit times have on parameter estimates in regression models. We build on these results with simulation studies that show how the informativeness of the censoring is affected by the right censoring rate, the clinic visit random effect variance, and the strength of the correlation between the survival and clinic visit random effects.

The next section presents a detailed description of the model, including random effect distributions and prior distributions. Section 3.2 calculates the marginal correlation between a person's survival time and clinic visits. Section 3.3 analytically demonstrates that informative censoring arises when the clinic visit rate is correlated with survival time and section 3.4 shows that the informativeness of the censoring is influenced by the strength of the correlation between survival times and clinic visits, the right censoring rate, and the distribution of clinic visits. Section 3.5 presents the results of our data analysis, and section 3.6 offers concluding remarks.

3.1 Notation and Model Formulation

For $i \in 1, \ldots, N$, where N is the number of patients, let y_{i1} be the log of patient i's unknown survival time, y_{i2} be patient i's number of clinic visits after their first visit during the study period, counted either until they become HIV positive or until the time of patient i's last visit, and x_i be an $(M+1) \times 1$ baseline covariate vector for person i. Let T_i be the time of their last visit measured in years since their initial visit and let $s(i) \in 1, \ldots, S$ be the ZIP Code Tabulation Area (ZCTA) for patient i, where $S = 270$ is the total number of ZCTAs in the data set.

We model log survival time y_{i1} as normal

$$y_{i1} = x_i'\boldsymbol{\alpha}_1 + \epsilon_{i1} + \delta_{i1} + b_{s(i)1} \tag{3.1}$$

$$\epsilon_{i1}|\sigma^2_{y1} \sim \mathrm{N}\left(0, \sigma^2_{y1}\right) \tag{3.2}$$

$$\delta_{i1}|\sigma^2_{\delta 1} \sim \mathrm{N}\left(0, \sigma^2_{\delta 1}\right), \tag{3.3}$$

where $\boldsymbol{\alpha}_1 = (\alpha_{10}, \alpha_{11}, \ldots, \alpha_{1M})'$ is an $(M+1) \times 1$ vector of unknown regression coefficients, $b_{s(i)1}$ is a ZCTA level random effect and δ_{i1} is a subject level random effect with variance $\sigma^2_{\delta 1}$. The random effect δ_{i1} will not be identifiable but is used to induce correlation between y_{i1} and y_{i2}. We discuss treatment of δ_{i1} in detail shortly.

Let patient i's clinic visit schedule be the result of a Poisson process. Then the total number of visits will follow a zero-truncated Poisson distribution. We approximate the zero-truncated Poisson distribution by modeling y_{i2} as a Poisson random variable with mean $\exp(\lambda_i)T_i$, with $\exp(\lambda_i)$ as the expected number of visits per year,

$$f(y_{i2}|\lambda_i) = \frac{[\exp(\lambda_i)T_i]^{y_{i2}} \exp(-\exp(\lambda_i)T_i)}{(y_{i2})!} \tag{3.4}$$

with

$$\lambda_i = x_i'\boldsymbol{\alpha}_2 + \delta_{i2} + b_{s(i)2}, \tag{3.5}$$

where $\boldsymbol{\alpha}_2 = (\alpha_{20}, \alpha_{21}, \ldots, \alpha_{2M})^T$ is an $(M+1) \times 1$ vector of unknown regression coefficients, $b_{s(i)2}$ is a second spatially correlated ZCTA level random effect that is also correlated with $b_{s(i)1}$ and δ_{i2} is a non-spatial random effect correlated with δ_{i1} and distributed

$$\delta_{i2}|\delta_{i1}, \sigma_{\delta 1}, \sigma_{\delta 2}, \rho_\delta \sim \mathrm{N}\left(\frac{\sigma_{\delta 2}}{\sigma_{\delta 1}}\rho_\delta \delta_{i1}, \left(1-\rho^2_\delta\right)\sigma^2_{\delta 2}\right), \tag{3.6}$$

where $\sigma^2_{\delta 2}$ is the variance of δ_{i2} unconditional on δ_{i1} and ρ_δ is the correlation between δ_{i1} and δ_{i2}. Including δ_{i2} in (3.5) induces overdispersion in y_{i2}, and induces non-spatial correlation between y_{i1} and y_{i2} when ρ_δ is nonzero and allows a test of whether visit rate and seroconversion time are independent by testing $H_0 : \rho_\delta = 0$. The Poisson approximation to the zero-truncated Poisson is justified in the appendix.

In model (3.1) and (3.3), none of δ_{i1}, ϵ_{i1}, $\sigma^2_{\delta 1}$, or σ^2_{y1} are identified. Only $\varepsilon_{i1} = \delta_{i1} + \epsilon_{i1}$ and $\sigma^2_1 = \sigma^2_{y1} + \sigma^2_{\delta 1}$ are identified with $\varepsilon_{i1} \sim \mathrm{N}(0, \sigma^2_1)$. Integrating out δ_{i1} gives a bivariate normal distribution for (y_{i1}, δ_{i2}),

$$\begin{pmatrix} y_{i1} \\ \delta_{i2} \end{pmatrix} \Bigg| \boldsymbol{\alpha}_1, b_{s(i)1}, \rho, \sigma^2_1, \sigma^2_{\delta 2} \sim \mathrm{N}\left(\begin{pmatrix} x_i'\boldsymbol{\alpha}_1 + b_{s(i)1} \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_1 & \rho\sigma_1\sigma_{\delta 2} \\ \rho\sigma_1\sigma_{\delta 2} & \sigma^2_{\delta 2} \end{pmatrix} \right), \tag{3.7}$$

where $\rho = \rho_\delta \left(\sigma^2_{\delta 1}/\sigma^2_1\right)^{1/2}$. For the observed survival data, we do not observe the exact survival time y_{i1}; instead, we only observe the event $l_i < y_{i1} < r_i$, where l_i is the log of the second to last visit time and $r_i = \log(T_i)$ if patients seroconvert, or for patients who do not seroconvert, $l_i = \log(T_i)$ and $r_i = \infty$. Therefore conditional on $\boldsymbol{\alpha}_1$, $b_{s(i)1}$, and σ^2_1, the observed data for survival are Bernoulli random variables with success probabilities

$$\begin{aligned} P(l_i < y_{i1} < r_i | \boldsymbol{\alpha}_1, b_{s(i)1}, \delta_{i1}, \sigma^2_{y1}) = \Phi &\left(\frac{r_i - \left(x_i'\boldsymbol{\alpha}_1 + b_{s(i)1}\right)}{\sigma_1} \right) - \\ \Phi &\left(\frac{l_i - \left(x_i'\boldsymbol{\alpha}_1 + b_{s(i)1}\right)}{\sigma_1} \right), \end{aligned} \tag{3.8}$$

where $\Phi(\cdot)$ denotes the CDF of a standard normal random variable.

3.1.1 Likelihood

Let $\boldsymbol{y}_1 = (y_{11}, \ldots, y_{N1})$, $\boldsymbol{y}_2 = (y_{12}, \ldots, y_{N2})$, $\boldsymbol{\delta}_2 = (\delta_{i1}, \ldots, \delta_{N2})$, $\boldsymbol{b}_1 = (b_{11}, \ldots, b_{S1})$, $\boldsymbol{b}_2 = (b_{12}, \ldots, b_{S2})$ $\boldsymbol{\theta}_1 = (\boldsymbol{\alpha}_1, \boldsymbol{b}_1, \sigma_1)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}_2, \boldsymbol{b}_2, \sigma_{\delta 2})$, where $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are $S \times 1$ vectors of HIV and clinic visit spatial random effects with s^{th} elements b_{s1} and b_{s2}, $s \in 1, \ldots, S$. In model (3.1) – (3.6), conditional on δ_i, y_{i1} is independent of y_{i2}. The likelihood of the observed data is

$$L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{\delta}_2) = \prod_i P(l_i < y_{i1} < r_i | \boldsymbol{\theta}_1, \delta_{i2}) f(y_{i2} | \boldsymbol{\theta}_2, \delta_{i2}) f(\delta_{i2} | \boldsymbol{\theta}_2), \tag{3.9}$$

where (3.9) follows from the conditional independence assumptions.

3.1.2 ZCTA Level Random Effect Distributions

Let $r, s \in 1, \ldots, S$ index ZCTAs and let d_{rs} be the distance in miles between the centroids of ZCTAs r and s. Let $\mathrm{Var}(b_{s1}) = \eta^2_1$, $\mathrm{Var}(b_{s2}) = \eta^2_2$, and $\mathrm{Corr}(b_{s1}, b_{s2}) = \rho_b$. Define the within

ZCTA covariance matrix $\boldsymbol{T}_b$ between the ZCTA level HIV and clinic visit random effects

$$\boldsymbol{T}_b = \text{Var}\begin{pmatrix} b_{s1} \\ b_{s2} \end{pmatrix} = \begin{pmatrix} \eta_1^2 & \rho_b\eta_1\eta_2 \\ \rho_b\eta_1\eta_2 & \eta_2^2 \end{pmatrix}. \tag{3.10}$$

We jointly model $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ as multivariate normal using the linear model of coregionalization (Matheron, 1982),

$$\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \Bigg| \, \eta_1^2, \eta_2^2, \phi_1, \phi_2, \rho_b \sim \text{N}\left(\begin{pmatrix} \boldsymbol{0}_{S\times 1} \\ \boldsymbol{0}_{S\times 1} \end{pmatrix}, \begin{pmatrix} \eta_1^2\boldsymbol{R}_1 & \rho_b\eta_1\eta_2\boldsymbol{R}_1 \\ \rho_b\eta_1\eta_2\boldsymbol{R}_1 & \eta_2^2\left(\rho_b^2\boldsymbol{R}_1 + (1-\rho_b^2)\boldsymbol{R}_2\right) \end{pmatrix} \right), \tag{3.11}$$

where the r, s entry of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are $\exp\left(-\phi_1 d_{rs}\right)$ and $\exp\left(-\phi_2 d_{rs}\right)$ respectively and ϕ_1 and ϕ_2 are decay parameters.

3.1.3 Prior Distributions For Regression Parameters

The maximum time a patient could be on the study is 6 years, and as 96% of the data are right censored, we specify a Normal$(3, 0.5^2)$ prior for the intercept term α_{10} in the survival model, which puts more than 95% of the prior mass for the survival times above 6 years, giving a high a priori right-censoring rate. Similarly, choosing a Normal$(1, 0.5^2)$ prior for α_{20} puts 95% of the prior mass between one and four visits per year. For the other regression coefficients α_{1m} and α_{2m}, $m \in 1, \ldots, M$, we set N$(0, 1)$ priors.

For the non-spatial variance parameters σ_1^2 and $\sigma_{\delta 2}^2$ and the spatial variance parameters η_1^2 and η_2^2, we set half standard normal priors $\sigma_j^2 \sim \text{N}(0,1)\mathbf{1}\left\{\sigma_j^2 > 0\right\}$, $\eta_j^2 \sim \text{N}(0,1)\mathbf{1}\left\{\eta_j^2 > 0\right\}$, $j = 1, 2$. For the decay parameters ϕ_1 and ϕ_2, larger values result in faster decay in the correlation between ZCTA level random effects with increasing distance. We set gamma priors $\phi_j \sim \text{Gamma}\,(2, 2)$, $j = 1, 2$, which specifies a priori that the correlation between ZCTA level random effects should decay to less than 0.05 between 1 and 25 miles with 95% prior probability. For the correlation parameters, ρ_b and ρ_δ, we re-scale Beta$(4, 4)$ distributions to the interval $(-1, 1)$,

$$p(\rho_b) \propto \left(\frac{\rho_b - 1}{2}\right)^3 \left(1 - \frac{\rho_b - 1}{2}\right)^3 \mathbf{1}\left\{-1 \leq \rho_b \leq 1\right\}, \tag{3.12}$$

$$p(\rho_\delta) \propto \left(\frac{\rho_\delta - 1}{2}\right)^3 \left(1 - \frac{\rho_\delta - 1}{2}\right)^3 \mathbf{1}\left\{-1 \leq \rho_\delta \leq 1\right\}. \tag{3.13}$$

3.2 Covariance Calculations

Because random effects and residuals are unobserved, the spatial and non-spatial correlations ρ_b and ρ_δ are not as easily interpretable for clinicians as the correlation $\text{Cor}(\exp(y_{i1}), y_{i2})$ between survival and clinic visits unconditional on residuals and random effects. We calculate $\text{Cor}(\exp(y_{i1}), y_{i2})$ using iterated expectations and covariances. First, the joint distribution of $\varepsilon_{i1} + b_{s(i)1}$ and $\delta_{i2} + b_{s(i)2}$ in the survival and clinic visits models is

$$\begin{pmatrix} \varepsilon_{i1} + b_{s(i)1} \\ \delta_{i2} + b_{s(i)2} \end{pmatrix} \Bigg| \rho, \rho_b, \sigma_1^2, \sigma_2^2, \eta_1^2, \eta_2^2 \sim \text{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \gamma_1^2 & \gamma_{12} \\ \gamma_{12} & \gamma_2^2 \end{pmatrix} \right) \tag{3.14}$$

where

$$\gamma_1^2 = \sigma_1^2 + \eta_1^2 \tag{3.15}$$

$$\gamma_{12} = \rho\sigma_1\sigma_2 + \rho_b\eta_1\eta_2 \tag{3.16}$$

$$\gamma_2^2 = \sigma_2^2 + \eta_2^2, \tag{3.17}$$

and $\text{Cor}(\epsilon_{i1} + \delta_{i1} + b_{s(i)1}, \delta_{i2} + b_{s(i)2}) \equiv \rho_\gamma = \gamma_{12}/(\gamma_1\gamma_2)$. This allows us to calculate the mean and variance of y_{i1} and y_{i2} unconditional on random effects. For y_{i1},

$$\text{E}\left[\exp(y_{i1}) | \boldsymbol{\alpha}_1, \gamma_1^2\right] = \exp\left(x_i'\boldsymbol{\alpha}_1 + \frac{1}{2}\gamma_1^2\right) \tag{3.18}$$

and

$$\text{Var}\left(\exp(y_{i1}) | \boldsymbol{\alpha}_1, \gamma_1^2\right) = \left\{\text{E}\left[\exp(y_{i1}) | \boldsymbol{\alpha}_1, \gamma_1^2\right]\right\}^2 \left(\exp\left(\gamma_1^2\right) - 1\right). \tag{3.19}$$

For y_{i2}, we use conditional expectation and variance formulas

$$\text{E}\left[y_{i2} | \boldsymbol{\alpha}_2, \gamma_2^2\right] = \exp\left(x_i'\boldsymbol{\alpha}_2 + \frac{1}{2}\gamma_2^2\right) \tag{3.20}$$

and

$$\begin{aligned} \text{Var}\left(y_{i2} | \boldsymbol{\alpha}_2, \gamma_2^2\right) &= \text{E}\left[\text{Var}\left(y_{i2} | \boldsymbol{\alpha}_2, \delta_{i2}, b_{s(i)2}\right)\right] + \text{Var}\left(\text{E}\left[y_{i2} | \boldsymbol{\alpha}_2, \delta_{i2}, b_{s(i)2}\right]\right), & (3.21) \\ &= \text{E}\left[\exp(\lambda_i)T_i\right] + \text{Var}\left(\exp(\lambda_i)T_i\right), & (3.22) \\ &= \text{E}\left[y_{i2} | \boldsymbol{\alpha}_2, \gamma_2^2\right] + \left\{\text{E}\left[y_{i2} | \boldsymbol{\alpha}_2, \gamma_2^2\right]\right\}^2 \left(\exp\left(\gamma_2^2\right) - 1\right). & (3.23) \end{aligned}$$

We use the law of total covariance to calculate $\mathrm{Cov}\left(\exp\left(y_{i1}\right), y_{i2}\right)$

$$\begin{aligned}
\mathrm{Cov}(\exp\left(y_{i1}\right), y_{i2}|\boldsymbol{\alpha}_1, \gamma_1^2, \boldsymbol{\alpha}_2, \gamma_2^2) &= \mathrm{E}\big[\mathrm{Cov}(\exp\left(y_{i1}\right), y_{i2}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_{i1}, \delta_{i2}, b_{s(i)1}, b_{s(i)2})\big] + \\
&\quad \mathrm{Cov}\big[\mathrm{E}\left\{\exp\left(y_{i1}\right)|\boldsymbol{\alpha}_1, \delta_{i1}, b_{s(i)1}, \sigma_{y1}^2\right\}, \mathrm{E}(y_{i2}|\boldsymbol{\alpha}_2, \delta_{i2}, b_{s(i)2})\big] \qquad (3.24) \\
&= \exp\left(x_i'\boldsymbol{\alpha}_1 + \frac{1}{2}\sigma_{y1}^2\right)\exp\left(x_i'\boldsymbol{\alpha}_2 + \log(T_i)\right) \times \\
&\quad \mathrm{Cov}\big[\exp\big(\delta_{i1} + b_{s(i)1}\big), \exp\big(\delta_{i2} + b_{s(i)2}\big)\big] \qquad (3.25) \\
&= \exp\left(x_i'\boldsymbol{\alpha}_1 + \frac{1}{2}\sigma_{y1}^2\right)\exp\left(x_i'\boldsymbol{\alpha}_2 + \log(T_i)\right) \times \\
&\quad \left\{\exp\left(\frac{1}{2}\left[\sigma_{\delta1}^2 + \sigma_{\delta2}^2 + \eta_1^2 + \eta_2^2 + 2\rho_\delta\sigma_{\delta1}\sigma_{\delta2} + 2\rho_b\eta_1\eta_2\right]\right) - 1\right\} \qquad (3.26) \\
&= \exp\left(x_i'\boldsymbol{\alpha}_1 + \frac{1}{2}\left[\sigma_{y1}^2 + \sigma_{\delta1}^2 + \eta_1^2\right]\right) \times \\
&\quad \exp\left(x_i'\boldsymbol{\alpha}_2 + \log(T_i) + \frac{1}{2}\left[\sigma_{\delta2}^2 + \eta_2^2\right]\right) \times \qquad (3.27) \\
&\quad \left\{\exp\left(\rho_\delta\sigma_{\delta1}\sigma_{\delta2} + \rho_b\eta_1\eta_2\right) - 1\right\} \\
&= \mathrm{E}\left[y_{i1}|\boldsymbol{\alpha}_1, \gamma_1^2\right]\mathrm{E}\left[y_{i2}|\boldsymbol{\alpha}_2, \gamma_2^2\right]\left[\exp\left(\gamma_{12}\right) - 1\right], \qquad (3.28)
\end{aligned}$$

where the first term in (3.24) is zero because y_{i1} and y_{i2} are independent conditional on the random effects $\delta_{i2}, b_{s(i)1}$, and $b_{s(i)2}$, equation (3.26) follows from the covariance between two lognormal random variables, and (3.28) follows by substituting $\rho\sigma_1$ for $\rho_\delta\sigma_{\delta1}$. We calculate the marginal correlation $\mathrm{Cor}\left(\exp\left(y_{i1}\right), y_{i2}\right)$ using (3.19), (3.23), and (3.28). The marginal correlation does not depend on $x_i'\boldsymbol{\alpha}_1$, but is monotone increasing (decreasing) in $x_i'\boldsymbol{\alpha}_2$ if $\rho_\gamma > 0$ ($\rho_\gamma < 0$). Furthermore,

$$0 \leq \left|\mathrm{Cor}\left(\exp\left(y_{i1}\right), y_{i2}\right)\right| \leq \left|\mathrm{Cor}\left(\exp\left(y_{i1}\right), \exp(\delta_{i2})\right)\right| \leq \rho_\gamma, \qquad (3.29)$$

where

$$\mathrm{Cor}\left(\exp\left(y_{i1}\right), \exp(\lambda_i)\right) = \frac{\exp\left(\gamma_{12}\right) - 1}{\sqrt{\left[\exp\left(\gamma_1^2\right) - 1\right]\left[\exp\left(\gamma_2^2\right) - 1\right]}}, \qquad (3.30)$$

is the correlation between two lognormal random variables. Figure 3.1 plots $\mathrm{Cor}\left(\exp\left(y_{i1}\right), \exp(\lambda_i)\right)$ as a function of γ_1 and γ_2, with $\gamma_1 = \gamma_2$ and $\rho_\gamma = 0.25$, 0.50, 0.75, and 0.90. Even when ρ_γ is large, the maximum value that $\mathrm{Cor}\left(y_{i1}, y_{i2}\right)$ can take decreases rapidly as the variances increase.

3.3 Correlation Between Survival and Clinic Visits Leads to Informative Censoring

Non-informative censoring implies that given the endpoints (l_i, r_i) of the censoring interval, the density for log survival y_{i1} is the marginal density $f(y_{i1}|\boldsymbol{\alpha}_1, b_{s(i)1}, \sigma_1^2)$ truncated to the interval (l_i, r_i). However, y_{i1} is correlated with subject i's visit rate λ_i through δ_{i2}. The joint conditional density for y_{i1} and δ_{i2} given the censoring interval is

$$f(y_{i1}, \delta_{i2}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, b_{s(i)1}, b_{s(i)2}, \sigma_1^2, \sigma_{\delta 2}^2, \rho_\delta, l_i, r_i) = \\ f(y_{i1}|\delta_{i2}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, b_{s(i)1}, b_{s(i)2}, \sigma_1^2, \sigma_{\delta 2}^2, \rho_\delta, l_i, r_i) f(\delta_{i2}|\boldsymbol{\alpha}_2, b_{s(i)2}, \sigma_{\delta 2}^2, l_i, r_i), \quad (3.31)$$

where the first term on the right hand side of (3.31) is the conditional density $f(y_{i1}|\delta_{i2}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, b_{s(i)1}, b_{s(i)2}, \sigma_1^2, \sigma_2^2, \rho_\delta)$ of y_{i1} given δ_{i2} truncated to the interval (l_i, r_i). To calculate the distribution of y_{i1} given the interval (l_i, r_i) *unconditional* on δ_{i2}, we need to integrate the joint density (3.31) with respect to δ_{i2},

$$f(y_{i1}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, b_{s(i)1}, b_{s(i)2}, \sigma_1^2, \sigma_2^2, \rho_\delta, l_i, r_i) = \\ \int f(y_{i1}|\delta_{i2}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, b_{s(i)1}, b_{s(i)2}, \sigma_1^2, \sigma_2^2, \rho_\delta, l_i, r_i) f(\delta_{i2}|\boldsymbol{\alpha}_2, b_{s(i)2}, \sigma_2^2, l_i, r_i) \quad (3.32)$$

If $\rho_\delta = 0$, then the first term in the integrand of (3.31) is $f(y_{i1}|\boldsymbol{\alpha}_1, b_{s(i)1}, \sigma_1^2)$ truncated to the interval (l_i, r_i) and the second term integrates to one. Therefore, when $\rho_\delta = 0$, the censoring is not informative.

Conversely, when ρ_δ is nonzero, then the censoring will be informative. We demonstrate this by comparing the integral (3.32) with $f(y_{i1}|\boldsymbol{\alpha}_1, b_{s(i)1}, \sigma_1^2)$ restricted to the interval (l_i, r_i). We cannot evaluate the integral in closed form so we approximate the integral numerically. We compare the densities for different values of ρ_δ and censoring intervals (l_i, r_i) and plot the results in Figure 3.2. For small interval widths the differences between the two densities are not large, irrespective of the value of ρ_δ. However, for large interval widths, even modest correlations make the censoring highly informative, which leads to biased estimates of model parameters and survival times. We further examine the behavior of the bias at the population level with the following simulation studies.

3.4 Simulation Studies: Informativeness of Censoring

We want to evaluate how the informativeness of the censoring varies with the correlation ρ_γ, the right censoring rate, and the distribution of the clinic visit mean λ_i. We simulate data from the model in section 3.1, setting the spatial random effects $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ equal to zero and fixing the linear predictors $x_i'\boldsymbol{\alpha}_1$ and $x_i'\boldsymbol{\alpha}_2$ at μ_1 and μ_2. In our observed data, patients' start of study time is random after a study start date, and data are collected until a fixed study end date $\exp(C_i)$. We give all simulated patients the same start of study time and generate lognormal survival times $\exp(y_{i1})$ and lognormal end of study times $\exp(C_i)$, where $\exp(y_{i1})$ and $\exp(C_i)$ are independent. We then generate exponentially distributed visit times with mean μ_2 until a visit time is generated that is greater than $\exp(y_{i1})$ or $\exp(C_i)$. The total visits for patient i is y_{i2}, and patients with $y_{i2} = 1$ are removed from the dataset.

Let t_{ij} be the i^{th} simulated patient's j^{th} visit time with final visit time $t_{iy_{i2}}$. Simulated patients can be right censored if $C_i < y_{i1}$, or if $\log(t_{iy_{i2}}) < y_{i1}$, and we set $l_i = \log(t_{iy_{i2}})$ and $r_i = \infty$. Simulated patients with $\log\left(t_{i(y_{i2}-1)}\right) < y_{i1} < \log(t_{iy_{i2}}) < C_i$ are interval censored, and $l_i = \log\left(t_{i(y_{i2}-1)}\right)$ and $r_i = \log(t_{iy_{i2}})$.

Under non-informative censoring,

$$\mathrm{E}\left[y_{i1} | l_i \leq y_{i1} \leq r_i\right] = \mathrm{E}\left[\frac{f(y_{i1})\mathbf{1}\left\{l_i < y_{i1} < r_i\right\}}{P(l_i < y_{i1} < r_i)}\right]. \tag{3.33}$$

We assess the effect of the clinic visit and right censoring distributions on the informativeness of the censoring under three simulation schema with 16 scenarios each. In all schema, we set $\mu_1 = 2.25$ and $\sigma_1^2 = 0.3$, which puts 95% of simulated survival times between 3.2 and 27.7 years.

In schema 1, we fix the clinic visit random effect variance $\sigma_{\delta 2}^2$ at 0.25, and consider right censoring rates of $0, 0.25, 0.5$, and 0.9, and average visits per year μ_2 of $0.5, 1, 2$ and 6. In schema 2, we fix the average number of visits per year μ_2 at 1, let the right censoring rates be the same as in schema 1, and let the random effect variance for visits $\sigma_{\delta 2}^2$ be $0.1, 0.25, 0.5$ and 0.75. In schema 3, we fix the censoring rate at 0.25, and let μ_2 and $\sigma_{\delta 2}^2$ vary, using the same values as in schema 1 and 2 respectively. Table 3.1 describes all 48 simulation scenarios.

For each of the 48 simulation scenarios, we construct $10,000$ simulated datasets with $10,000$ IID log survival times and clinic visit schedules. For dataset $g, g \in 1, \ldots, 10,000$ for a given scenario, we calculate the difference between the patients' known survival times y_{i1} and expected survival times $\int_{l_i}^{r_i} y f_{y_{i1}}(y)/P(l_i < y_{i1} < r_i)dy$ under non-informative censoring. We then calculate the average difference across patients within a dataset as

$$\text{Diff}_g = \frac{1}{10,000} \sum_{i=1}^{10,000} y_{i1} - \int_{l_i}^{r_i} y f_{y_{i1}}(y)/P(l_i < y_{i1} < r_i)dy, \tag{3.34}$$

which gives us in years the difference between the true expected value of y_{i1} and the expected value of y_{i1} assuming non-informative censoring. For each scenario, we get a distribution of averaged differences Diff_g, and if the distribution is not centered at zero, then the censoring is informative. We conduct the simulation for all three schema three times, one for each $\rho_\gamma \in \{-0.5, 0, 0.5\}$.

3.4.1 Simulation Results

The results for interval-censored patients and right-censored patients are presented in Figures 3.3 and 3.4 for $\rho_\gamma = 0$ and σ_2^2 fixed (schema 1). Consistent with non-informative censoring, in every scenario the distribution of the averaged differences (3.34) is centered at zero, showing that the censoring is not informative. The results are similar when we fix the censoring rate (schema 2), and μ_2 (schema 3).

The results for interval-censored patients and right-censored patients are presented in Figures 3.5 and 3.6 for $\rho_\gamma = 0.5$ and $\sigma_{\delta 2}^2$ fixed (schema 1). The dotted lines show the distribution of the averaged differences (3.34), and the solid lines show the distribution of the average differences when we replace the marginal density $f_{y_{i1}}(y)$ in (3.34) with the conditional density $f_{y_{i1}|\lambda_i}(y|\lambda)$. We see bias for the interval-censored subjects or the right censored subjects in every scenario. When the interval censored predictions are biased, the marginal density consistently underestimates mean survival time. For right censored patients, when the right censoring rate is low the marginal density underestimates survival, and as the right censoring rate increases, the bias becomes more and more positive. In all scenarios, the bias decreases with increased rate of clinic visits.

The results for schema 2 (μ_2 fixed) for $\rho_\gamma = 0.5$ are similar (Figures 3.7 and 3.8). For interval censored patients, survival times are consistently underestimated except at very high right censoring rates, and the magnitude of the bias increases with the random effect variance σ_2^2. Estimates of survival for right censored patients are negatively biased at modest right censoring rates, and the bias becomes more positive as the right censoring rate increases.

The results for schema 3 (right censoring rate fixed) are consistent with the results for schemas 1 and 2 (Figures 3.9 and 3.10). The magnitude of the bias increases with μ_2 and with $\sigma_{\delta 2}^2$. For all scenarios and schemas, when $\rho_\gamma = -0.5$ (not shown), the results are similar, but the direction of the bias is reversed. Additionally, when we replace the marginal density $f_{y_{i1}}(y)$ in (3.34) with the conditional densitiy $f_{y_{i1}|\lambda_i}(y|\lambda)$, the bias goes away.

Thus, when ρ_γ is zero, the censoring is not informative irrespective of the censoring and clinic visit distributions. Conversely, the censoring is informative when the censoring is non-zero. In general, the informativeness of the censoring increases with the right censoring rate and clinic visit variance, and decreases with the average number of clinic visits.

3.5 Data Analysis

We ran the model in section 3.1 using a Random Walk Metropolis-Hastings algorithm coded in `C++` using `R` Version 3.3.1 and the `Rcpp` package Version 0.12.7. We ran four chains in parallel to generate $10,000$ samples from the posterior. Convergence within chains was checked using Geweke statistics (Geweke, 1992) and Raftery and Lewis statistics (Raftery and Lewis, 1992). Convergence between chains was checked using Gelman diagnostics (Gelman and Rubin, 1992), and convergence was deemed satisfactory.

3.5.1 Results

The results of the regression analysis are presented in Table 3.2. For HIV, results are consistent with what we would typically expect; African Americans and Hispanics have shorter seroconversion times than Whites, reflecting their increased risk of HIV. Drug use

and having an STI are associated with shorter survival times. For clinic visits, Hispanics have significantly fewer visits per year than whites, and illicit drug users come to the clinic more frequently.

3.5.2 Covariance Matrix Parameters

Table 3.2 shows that the non-spatial variances are much larger than the spatial variances for both HIV survival time and clinic visits. Further, the correlation between the spatial random effects drops off rapidly with distance for both outcomes (not shown in the table). For HIV seroconversion, the correlation decays to 0.05 between 1 and 17 miles, and for clinic visits the correlation decays to 0.05 between 0.7 and 2.5 miles.

Posterior densities for the within-person correlation ρ, within-ZCTA random effects correlation ρ_b, and total random effects correlation ρ_γ are plotted in Figure 3.11. The posterior 95% interval for ρ_b is $(-0.25, 0.24)$, and the log Bayes Factor testing $\rho_b = 0$ versus the alternative that $\rho_b \neq 0$ is 0.98, suggesting that the spatial correlation between the survival time and clinic random effects is negligible. However, the posterior 95% interval for ρ is $(-0.3, -0.1)$, and the log Bayes Factor comparing $\rho = 0$ versus the alternative that $\rho \neq 0$ is -21. This is strong evidence in favor of survival time being negatively associated with frequency of clinic visits suggesting that people who come in more frequently tend to have shorter seroconversion times. Similarly, for ρ_γ the posterior 95% interval is $(-0.24, -0.11)$, and the log Bayes Factor comparing $\rho_\gamma = 0$ versus the alternative that $\rho_\gamma \neq 0$ is -13.

We plot the marginal correlation between $\exp(y_{i1})$ and y_{i2} as a function of the linear predictor $x_i^T \boldsymbol{\alpha}_2$ along with the limiting correlation (3.30) in Figure 3.12. The Lognormal variance γ_1^2 is quite large and as discussed in section 3.2, despite the strong posterior evidence that $\rho_\gamma < 0$, the correlation between y_{i1} and y_{i2} unconditional on the random effects is small.

3.6 Discussion

Our data analysis shows that patients who come to the clinic more frequently have lower survival times on average. A plausible mechanism for the negative correlation is that patients may tend to come into the clinic quickly after engaging in behaviors that put them at risk for HIV, suggesting that patients are accurately assessing their risk of HIV. The negative correlation also implies that if we ran the model without clinic visits that the censoring would be informative, which could potentially bias inferences and estimates of seroconversion times.

Modeling $\log(y_{i1})$ and λ_i as bivariate Normal gives the Poisson process model a number of useful properties. First, the properties of the Multivariate Normal distribution have been well studied (Crow and Shimizu, 1987; Aitchison and Ho, 1989; Mostafa and Mahmoud, 1964), allowing us to analytically calculate the likelihood (3.9) which facilitates sampling from the joint posterior of y_{i1} and δ_{i2}. Another benefit of the bivariate Normal specification for y_{i1} and λ_i is that it simplifies the calculations in sections 3.2 and 3.3.

Other researchers have used shared frailty parameters to model dependency between the visit process and the survival times (Cai et al., 2012; Liang et al., 2009; Liu et al., 2008; Sun et al., 2007, 2012; Zhang et al., 2007), and we have tried to expand upon the previous work by considering more detailed simulation scenarios. We have also presented the results in terms of the bias in our estimates of the survival times as opposed to individual parameter estimates, as this is usually more of interest to clinicians, and found that even in the most extreme scenarios, the bias is not so large that ignoring the informativeness of the censoring would render results useless.

One limitation of our work is that for rare events, corresponding survival times are long, and the survival time variance can be large, so even given strong posterior evidence for the sign of ρ_γ, the correlation between survival and clinic visits can be small. Another limitation is that in the simulation studies, we are making two potentially strong assumptions. First, we assume a homogeneous visit process where the visits are iid exponential, which has the well-known and probably unrealistic property of being memoryless. Further, we are assuming that survival is only correlated with the expected value of the visit process, and not with

the visits themselves. It may in fact be the case that patients visit the clinic more (or less) frequently when they are engaging in more (or less) risky behavior. If we believe this to be the case, modeling a visit process where visit times are independent, but not necessarily identically distributed, may make the model a more accurate representation of the data generating process.

Tables and Figures

Schema 1 ($\sigma_2^2 = 0.5$)		Schema 1 ($\mu_2 = 1$)		Schema 3 (RCR = 0.25)	
RCR	μ_2	RCR	σ_2^2	μ_2	σ_2^2
0	0.5	0	0.10	0.5	0.10
0	1.0	0	0.25	0.5	0.25
0	2.0	0	0.50	0.5	0.50
0	6.0	0	0.75	0.5	0.75
0.25	0.5	0.25	0.10	1.0	0.10
0.25	1.0	0.25	0.25	1.0	0.25
0.25	2.0	0.25	0.50	1.0	0.50
0.25	6.0	0.25	0.75	1.0	0.75
0.50	0.5	0.50	0.10	2.0	0.10
0.50	1.0	0.50	0.25	2.0	0.25
0.50	2.0	0.50	0.50	2.0	0.50
0.50	6.0	0.50	0.75	2.0	0.75
0.90	0.5	0.90	0.10	6.0	0.10
0.90	1.0	0.90	0.25	6.0	0.25
0.90	2.0	0.90	0.50	6.0	0.50
0.90	6.0	0.90	0.75	6.0	0.75

Table 3.1: Parameter values for simulation study in Section 3.4. RCR is Right Censoring Rate. For all scenarios and schemas, we use a lognormal distribution with mean parameter 2.25 and variance parameter 0.3 for y_{i1}.

	Mean (95% CI)	
Parameter	HIV Survival	Clinic Visits
Race		
White	REF	REF
African American	-0.59 (-1.00, -0.18)	0.03 (-0.03, 0.09)
Hispanic	-0.50 (-0.76, -0.25)	-0.10 (-0.13, -0.06)
Other	0.17 (-0.23, 0.58)	-0.05 (-0.10, 0.001)
STI	-0.75 (-1.04, -0.47)	0.03 (-0.01, 0.06)
Drug	-0.80 (-1.04, -0.56)	0.05 (0.02, 0.08)
Spatial variance	1.02 (0.08, 2.52)	0.05 (0.04, 0.07)
Residual Variance	6.31 (5.55, 7.17)	0.33 (0.31, 0.34)

Table 3.2: Posterior summaries for regression parameters. Spatial variance parameters are η_1^2 and η_2^2. Non-spatial residual variance parameters are σ_1^2 and σ_2^2.

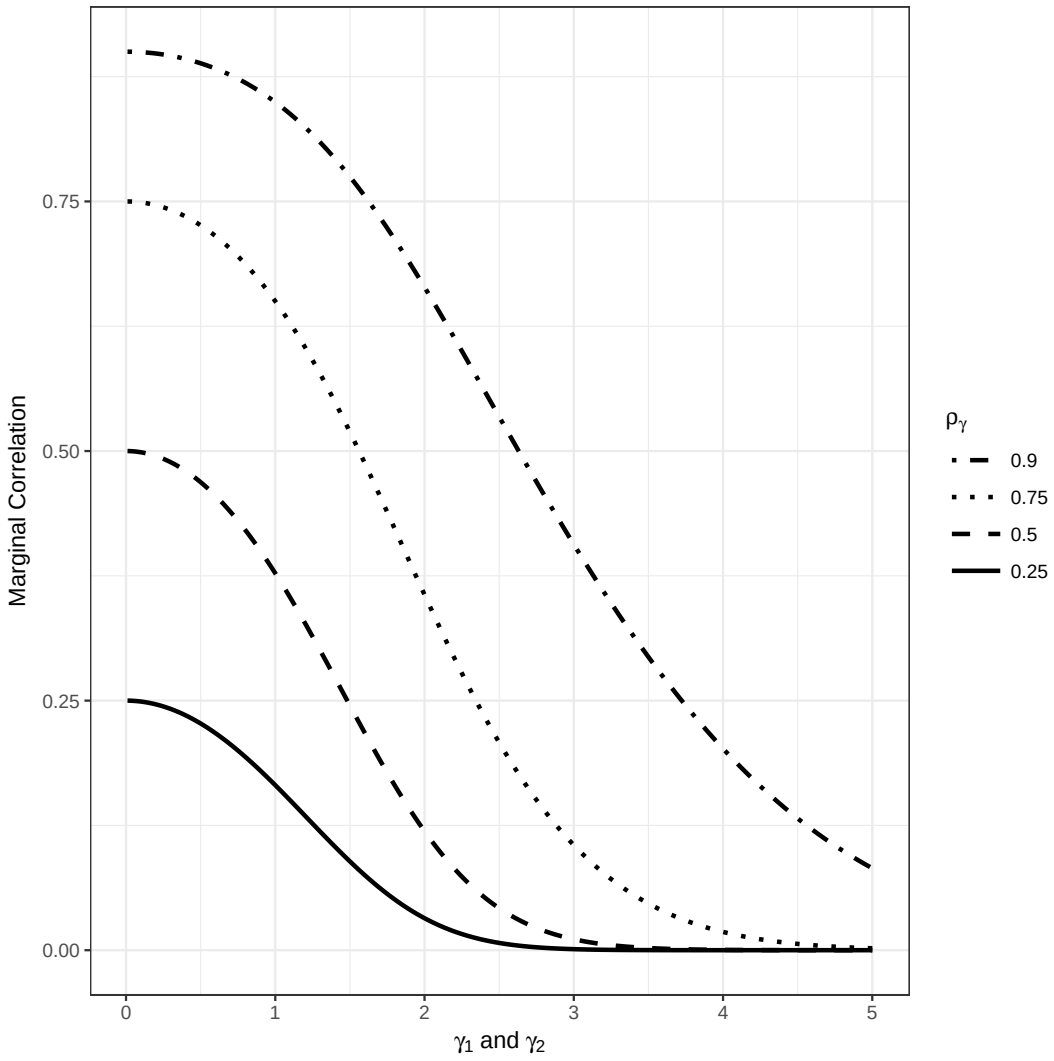

Figure 3.1: Plots of limiting Lognormal correlation for Normal correlations of 0.25 (Solid), 0.50 (Dashed), 0.75 (Dotted), and 0.90 (Dot-dash). Standard deviations γ_1 and γ_2 are equal.

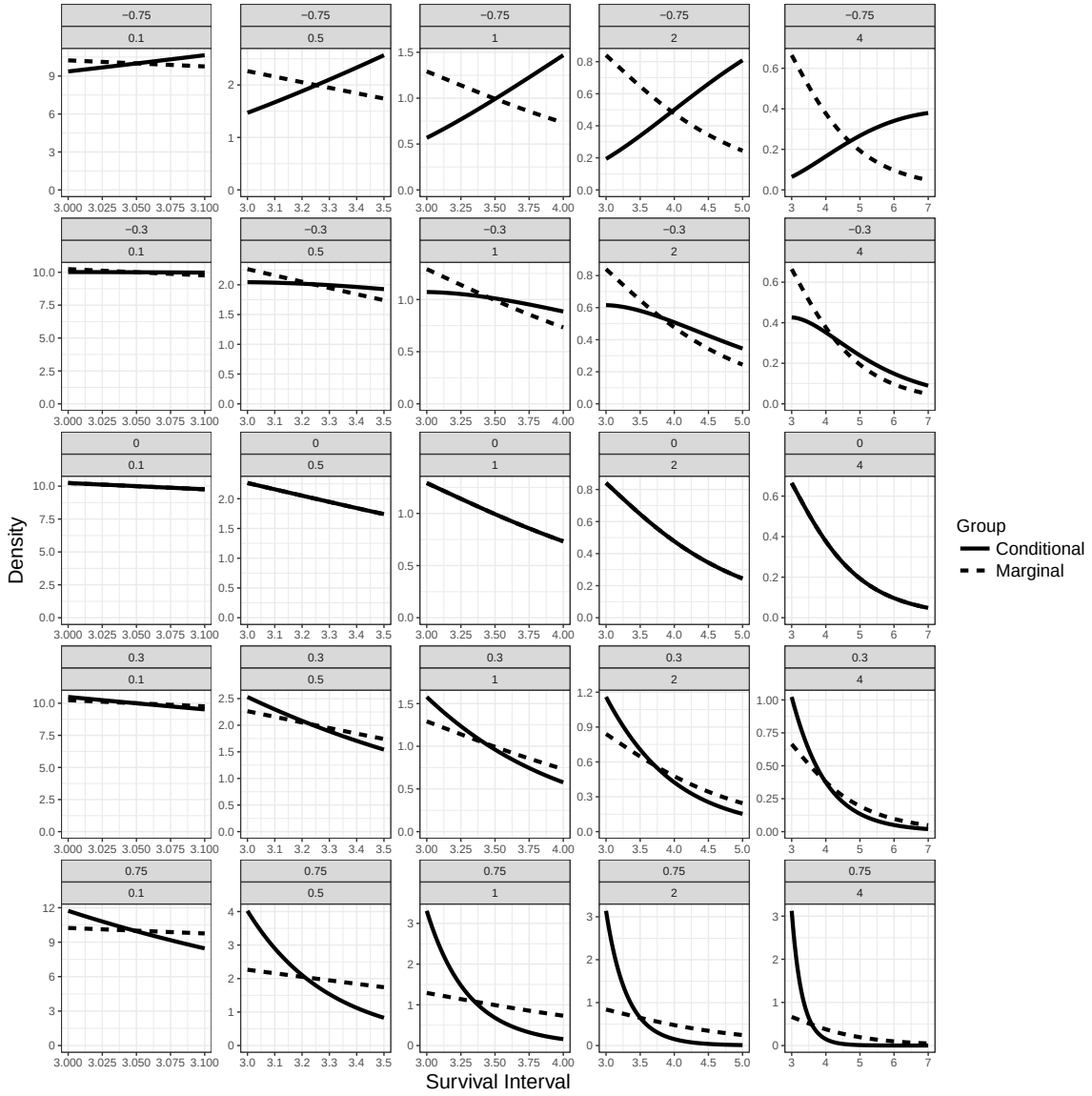

Figure 3.2: Differences between truncated marginal density for y_{i1} (dotted), and truncated conditional density of $y_{i1}|\lambda_i$ after integrating out δ_{i2} (solid). Top label for each plot indicates the correlation ρ_γ between y_{i1} and δ_{i2}, which ranges from -0.75 to 0.75. Bottom label indicates width of the censoring interval, which increases from 0.1 to 4. In all plots, $y_{i1} \sim \mathrm{N}\left(1, 0.25^2\right)$, and $\delta_{i2} \sim \mathrm{N}\left(2, 0.25^2\right)$.

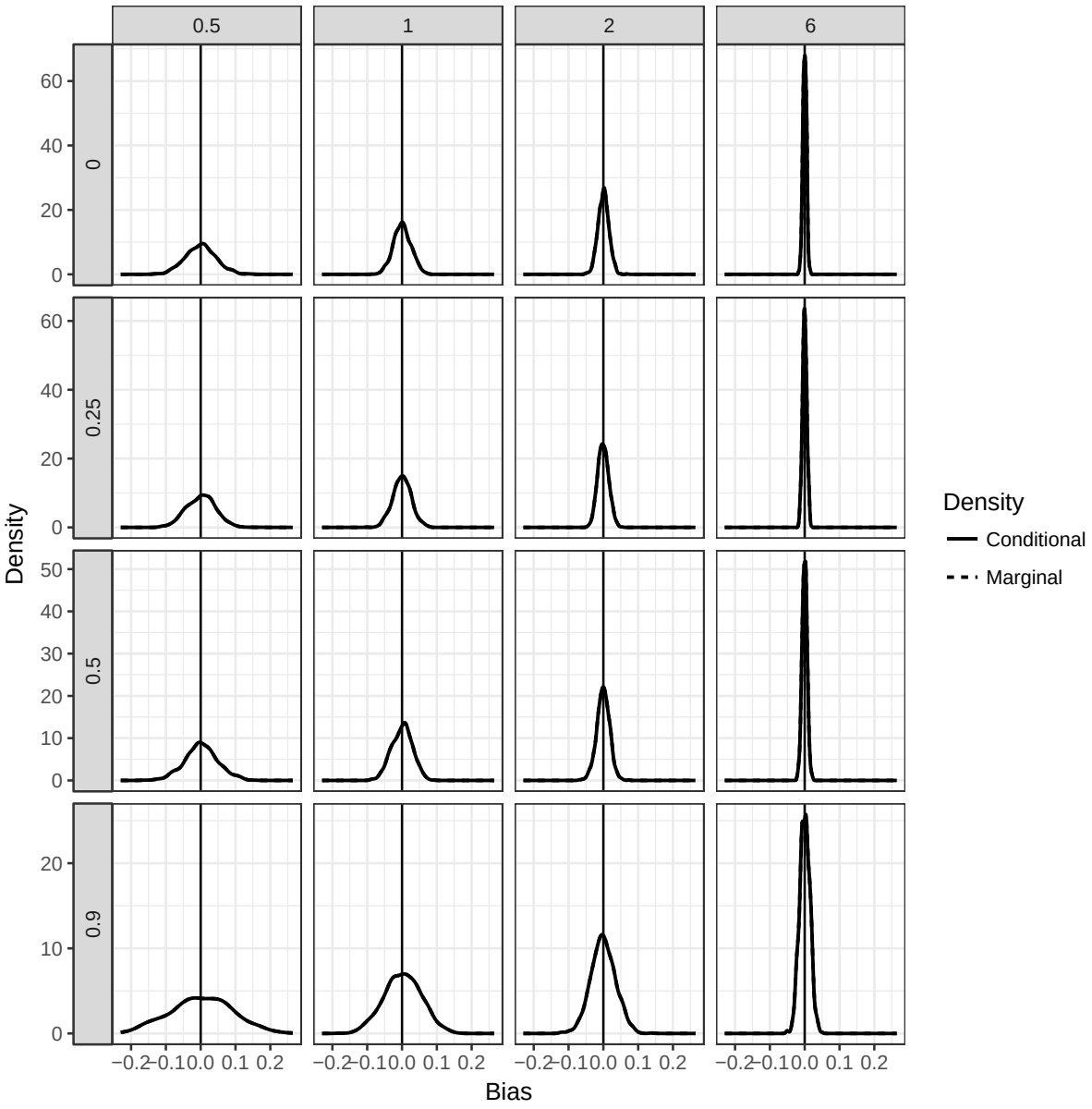

Figure 3.3: Simulation results estimating bias of $\mathrm{E}\left[y_{i1}|l_i \leq y_{i1} \leq r_i\right]$ for predicting survival for interval-censored patients when $\rho_\gamma = 0$; vertical line at zero. When $\rho_\gamma = 0$, $\mathrm{E}\left[y_{i1}|l_i \leq y_{i1} \leq r_i\right]$ is an unbiased estimate of survival for interval-censored patients.

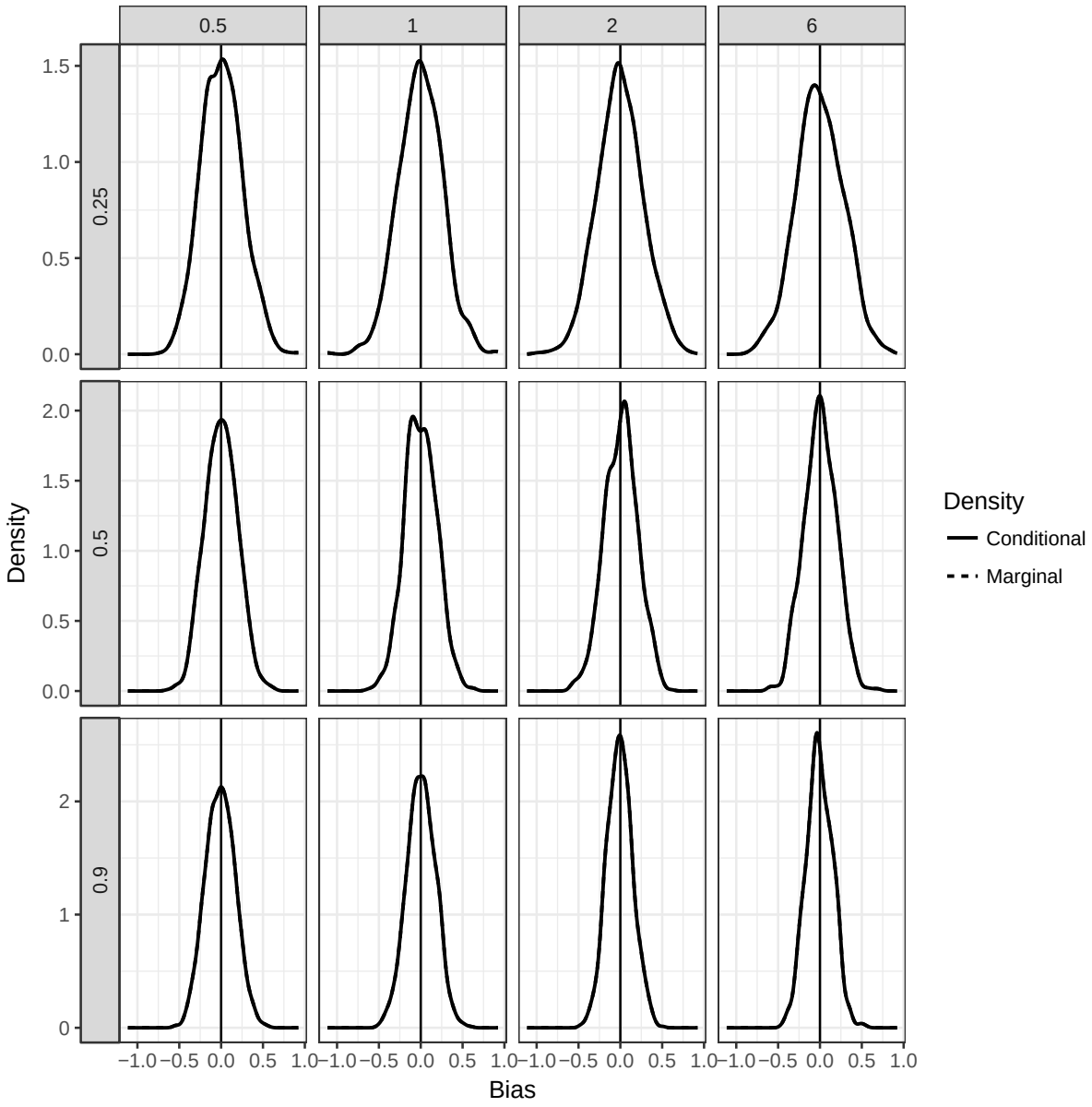

Figure 3.4: Simulation results estimating bias of $\mathrm{E}\left[y_{i1}|l_i \leq y_{i1} \leq r_i\right]$ for predicting survival for right-censored patients when $\rho_\gamma = 0$; vertical line at zero. When $\rho_\gamma = 0$, $\mathrm{E}\left[y_{i1}|l_i \leq y_{i1} \leq r_i\right]$ is unbiased estimate of survival for right censored patients.

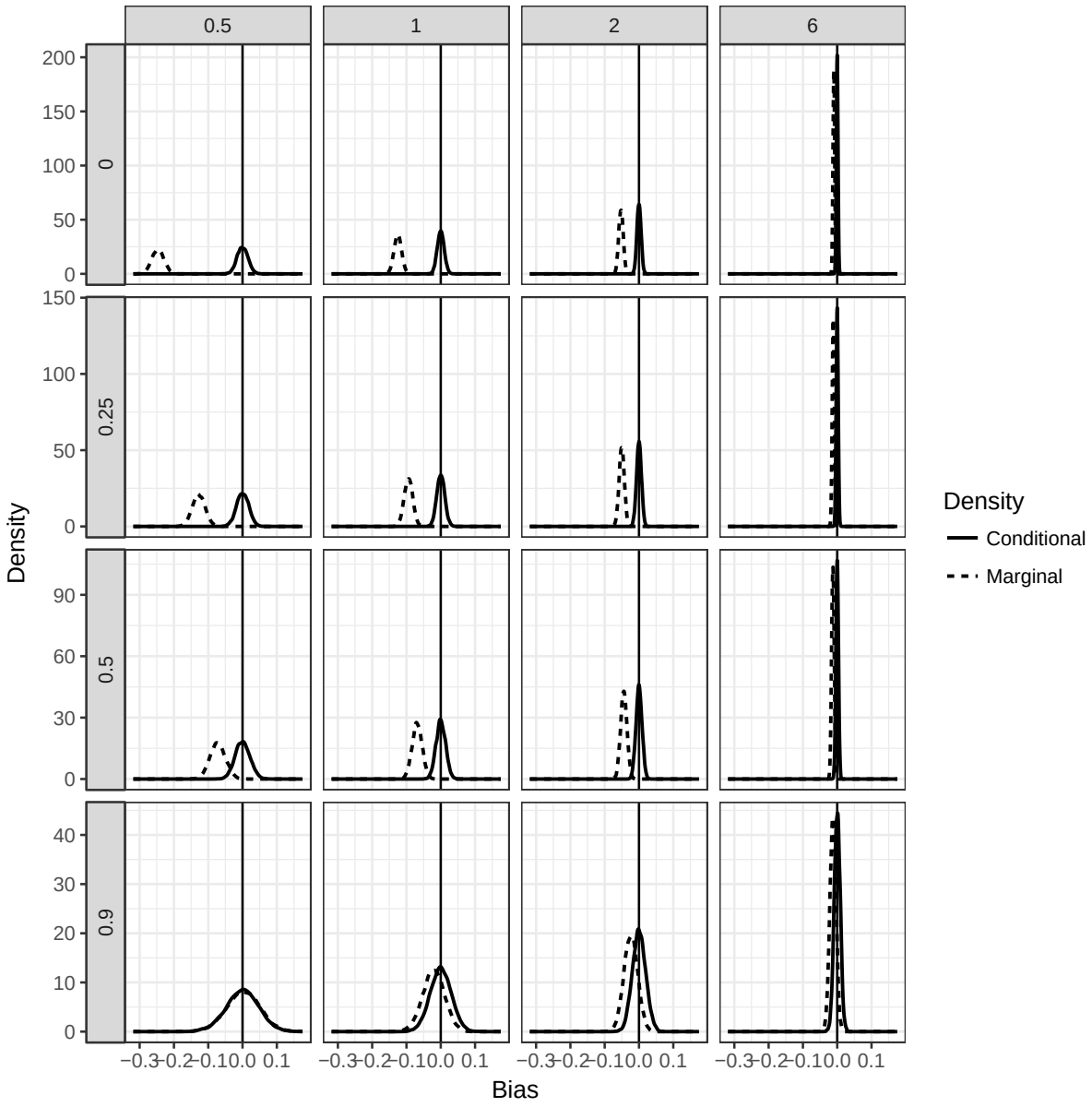

Figure 3.5: Simulation results estimating bias of $\mathrm{E}[y_{i1}|l_i \leq y_{i1} \leq r_i]$ for predicting survival for interval-censored patients when $\rho_\gamma = 0.5$; vertical line at zero. Top label is the average number of visits per year, and the left label is the right censoring rate. Bias decreases with increasing frequency of clinic visits, and is unaffected by right censoring rate

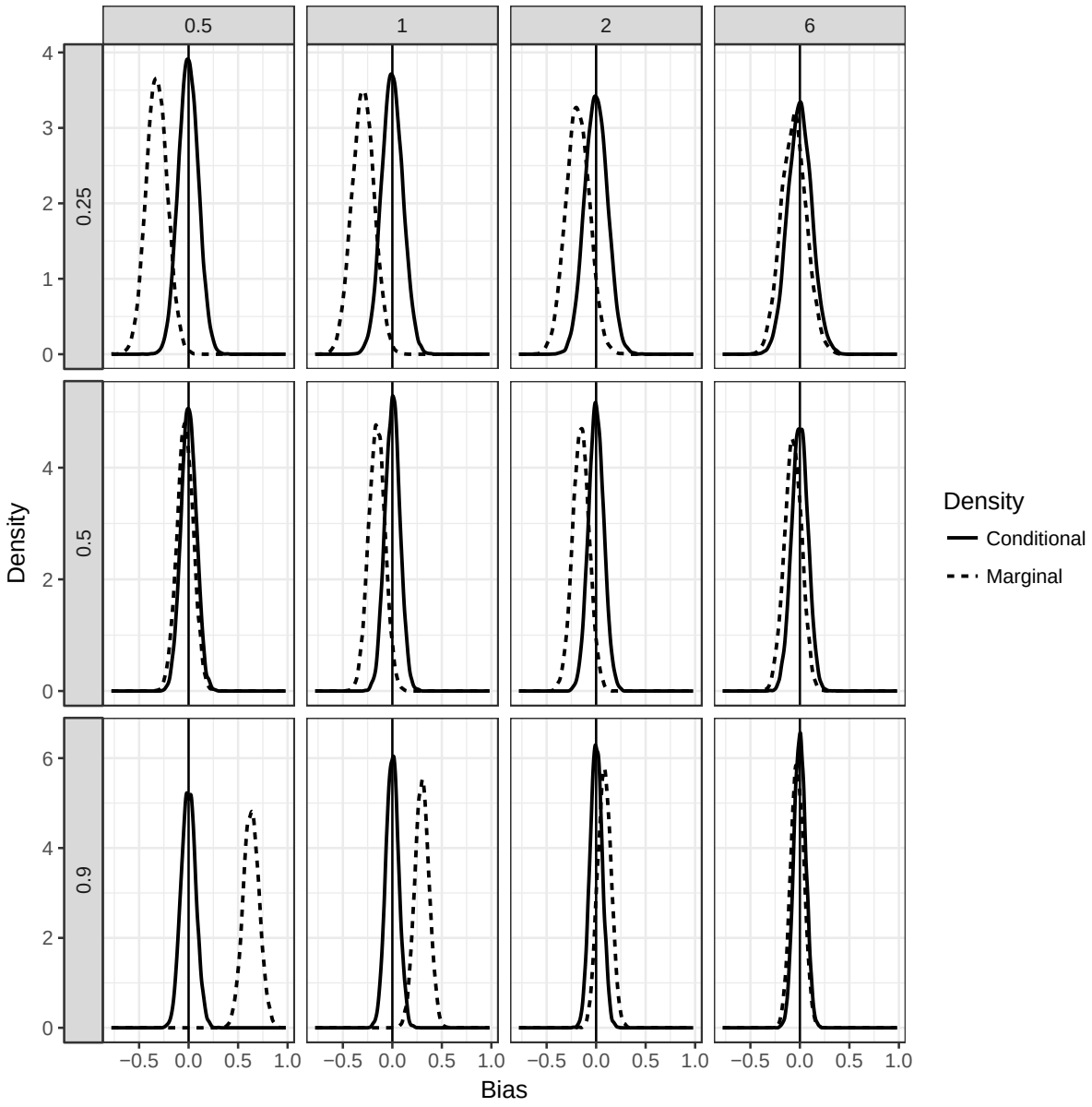

Figure 3.6: Simulation results estimating bias of $\mathrm{E}[y_{i1}|l_i \leq y_{i1} \leq r_i]$ for predicting survival for right-censored patients when $\rho_\gamma = 0.5$; vertical line at zero. Top label is the average number of visits per year, and the left label is the right censoring rate. Bias decreases with increasing frequency of clinic visits, and becomes more positive as right censoring rate increases.

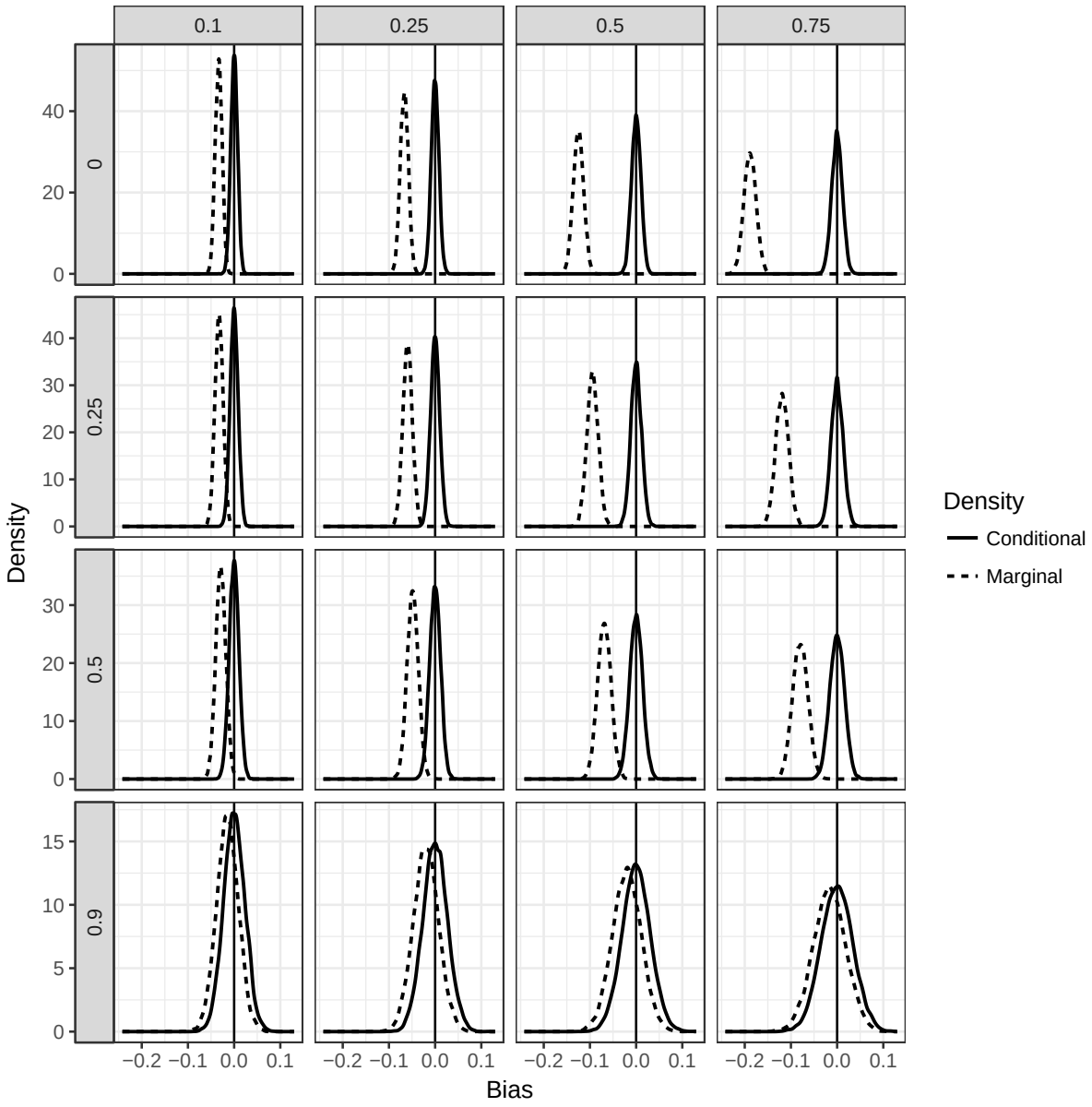

Figure 3.7: Simulation results estimating bias of $\mathrm{E}\left[y_{i1}|l_i \leq y_{i1} \leq r_i\right]$ for predicting survival for interval-censored patients when $\rho_\gamma = 0.5$; vertical line at zero. Top label is the random effect variance $\sigma^2_{\delta 2}$, and the left label is the right censoring rate. Bias increases with $\sigma^2_{\delta 2}$, and is unaffected by right censoring rate

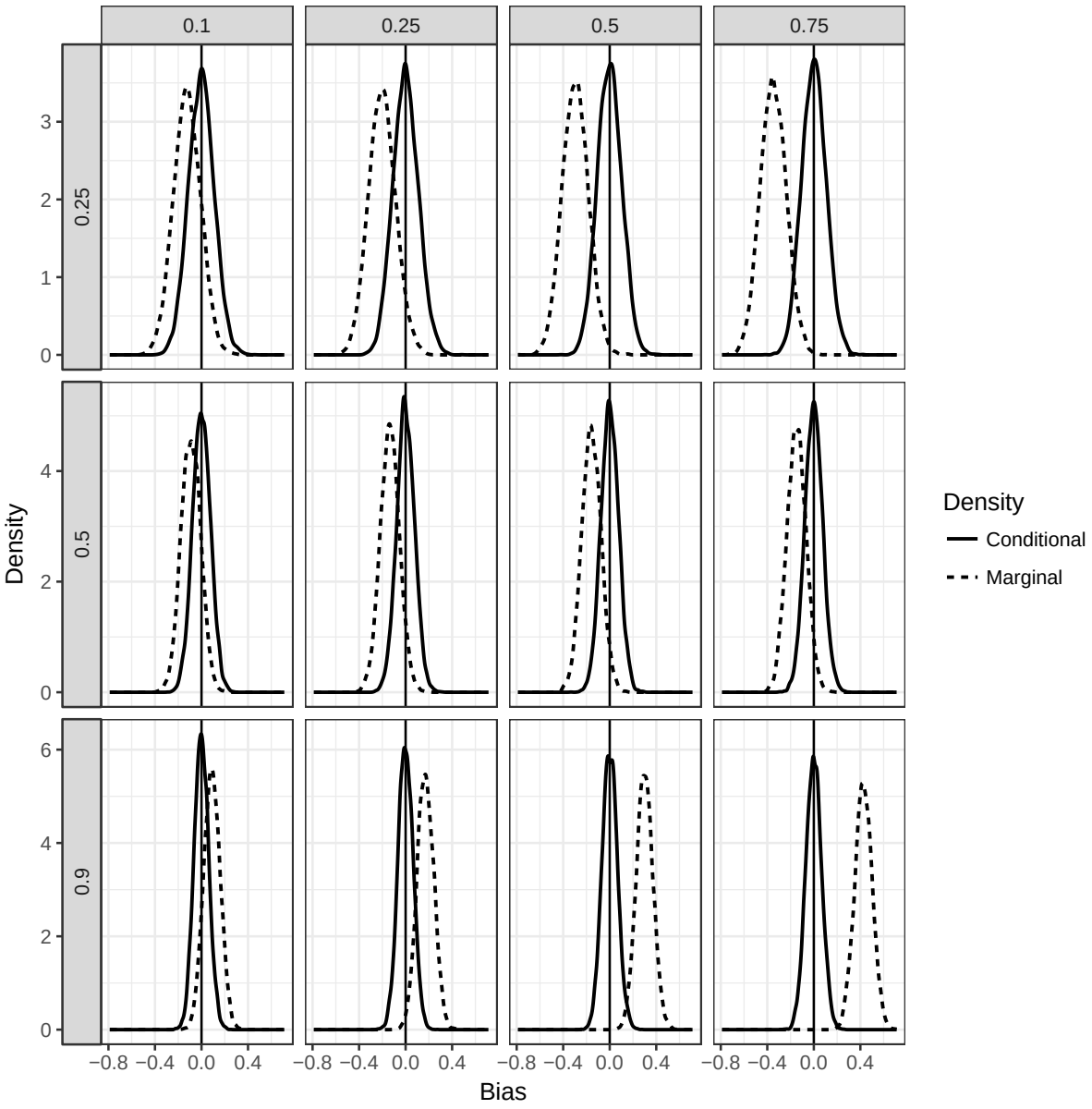

Figure 3.8: Simulation results estimating bias of $\mathrm{E}\left[y_{i1}|l_i \leq y_{i1} \leq r_i\right]$ for predicting survival for right-censored patients when $\rho_\gamma = 0.5$; vertical line at zero. Top label is the random effect variance $\mathrm{Var}(\delta_{i2})$, and the left label is the right censoring rate. Bias increases with increasing $\sigma^2_{\delta 2}$, and becomes more positive as right censoring rate increases.

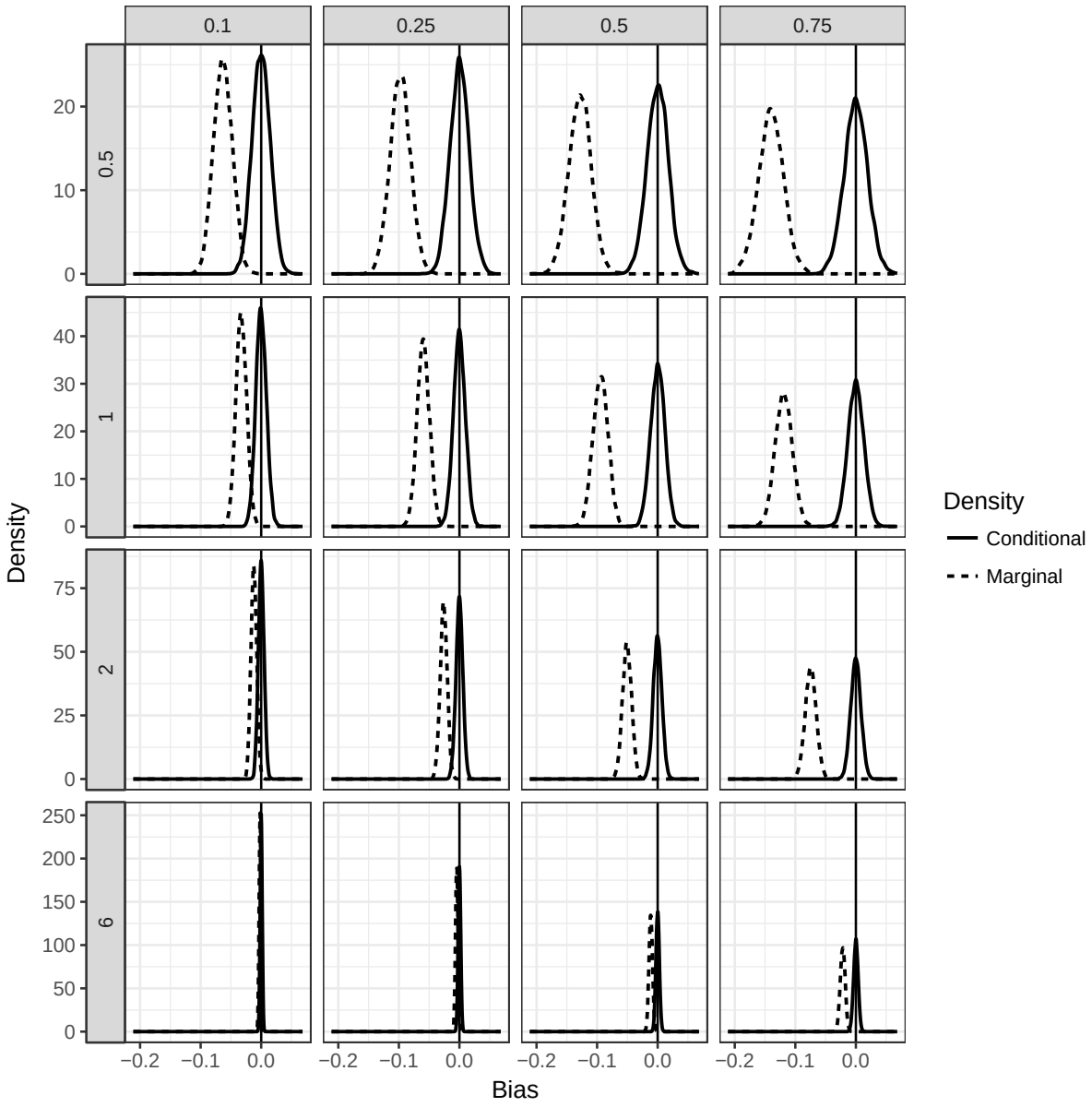

Figure 3.9: Simulation results estimating bias of $\mathrm{E}[y_{i1}|l_i \leq y_{i1} \leq r_i]$ for predicting survival for interval-censored patients when $\rho_\gamma = 0.5$; vertical line at zero. Top label is the random effect variance $\sigma^2_{\delta 2}$, and the left label is the average visits per year. Bias increases with $\sigma^2_{\delta 2}$, and decreases with increasing frequency of clinic visits.

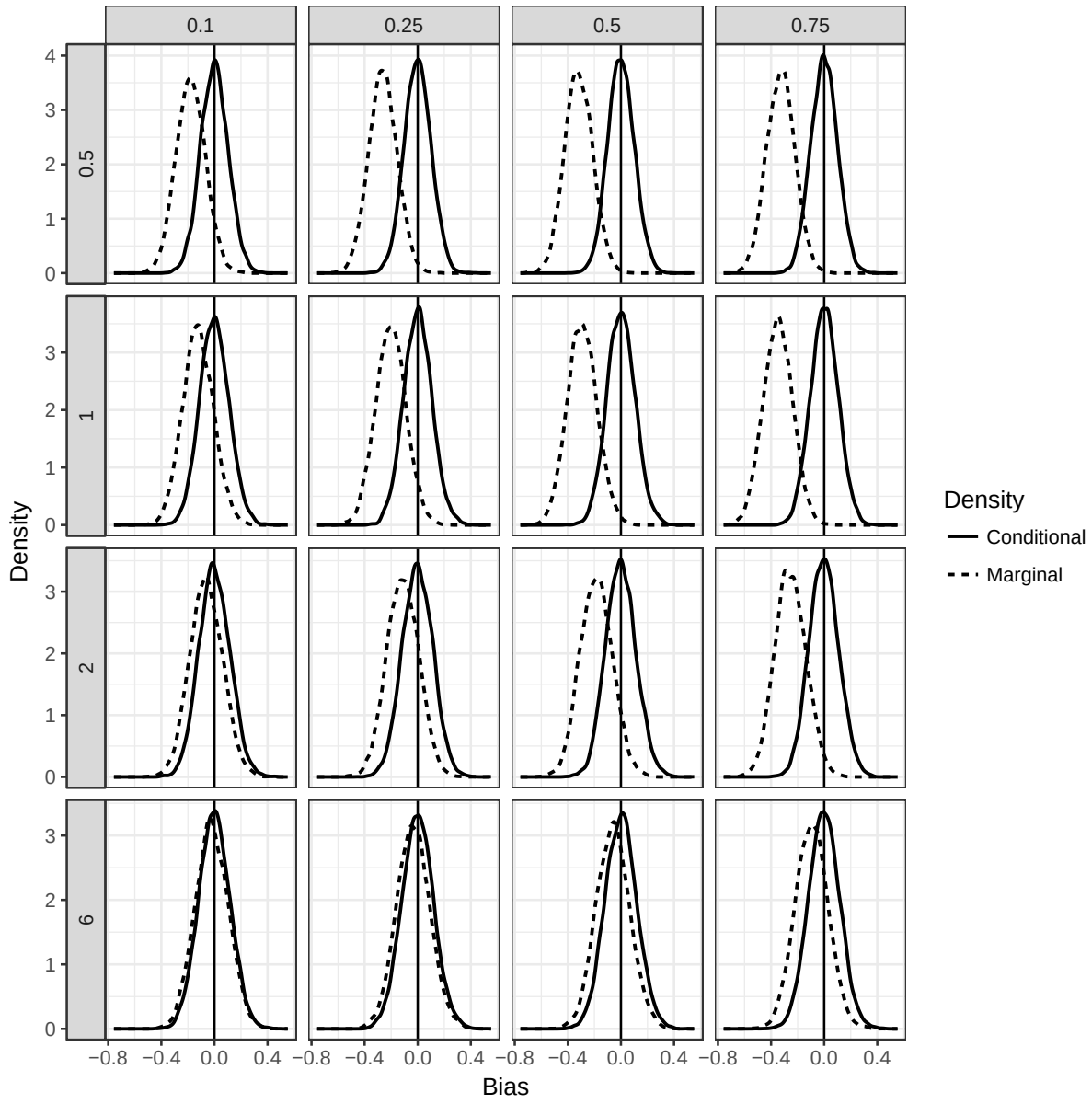

Figure 3.10: Simulation results estimating bias of $\mathrm{E}\left[y_{i1}|l_i \leq y_{i1} \leq r_i\right]$ for predicting survival for right-censored patients when $\rho_\gamma = 0.5$; vertical line at zero. Top label is the random effect variance $\mathrm{Var}(\delta_{i2})$, and the left label is the average visits per year. Bias increases with $\sigma^2_{\delta 2}$, and decreases with increasing frequency of clinic visits.

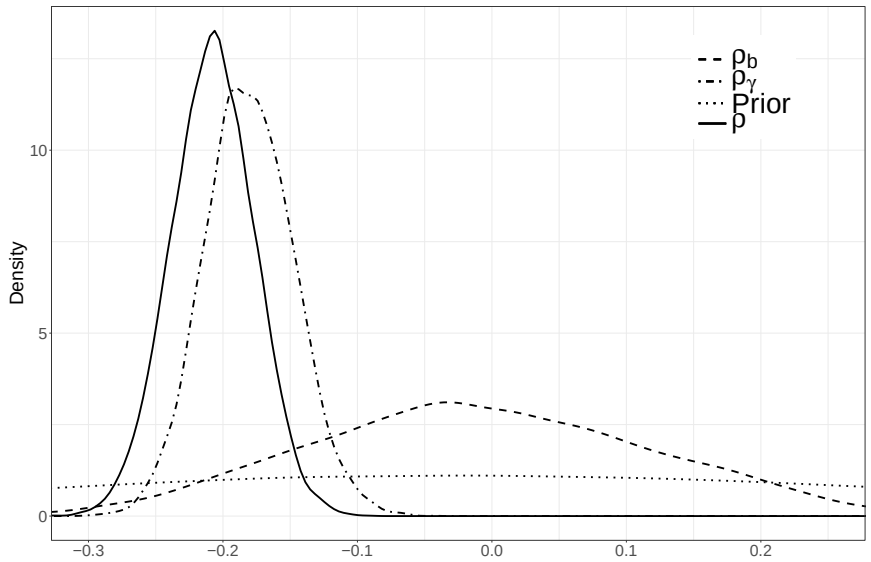

Figure 3.11: Prior distribution for within-person and within-ZCTA code random effect correlation parameters ρ_δ and ρ_b (Dotted). Posterior densities for ρ_δ (Solid), ρ_b (Dashed), and total random effect correlation ρ_γ (Dot-dash).

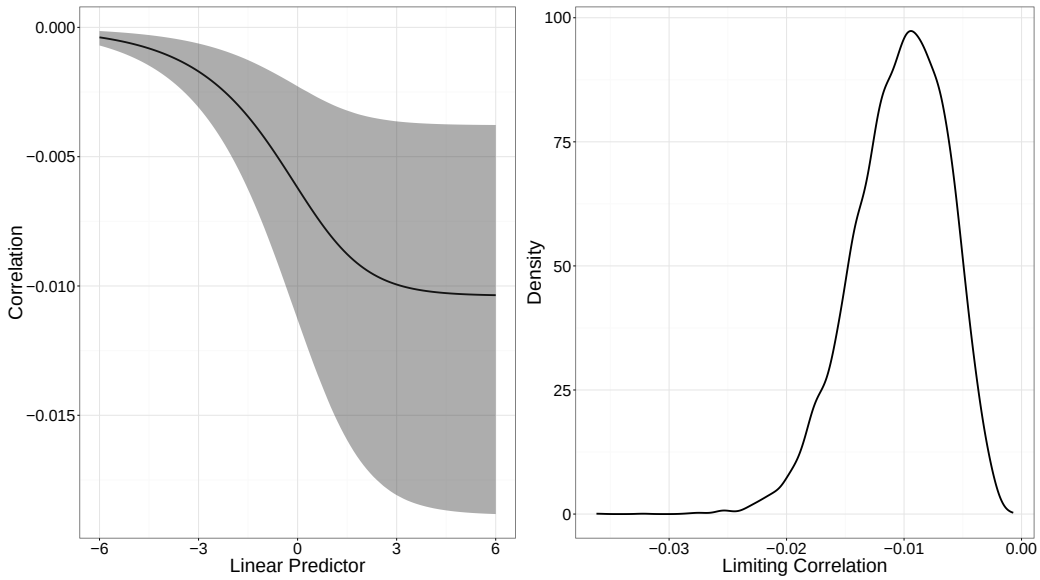

Figure 3.12: Marginal correlation between y_{i1} and y_{i2} as a function of linear predictor $x_i^T \boldsymbol{\alpha}_2$ (left), and limit of the correlation as linear predictor approaches infinity (right).

Appendix: Approximating the Zero Truncated Poisson Distribution

Suppose we have count data $Y = 1, 2, \ldots$ which we model as zero-truncated Poisson,

$$f(Y|\lambda) = \frac{\lambda^Y}{(\exp(\lambda) - 1)Y!} \tag{3.35}$$

so that

$$\mathrm{E}[Y|\lambda] = \frac{\exp(\lambda)}{\exp(\lambda) - 1}\lambda \tag{3.36}$$

$$= c\lambda, \tag{3.37}$$

where $c = \exp(\lambda)(\exp(\lambda) - 1)^{-1}$. Let $X = Y - 1$. Then

$$f(X|\lambda) = \frac{\lambda^{X+1}}{(\exp(\lambda) - 1)(X + 1)!}, \tag{3.38}$$

and

$$\mathrm{E}[X|\lambda] = \mathrm{E}[Y|\lambda] - 1 = c\lambda - 1. \tag{3.39}$$

Suppose conditional on λ, we approximate X as a Poisson random variable with mean $c\lambda - 1$, so that

$$g(X|\lambda) = \frac{(c\lambda - 1)^X}{\exp(c\lambda - 1)\, X!}.$$

To assess how well g approximates f, we calculate the Kullback–Leibler divergence from g to f,

$$D_{KL}(f||g) = \sum_{k=0}^{\infty} f(k|\lambda) \times \log\left(\frac{f(k|\lambda)}{g(k|\lambda)}\right)$$

for values of λ ranging from 0.001 up to 20. The results are plotted in Figure 3.13. For very small values of λ, the divergence is approximately zero. The divergence increases to a maximum of 0.015 when $\lambda = 3.4$, and then decreases monotonically after.

A table of the density values for $f(X|\lambda = 3.4)$ and $g(X|\lambda = 3.4)$ are presented in Table 3.3 for X ranging from 0 to 8. The divergence between the two densities is always less than

about 0.03, and is generally much smaller. Therefore, even at the maximum divergence, the Poisson approximation to the zero-truncated Poisson is still decent. Therefore a zero-truncated Poisson distribution for Y can be approximated by using a Poisson distribution for $Y - 1$.

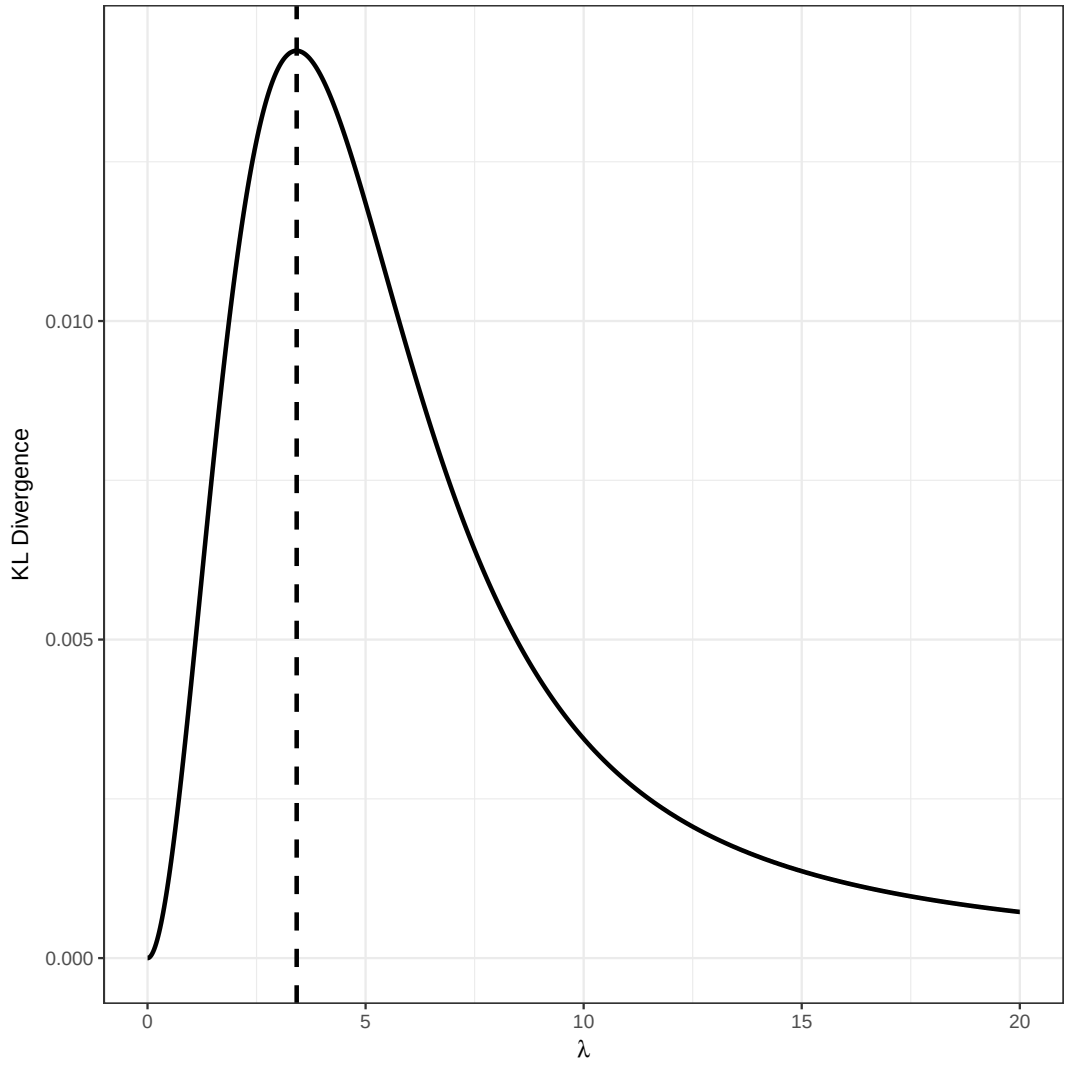

Figure 3.13: Kullback–Leibler divergence from g to f. Vertical dotted line at maximum divergence.

X	ZTP	Poisson
0	0.117	0.081
1	0.200	0.203
2	0.226	0.256
3	0.192	0.214
4	0.131	0.135
5	0.074	0.068
6	0.036	0.029
7	0.015	0.010
8	0.006	0.003

Table 3.3: Densities for $f(X|\lambda = 3.4)$ and $g(x|\lambda = 3.4)$ for $X \in \{0, \ldots, 8\}$.

CHAPTER 4

Developing a risk profile for HIV seroconversion using a spatio-temporal factor analysis

In behavioral studies of HIV, researchers often collect large amounts of data on each patient with the goal of identifying which measurements are associated with a patient's risk of contracting HIV. It is often assumed that a low dimensional set of latent factors explains most of the correlation among these measurements. When the measurements are discrete, or a mix of discrete and continuous data, it is common to model them using a latent multivariate normal distribution (Ansari and Jedidi, 2000; Conti et al., 2014; Hu et al., 2004; Quinn, 2004). Factors within and between records are usually assumed to be uncorrelated, however this is not a reasonable assumption when the data are correlated over space and/or time.

One method to introduce spatio-temporal correlation among factors is through the use of separable models, which are models where the spatial and temporal processes can be decomposed into a sum or product of spatial and temporal correlations. Among multiplicative separable models, unique factor loadings are typically assigned to each spatial unit and the loadings are correlated over space, while the factors evolve over time within a subject (Luttinen and Ilin, 2009; Schmidt and Laurberg, 2008; Schmidt, 2009). These models are similar to spatial dynamic factor models, which use an autoregressive process to model the evolution of factors over time, and then model the spatial correlation in the factor loadings (Lopes et al., 2008, 2011; Strickland et al., 2011; Thorson et al., 2016).

Despite their flexibility, multiplicative separable models and spatial dynamic factor models are not stationary, making them difficult to interpret. In contrast, additive separable models treat the factors as a sum of independent spatial and temporal processes (Cramb

et al., 2015; Richardson et al., 2006; Schliep et al., 2018). As long as the spatial and temporal processes themselves are stationary, the total process will also be stationary. The factor variances can be fixed by treating them as a convex combination of the spatial and temporal factor variances(Abellan et al., 2008; Cheng et al., 2018; Richardson and Green, 1997).

We analyze repeated clinic visit data from the Los Angeles LGBT Center collected between 2008 and 2014. When patients visit the clinic they are given an 82-item risk assessment questionnaire that asks for basic demographic information and risk behaviors such as drug use and history of sexually transmitted infections for the patients and their last two sexual partners.

One of our goals is to reduce the outcomes at one visit to a set of latent factors that identify features of a patient's propensity for risky behavior which in turn are used to predict their HIV status at the next visit. We model correlation among the outcomes through latent factors and specify a lower triangular structure for the loadings matrix to fix rotational identifiability problems (Dunson, 2007; Geweke and Zhou, 1996). We know when patients come to the clinic and the ZIP code where they live at each visit. We assume that factors are uncorrelated within visits and ZIP Codes and model spatio-temporal correlation among the factors using a weighted additive separable model for the spatial and temporal processes. We treat the factor scores from the current visit as covariates in a probit model to predict a patient's probability of being HIV positive at the next follow-up visit and show that this is equivalent to a factor model that allows the factors to load onto different outcomes at different times.

The next section presents the factor model and how we use the factors to predict a patient's risk of becoming HIV positive. Section 4.2 describes the spatio-temporal correlation in the factor model. Section 4.3 presents the prior distributions. Section 4.4 presents the results of the analysis and section 4.5 concludes the paper with a brief discussion.

4.1 Data Structure and Model

We only consider men who have sex with men (MSM) who reside in Los Angeles County with at least two clinic visits. The first visit establishes seronegativity at baseline, and later visits determine whether or not they have become HIV positive since the previous visit. It can take at least three months for a patient to test positive after becoming infected, so to reduce the probability of including patients in our sample who actually entered the study HIV positive, patients need to have visits spanning at least six months. The final data set we use has $7,890$ patients and $29,737$ records.

4.1.1 Factor Model

The data contains $K = 25$ risk-associated outcomes that are a mix of binary and continuous variables. Let $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijK})$ be a $K \times 1$ vector of risk outcomes and x_i be an $(M+1) \times 1$ covariate vector with first element equal to one, where $i = 1, \ldots, N$ indexes patients, $j = 1, \ldots, n_i$ indexes visits where n_i as the number of visits for patient i, and $k = 1, \ldots, K$ indexes outcomes. Define t_{ij} as the time from baseline of person i's j^{th} visit, and $s(i,j)$ as the ZCTA where person i lives at visit j, where $s(i,j) \in \{1, \ldots, S\}$. We want to reduce $\boldsymbol{y}_{ij}$ to a lower dimensional set of $P \ll K$ unobserved factors $\boldsymbol{\xi}_{ij} = (\xi_{ij1}, \ldots, \xi_{ijP})^T$. We model the $\boldsymbol{y}_{ij}$ as functions of latent normal random variables $\boldsymbol{y}^*_{ij} = (y^*_{ij1}, \ldots, y^*_{ijK})$

$$y_{ijk} = \begin{cases} \boldsymbol{I}\left[y^*_{ijk} > 0\right], & \text{if } y_{ijk} \text{ is binary,} \\ y^*_{ijk}, & \text{otherwise.} \end{cases} \tag{4.1}$$

Conditional on latent factors $\boldsymbol{\xi}_{ij}$, outcomes y^*_{ijk} are independent and modeled longitudinally as

$$y^*_{ijk} = x_i^T \boldsymbol{\alpha}_k + \Lambda_k^T \boldsymbol{\xi}_{ij} + \epsilon_{ijk}, \tag{4.2}$$

where $\boldsymbol{\alpha}_k^T = (\alpha_{k0}, \alpha_{k1}, \ldots, \alpha_{kM})$ is a vector of $M+1$ unknown regression coefficients, and Λ_k^T is the k^{th} row of the $K \times P$ loadings matrix Λ and the residuals ϵ_{ijk} are independent and

identically distributed

$$\epsilon_{ijk} \sim \mathrm{N}\left(0, \sigma_k^{*2}\right), \tag{4.3}$$

with variances σ_k^{*2}. For binary outcomes, we fix σ_k^{*2} at 1, but any constant value can be chosen without loss of generality.

For HIV seroconversion, once a patient contracts HIV they can never be HIV negative. Further, patients who are on the study for longer have more time to contract HIV. Let $z_{ij} = 1$ if patient i is HIV positive at visit j, and zero otherwise. For $j = 1$, necessarily $z_{ij} = 0$. For $j \in 2, \ldots, n_i$, we model z_{ij} as a function of a latent random variable z_{ij}^*,

$$z_{ij} = \mathbf{1}\left[z_{ij}^* > 0\right], \tag{4.4}$$

$$z_{ij}^* = x_i^T \boldsymbol{\beta} + \boldsymbol{\gamma}^T \boldsymbol{\xi}_{i(j-1)} + \log\left(\left|t_{ij} - t_{i(j-1)}\right|\right) + \delta_{ij}, \tag{4.5}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_M)^T$ is a set of M regression coefficients and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_P)^T$ is a set of P regression coefficients multiplying the latent factors and $\log\left(\left|t_{i(j+1)} - t_{ij}\right|\right)$ is an offset term. Modeling δ_{ij} as a standard logistic random variable would give a logit model for z_{ij}, allowing us to model the odds of HIV seroconversion per unit time. However, one of the benefits of using a probit model for HIV seroconversion is that we can then do Gibbs sampling. Therefore we approximate a standard logistic distribution for δ_{ij} by matching the moments of a normal distribution to the moments of the standard logistic distribution so that

$$\delta_{ij} \sim \mathrm{N}\left(0, \frac{\pi^2}{3}\right). \tag{4.6}$$

Finally, appending z_{ij}^* to $\boldsymbol{y}_{i(j-1)}^*$ and we can re-write (4.2) and (4.5) as

$$\begin{pmatrix} \boldsymbol{y}_{i(j-1)}^* \\ z_{ij}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{\beta}^T \end{pmatrix} x_i + \begin{pmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{\gamma}^T \end{pmatrix} \boldsymbol{\xi}_{i(j-1)} + \begin{pmatrix} \mathbf{0}_K \\ \log\left(\left|t_{ij} - t_{i(j-1)}\right|\right) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_{i(j-1)} \\ \delta_{ij} \end{pmatrix}, \tag{4.7}$$

for $j \in 2, \ldots, n_i$ where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$, showing that model (4.1) – (4.6) is equivalent to a single larger factor model.

4.1.2 Identification of the Parameters

From (4.7), the variance of $\left(\boldsymbol{y}_{i(j-1)}^{*T}, z_{ij}^{*}\right)^{T}$ is

$$\operatorname{Var}\begin{pmatrix}\boldsymbol{y}_{i(j-1)}^{*}\\ z_{ij}^{*}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Lambda}\\ \boldsymbol{\gamma}^{T}\end{pmatrix}\boldsymbol{\xi}_{i(j-1)}\begin{pmatrix}\boldsymbol{\Lambda}\\ \boldsymbol{\gamma}^{T}\end{pmatrix}^{T} + \begin{pmatrix}\sigma_1^{*2} & 0 & \dots & 0 & 0\\ 0 & \sigma_2^{*2} & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & \sigma_K^{*2} & 0\\ 0 & 0 & \dots & 0 & \frac{\pi^2}{3}\end{pmatrix}, \tag{4.8}$$

$$= \begin{pmatrix}\boldsymbol{\Lambda}\\ \boldsymbol{\gamma}^{T}\end{pmatrix}\boldsymbol{\Omega}\begin{pmatrix}\boldsymbol{\Lambda}\\ \boldsymbol{\gamma}^{T}\end{pmatrix}^{T} + \begin{pmatrix}\boldsymbol{\Sigma}_y & \mathbf{0}_K\\ \mathbf{0}_K^{T} & \frac{\pi^2}{3}\end{pmatrix}, \tag{4.9}$$

where for all j, $\boldsymbol{\Omega} = \operatorname{Var}(\boldsymbol{\xi}_{ij})$ and $\boldsymbol{\Sigma}_y = \operatorname{Var}(\boldsymbol{\epsilon}_{ij})$. Pre-multiplying $\boldsymbol{\xi}_{ij}$ and post multiplying $\left(\boldsymbol{\Lambda}^{T}, \gamma\right)^{T}$ by an orthogonal matrix leaves $\operatorname{Var}\left(\boldsymbol{y}_{ij}^{*}|\boldsymbol{\Lambda}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_y\right)$ unchanged, so (4.9) is unique only up to orthogonal rotations of the factors $\boldsymbol{\xi}_{ij}$ and loadings $(\boldsymbol{\Lambda}^{T}, \gamma)^{T}$.

We fix the scale of the factors by setting $\boldsymbol{\Omega}$ to be the $P \times P$ identity matrix $\boldsymbol{I}_P$. To remove the remaining rotational identifiability problems, at least $P(P-1)/2$ additional constraints need to be placed on the elements of $\boldsymbol{\Lambda}$. We choose a lower triangular specification for $\boldsymbol{\Lambda}$,

$$\boldsymbol{\Lambda} = \begin{pmatrix}\lambda_{11} & 0 & \dots & 0\\ \lambda_{21} & \lambda_{22} & \dots & 0\\ \vdots & \vdots & \vdots & \vdots\\ \lambda_{K1} & \lambda_{K2} & \dots & \lambda_{KP}\end{pmatrix}, \tag{4.10}$$

and restricting the diagonal elements to be positive for $\lambda_{pp} > 0$, $p \in 1, \dots p$.

4.2 Spatio-temporal Factors

Each patient has multiple records indexed by ZCTA. To model spatiotemporal correlation among the latent normal random variables, we decompose the factors $\boldsymbol{\xi}_{ij}$ additively into a weighted sum of independent temporal processes $\boldsymbol{\tau}_{ij}$ and spatial processes $\boldsymbol{\psi}_{s(i,j)}$. Define $\boldsymbol{R} = \operatorname{diag}(\rho_1, \dots, \rho_P)$, where for $p \in 1, \dots, p$, $\rho_p \in (0, 1)$ is a weighting factor that indicates

the relative strength of the temporal factors in comparison to the spatial factors. Then

$$\boldsymbol{\xi}_{ij} = \boldsymbol{R}^{1/2}\boldsymbol{\tau}_{ij} + (\boldsymbol{I}_P - \boldsymbol{R})^{1/2}\,\boldsymbol{\psi}_{s(i,j)}. \tag{4.11}$$

We have no reason to believe a priori that the effect of the spatial and temporal factors on each of the outcomes should be equivalent, so a priori we model

$$\boldsymbol{\tau}_{ij} \sim \mathrm{N}\left(\mathbf{0}_{P\times 1}, \boldsymbol{I}_P\right), \tag{4.12}$$

and

$$\boldsymbol{\psi}_{s(i,j)} \sim \mathrm{N}\left(\mathbf{0}_{P\times 1}, \boldsymbol{I}_P\right). \tag{4.13}$$

This fixes the prior variance of $\boldsymbol{\xi}_{ij}$ while simultaneously allowing inference about the relative importance of the spatial and temporal processes in predicting outcomes. Substituting (4.11) into (4.7) gives

$$\begin{pmatrix} \boldsymbol{y}_{i(j-1)} \\ z_{ij} \end{pmatrix} = \boldsymbol{\mu}_{ij} + \begin{pmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{\gamma}^T \end{pmatrix} \left(\boldsymbol{R}^{1/2}\boldsymbol{\tau}_{i(j-1)} + (\boldsymbol{I}_P - \boldsymbol{R})^{1/2}\,\boldsymbol{\psi}_{s(i,(j-1))} \right) + \begin{pmatrix} \boldsymbol{\epsilon}_{i(j-1)} \\ \delta_{ij} \end{pmatrix}, \tag{4.14}$$

$$= \boldsymbol{\mu}_{ij} + \begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{\Lambda} \\ \boldsymbol{\gamma}^T & \boldsymbol{\gamma}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{R}^{1/2}\boldsymbol{\tau}_{i(j-1)} \\ (\boldsymbol{I}_P - \boldsymbol{R})^{1/2}\,\boldsymbol{\psi}_{s(i,(j-1))} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_{i(j-1)} \\ \delta_{ij} \end{pmatrix}, \tag{4.15}$$

$$= \boldsymbol{\mu}_{ij} + \boldsymbol{\Lambda}_{\text{new}}\boldsymbol{\xi}_{\text{new},i(j-1)} + \boldsymbol{\epsilon}_{\text{new},i(j-1)}, \tag{4.16}$$

where

$$\boldsymbol{\mu}_{ij} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} x_i + \begin{pmatrix} \mathbf{0}_K \\ \log\left(\left|t_{ij} - t_{i(j-1)}\right|\right) \end{pmatrix}, \tag{4.17}$$

are the fixed effects plus offset in (4.14),

$$\boldsymbol{\Lambda}_{\text{new}} = \begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{\Lambda} \\ \boldsymbol{\gamma}^T & \boldsymbol{\gamma}^T \end{pmatrix}, \tag{4.18}$$

is a $(K+1)\times 2P$ matrix of factor loadings,

$$\boldsymbol{\xi}_{\text{new},i(j-1)} = \begin{pmatrix} \boldsymbol{\tau}_{i(j-1)} \\ \boldsymbol{\psi}_{s(i,(j-1))} \end{pmatrix}, \tag{4.19}$$

is a $2P \times 1$ vector of factors, and

$$\boldsymbol{\epsilon}_{\text{new},i(j-1)} = \begin{pmatrix} \boldsymbol{\epsilon}_{i(j-1)} \\ \delta_{ij} \end{pmatrix}, \tag{4.20}$$

is a $(K+1) \times 1$ vector of residuals.

Thus, equation (4.14) is equivalent to a confirmatory factor model with factors $\boldsymbol{\xi}_{\text{new},i(j-1)}$ with $\text{Var}\left(\boldsymbol{\xi}_{\text{new},i(j-1)}\right) = \boldsymbol{I}_{2P}$, and loadings matrix $\boldsymbol{\Lambda}_{\text{new}}$. Because we have $P(P-1)/2$ constraints on the first P columns of $\boldsymbol{\Lambda}_{\text{new}}$ and the second P columns of $\boldsymbol{\Lambda}_{\text{new}}$ are constrained to be equal to the first P, the parameters are still identified.

4.2.1 Factor Prior Distributions

Let $\boldsymbol{\tau}_i = \left(\boldsymbol{\tau}_{i1}^T, \ldots, \boldsymbol{\tau}_{in_i}^T\right)^T$ be the $(n_i P) \times 1$ collection of factor scores for patient i. Also let $\boldsymbol{H}_{ip}$ be an exponential decay correlation matrix with decay parameter ϕ_p for the p^{th} temporal factor over time, where the (j, j') element of $\boldsymbol{H}_{ip}$ is $\exp\left(\phi_p |t_{ij} - t_{ij'}|\right)$ and the subscript i is needed because each patient has a unique visit schedule. Then $\boldsymbol{\tau}_i$ has a multivariate normal prior,

$$\boldsymbol{\tau}_i \sim \text{N}\left(\mathbf{0}_{n_i P}, \boldsymbol{\Sigma}_{\boldsymbol{\tau} i}\right), \tag{4.21}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\tau} i}$ is a block diagonal matrix,

$$\boldsymbol{\Sigma}_{\boldsymbol{\tau} i} = \begin{pmatrix} \boldsymbol{H}_{i1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_{i2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{H}_{iP} \end{pmatrix}. \tag{4.22}$$

We define the distribution of the spatial factors analogously. Let $\boldsymbol{\psi}_s = (\psi_{s1}, \ldots, \psi_{sP})$ be the P vector of factor scores for ZIP code s, where $s \in 1, \ldots, S$ and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^T, \ldots, \boldsymbol{\psi}_S^T)^T$ be the collection of factors over all ZIP codes. Also for $p \in 1, \ldots, P$, let $\boldsymbol{G}_p$ be an exponential decay correlation matrix with decay parameter ν_p for the p^{th} spatial factor over space, where the (j, j') element of $\boldsymbol{G}_p$ is equal to $\exp\left(\nu_p |s_j - s'_j|\right)$ and $|s_j - s'_j|$ is the distance in miles

between the centroids of ZIP codes j and j' for $j, j' \in 1, \ldots S$. Then

$$\boldsymbol{\psi} \sim \mathrm{N}\left(\mathbf{0}_{SP}, \boldsymbol{\Sigma}_{\boldsymbol{\psi}}\right), \tag{4.23}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\psi}}$ is a block diagonal matrix,

$$\boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \begin{pmatrix} \boldsymbol{G}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{G}_P \end{pmatrix}. \tag{4.24}$$

4.3 Prior Distributions

We choose $\mathrm{N}(0, 10^2)$ priors for intercepts α_{k0}, $k \in 1, \ldots K$, and for β_0. For the factor loadings $\lambda_{pp}, p = 1, \ldots, P$ on the diagonal of $\boldsymbol{\Lambda}$ we specify positive half standard normal priors. For all other regression coefficients and factor loadings we choose standard normal priors. For residual variances σ_k^2 we set Inverse-Gamma(4,3) priors which puts 95% of the prior mass between 0.34 and 2.75, which is relatively uninformative on the probit scale.

For the temporal decay parameters ϕ_p, roughly 95% of patients are in the study for between 0.25 and 5.5 years. Since $\mathrm{Cor}\left(\tau_{ijp}, \tau_{ij'p}\right) = \exp\left(\phi_p |t_{ij} - t_{ij'}|\right)$ for factor p for patient i, we want to choose a prior for ϕ_p that places appreciable mass on $\mathrm{Cor}\left(\tau_{ijp}, \tau_{ij'p}\right) > 0.05$ given a $|t_{ij} - t_{ij'}| > 0.25$. Choosing Gamma(2,1) priors for ϕ_p a priori sets a prior probability of more than 99% that the $\mathrm{Cor}\left(\tau_{ijp}, \tau_{ij'p}\right)$ decays to less than 0.05 after 0.25 years.

Similarly, for the spatial decay parameters ν_p, the minimum and maximum distance between ZCTA centroids in Los Angeles county are 0.5 and 72 miles respectively. Putting Gamma(2, 2) priors on the ν_p means that $\mathrm{Cor}\left(\psi_{s(i,j)p}, \psi_{s(i,j)p'}\right)$ will decay to less than 0.05 between 1 mile and 25 miles with 95% prior probability. Finally, for the spatio-temporal scale parameters ρ_p, we choose Unif$(0, 1)$ priors.

4.4 Results

We ran the model described in section 4.1 with one factor using a combination of Gibbs sampling and Random Walk Metropolis Hastings steps using 8 chains until an effective sample size of at least $2,000$ was reached for all parameters including the temporal factors $\boldsymbol{\tau}_{ij}$ and the spatial factors $\boldsymbol{\psi}_{s(i,j)}$. Convergence between chains was checked using Geweke diagnostics Geweke (1992), and convergence within chains was checked using Raftery and Lewis Diagnostics (Raftery and Lewis, 1992) and Gelman diagnostics (Gelman and Rubin, 1992). No issues were found. Details of the sampler are provided in the appendix.

4.4.1 Regressions for Risk Outcomes

Covariates included in the model are indicators for race and a cubic B-spline for age at initial visit, with knots at the quintiles. Posterior summaries of the regression coefficients $\alpha_{mk}, m \in \{0, \ldots, M\}$, and $k \in \{1, \ldots, 25\}$ are provided in the web appendix. Generally, a patient's age at baseline is inversely related to the age of their partners, so younger patients tend to have slightly older partners, and older patients tend to have younger partners. Younger patients also tended to have more partners than older patients.

In comparison to Whites, African Americans and Hispanics had fewer partners and slightly older partners, and were much more likely to have partners of the same race. African Americans had lower risk of all STIs aside from Syphilis, and lower rates of all hard drug use among the drugs we considered. African Americans were also more likely to either have HIV positive partners or partners of unknown status. Hispanics had higher rates of STIs and lower rates of use of most drugs aside from meth.

4.4.2 Results for One Factor Model

Posterior summaries for the factor loadings are presented in Table 4.1. Other than having had partners of unknown HIV status or of a different race, all outcomes load substantially and positively onto the factor except for the indicators, which suggests that the factor corre-

sponds to a patient's overall propensity for risky behavior. Summaries of the factor covariance parameters are presented in Table 4.3. The 95% posterior interval for the weighting parameter ρ is $(0.988, 0.996)$, suggesting that spatial factors do not contribute much to a patient's overall riskiness.

Table 4.3 presents the longitudinal and spatial decay parameters ϕ_t and ϕ_s and their associated effective ranges, defined as the distance (in miles) or time (in years) at which the correlation decays to less than 0.05. The estimated correlation between a person's longitudinal factors that are one year apart is 0.72, and does not decay to 0.05 until about nine years, which is longer than the maximum possible time between visits within the study. The spatial correlation decays relatively rapidly, decreasing to less than 0.05 after about 10 miles. Taken together, these results suggest that a patient's propensity for risky behavior remains relatively constant over time, and is essentially uncorrelated with other patients irrespective of how close they live.

Posterior summaries for the HIV regression coefficients $\boldsymbol{\beta}$ and γ are presented in Table 4.4. African Americans and Hispanics are at increased risk of becoming HIV positive in comparison to whites, which is consistent with the literature. Further, a patient's factor score at one visit is significantly associated with whether or not they become HIV positive by their next visit.

To evaluate the ability of our model to discriminate between seroconverters and non-seroconverters, we take the posterior MCMC samples and for each iteration, calculate all of the patients predicted probabilities of becoming HIV positive at their last visit. This gives us posterior samples of each patient's predicted probability. We then calculate a kernel density estimate of the predicted probabilities across patients for each of the samples. The posterior mean density and 95% pointwise confidence bands for seroconverters' and non-seroconverters' predicted probabilities are presented in Figure 4.1. The two solid lines show the pointwise mean density curve. Comparing the mean curves, the predicted probabilities for non-seroconverters is between $(0.001, 0.055)$ with 95% posterior probability. For non-seroconverters, the predicted probabilities are between $(0.004, 0.087)$ with 95% posterior probability. So while our model predicts higher probabilities for seroconverters than non-

seroconverters on average, there is still a large amount of overlap.

To further assess the performance of our model in discriminating between seroconverters and non-seroconverters, Figure 4.2 plots ROC curves and 95% credible bands for models with and without factors. For the model without factors the area under the ROC curve (AUC) is 0.60 (95% posterior interval $(0.58, 0.61)$). When we include the factors into the model, the AUC increases to 0.73 (95% posterior interval $(0.71, 0.75)$), which shows that the factors substantially improve the performance of our model.

To calculate the positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity, we calculate Youden's Index (Youden, 1950) to classify patients as predicted seroconverters and non-seroconverters. The posterior mean (95% CI) PPV, NPV, sensitivity, and specificity are 0.073 $(0.066, 0.081)$, 0.982 $(0.979, 0.985)$, 0.704 $(0.636, 0.770)$, and 0.51 $(0.575, 0.70)$, respectively. Despite the relatively low specificity, we are able to capture 75% of the total HIV cases while simultaneously maintaining reasonable assurances that the patients we classified as being non-seroconverters are in fact overwhelmingly likely to be non-seroconverters.

4.4.3 Results for the Two Factor Model

Posterior summaries for the factor loadings and HIV regression coefficients for the two-factor model are presented in Table 4.2. Qualitatively, the results are very similar to the one factor model. The key difference is that the factor loadings for the indicators of having partners of a different race are non-zero in column two of the loadings matrix. However, the HIV regression coefficient for factor two is tightly centered at zero, which suggests that the race difference indicators do not predict HIV. Both of the weighting parameters ρ_1 and ρ_2 are greater than 0.99 with posterior probability 1, which implies that the spatial correlation among the factors is negligible.

4.5 Discussion

We propose a fully Bayesian two-stage factor analysis model for spatio-temporally correlated data. In the first stage, we reduced 25 outcomes that are known or thought to be associated with increased risk of contracting HIV into a latent factor with temporal and spatial correlation. In the second stage, we used the factor to predict a patient's probability of becoming HIV positive by their next visit. We show that this is equivalent to a factor model where the factors can load on to outcomes measured at different visits.

An alternative to a factor analysis model would be to treat the K outcomes as covariates in a regression model. One of the advantages that a factor analysis provides for this data is that it is a parsimonious model of the correlation among outcomes within patients over time and between patients over space. This allows us to predict HIV at follow-up visits. If we were to directly include the outcomes into a model for HIV, we would still be able to assess which risk factors are associated with increased risk of HIV, but we would not have a mechanism for predicting risk at future visits.

There are several ways to extend our model. First, we assume that the spatial and temporal variances are constant. Aguilar and West (2000) proposed to use stochastic volatility modeling to model the loadings as constant over time while modeling the log of the factor variances and error variances using an autoregressive process. Further, if it is believed that the spatial or temporal processes are non-stationary, a Kalman filter model can also be used to estimate the variances at each time point or each spatial location (Zuur et al., 2003). In addition, Banerjee et al. (2014) provide a number of other non-separable spatio-temporal processes that we can use in place of the additive specification used in this paper.

One way in which clinicians can use our results is by simply presenting patients with their model-based estimates of contracting HIV so that they can be more informed. However, an important point is that the model-based estimates of risk are not equivalent to a test for HIV, so clinicians can also use these results in relatively inexpensive behavioral interventions, including more frequent reminders to come to the clinic or counseling for high risk patients.

Tables and Figures

Outcome[1]	Mean (95% CI) Loading
Log(Partners + 1) (Past 30 Days)	0.17 (0.16, 0.18)
Partner 1 Age Difference	2.85 (2.69, 2.99)
Partner 2 Age Difference	2.61 (2.44, 2.77)
Gonorrhea / Chlamydia (Past Year)	0.14 (0.11, 0.16)
Syphilis (Past Year)	0.17 (0.15, 0.22)
Herpes (Past Year)	0.17 (0.13, 0.20)
Genital Warts (Past Year)	0.13 (0.09, 0.17)
Anal Sex (Past 3 Months)	0.24 (0.22, 0.26)
Sex with IDU (Past 3 Months)	0.94 (0.85, 1.02)
Sex with HIV+ Person (Past 3 Months)	0.76 (0.72, 0.79)
Sex with Sex Worker (Past 3 Months)	0.34 (0.36, 0.43)
Ecstasy (Past Year)	0.55 (0.52, 0.59)
Meth (Past Year)	1.11 (1.06, 1.20)
Nitrates (Past Year)	0.61 (0.58, 0.64)
ED Drugs (Past Year)	0.52 (0.48, 0.56)
Cocaine (Past Year)	0.57 (0.54, 0.6)
Alcohol (Past Year)	0.07 (0.05, 0.09)
Used IV Drugs (Past 3 Months)	0.85 (0.75, 0.93)
Partner 1 HIV positive	0.64 (0.59, 0.68)
Partner 1 HIV Status unknown	-0.01 (-0.04, 0.03)
Partner 1 different race	0.04 (0.01, 0.07)
Partner 2 HIV Positive	0.60 (0.55, 0.64)
Partner 2 HIV status Unknown	0.02 (0.00, 0.03)
Partner 2 different race	0.01 (-0.02, 0.03)
Intimate Partner Violence	0.26 (0.24, 0.28)

Table 4.1: Posterior summaries of loadings in one-factor model. Abbreviations: IDU means intravenous drug user, IV means intravenous, and ED means erectile dysfunction. Partner 1 is last partner, and Partner 2 is next to last partner.

Outcome[1]	Loading 1 (Mean, 95% CI)	Loading 2 (Mean, 95% CI)
Log(Partners + 1)	0.127 (0.087, 0.151)	-0.011 (-0.027, 0.003)
Partner 1 Age Difference	3.511 (2.946, 4.355)	0.255 (-0.144, 0.72)
Partner 2 Age Difference	3.211 (2.705, 3.957)	0.249 (-0.115, 0.669)
Gonorrhea / Chlamydia	0.082 (0.026, 0.121)	-0.014 (-0.028, -0.001)
Syphilis	0.105 (0.051, 0.15)	-0.008 (-0.023, 0.013)
Herpes	0.119 (0.068, 0.166)	0.027 (0, 0.055)
Genital Warts	0.101 (0.058, 0.144)	0.028 (0, 0.058)
Anal Sex Past 3 Months	0.206 (0.171, 0.232)	-0.005 (-0.03, 0.02)
Sex with IDU	0.658 (0.489, 0.772)	-0.003 (-0.087, 0.081)
Sex with HIV+ Person	0.613 (0.465, 0.704)	0.007 (-0.065, 0.076)
Sex with Sex Worker	0.291 (0.193, 0.361)	-0.034 (-0.072, 0.002)
Ecstasy	0.377 (0.242, 0.469)	0.015 (-0.032, 0.058)
Meth	0.523 (0.251, 0.797)	0.026 (-0.05, 0.095)
Nitrates	0.44 (0.307, 0.531)	0.009 (-0.045, 0.057)
ED Drugs	0.364 (0.241, 0.448)	0.005 (-0.042, 0.048)
Cocaine	0.39 (0.252, 0.482)	0.007 (-0.041, 0.052)
Alcohol	0.021 (-0.028, 0.058)	0.003 (-0.007, 0.013)
Used IV Drugs	0.624 (0.491, 0.728)	0.013 (-0.071, 0.099)
Partner 1 HIV positive	0.532 (0.412, 0.61)	0 (-0.065, 0.061)
Partner 1 HIV Status unknown	-0.01 (-0.032, 0.009)	-0.022 (-0.033, -0.012)
Partner 2 HIV Positive	0.493 (0.379, 0.58)	0.002 (-0.06, 0.061)
Partner 2 HIV status Unknown	0.01 (-0.012, 0.032)	-0.021 (-0.032, -0.01)
Intimate Partner Violence	0.182 (0.117, 0.227)	-0.007 (-0.033, 0.015)
Partner 1 different race	0.015 (-0.037, 0.065)	0.408 (0.375, 0.439)
Partner 2 different race	-0.016 (-0.07, 0.037)	0.446 (0.412, 0.479)

Table 4.2: Posterior summaries of loadings in two-factor model. Abbreviations: IDU means intravenous drug user, IV means intravenous, and ED means erectile dysfunction. Partner 1 is last partner, and Partner 2 is next to last partner.

Factor Type	Covariance Parameters		Decay Parameters		Effective Range
Longitudinal	σ_t	0.996 (0.994, 0.998)	ϕ_t	0.32 (0.30, 0.35)	9.3 (8.5, 10.1)
Spatial	σ_s	0.085 (0.060, 0.110)	ϕ_s	0.29 (0.14, 0.51)	11.0 (4.9, 21.7)

Table 4.3: Factor distribution parameters and derived quantities. Effective range is in years for longitudinal factors and miles for spatial factors, posterior mean (95% CI) for weighting parameter ρ_1 is $0.993(0.988, 0.996)$.

Covariate	Mean (95% CI)
Race	
White	REF
Black	0.34 (0.16, 0.50)
Hispanic	0.32 (0.23, 0.42)
Other	0.01 (-0.16, 0.18)
Factor Score	0.37 (0.32, 0.42)

Table 4.4: Posterior summaries of regression coefficients for HIV model. Age effects (not shown) were small and not significant.

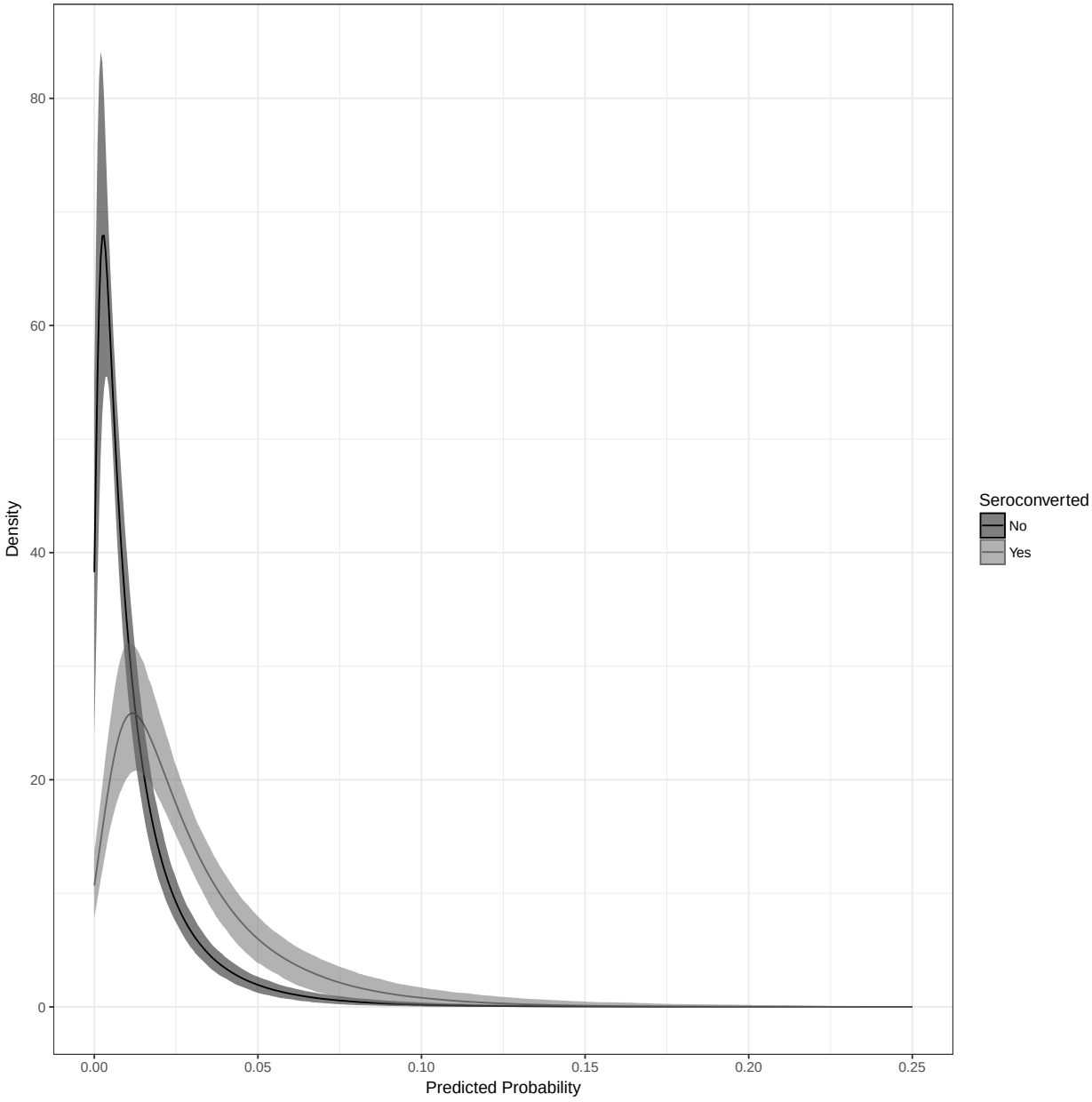

Figure 4.1: Plot of posterior densities for predicted probabilities for seroconveters (light gray) and non-seroconverters (dark gray). Solid lines are mean density for the samples at each point, and shaded regions are 95% pointwise credible bands for the density at each point. Probabilities calculated using each patient's final two visit times.

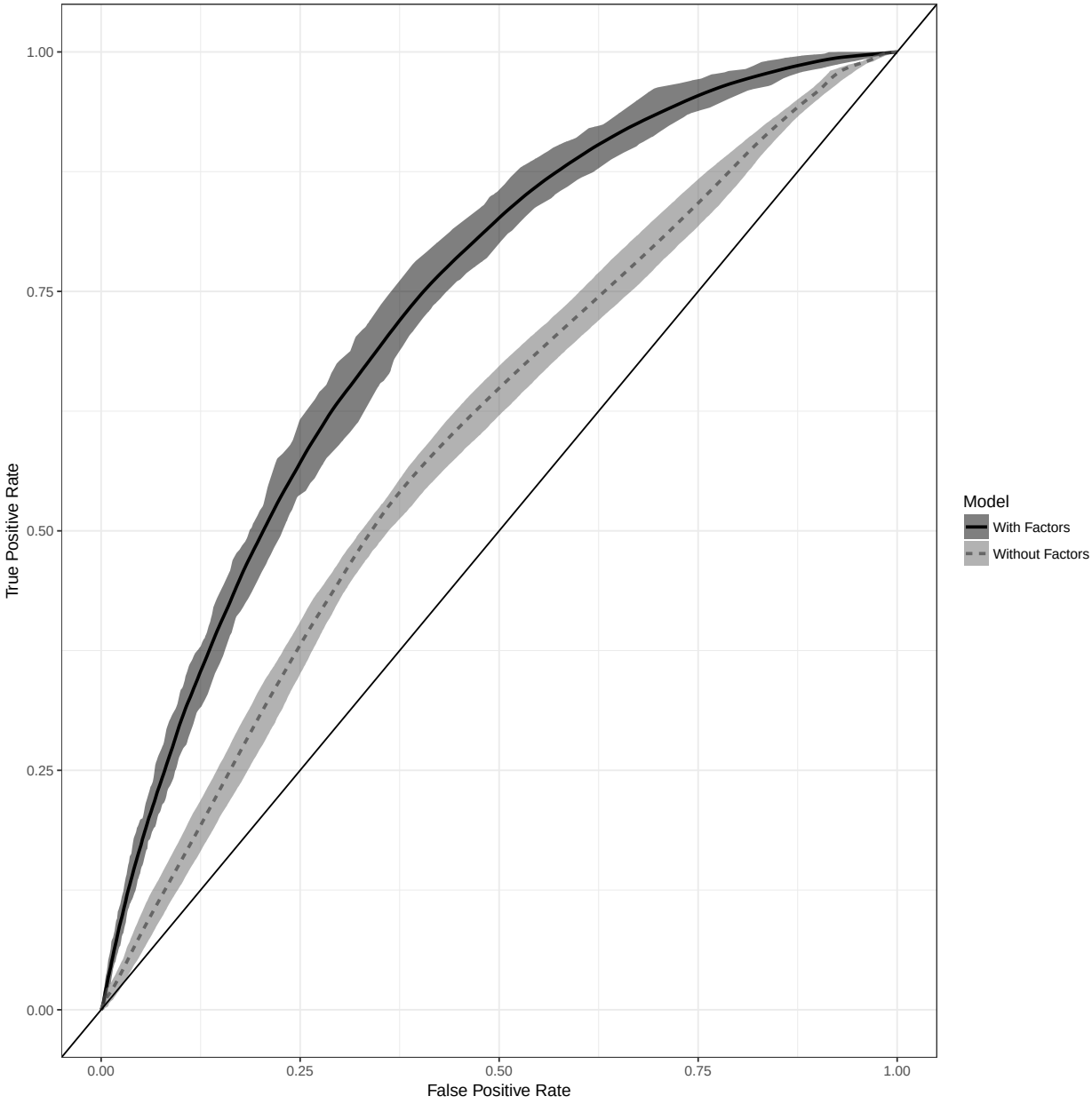

Figure 4.2: Plots of ROC curves with 95% pointwise credible bands for the model with factors (solid line, dark grey bands) and without factors (dashed, light gray bands).

Appendix A - MCMC Sampler Details

Let $i \in 1, \ldots, N$, index patients, $j \in 1, \ldots, n_i$ index visits, $p \in 1, \ldots, P$ index factors, and $k \in 1, \ldots, K$ index outcomes. We use random walk Metropolis Hastings steps for the temporal decay parameters ϕ_p, spatial decay parameters ν_p, and spatio-temporal scaling factor ρ. For all other parameters including the temporal factors $\boldsymbol{\tau}_{ij}$ and spatial factors ψ_s, we use Gibbs steps.

Let y_{ijk} be the k^{th} outcome for patient i at visit j, and $\boldsymbol{y}_k = (y_{11k}, \ldots, y_{Nn_Nk})^T$ be the vector of measurements for all patients and visits for outcome $k, k \in 1, \ldots, K$, where n_N is the last visit for patient N. Let z_{ij} be a binary indicator for patient i becoming HIV positive at visit j, and let $\boldsymbol{z} = (z_{11}, \ldots, z_{Nn_N})^T$. Let X_i be an $n_i \times M$ covariate matrix with rows x_{ij}^T for person i, and $X^T = (X_1^T, \ldots, X_N^T)$ be the full covariate matrix over all patients and visits. For the factors at visit j, let $\boldsymbol{\xi} = \left(\xi_{11}^T, \ldots, \xi_{1n_1}^T, \ldots, \xi_{Nn_n}^T\right)^T$ as a $n \times P$ matrix with rows $\boldsymbol{\xi}_{ij}$.

Finally, define the matrix P_i as an $(n_i - 1) \times n_i$ subsetting matrix which is the first n_i rows of the $n_i \times n_i$ identity matrix, and define $X_i^* = P_i X_i$, $X^* = \left(\bigoplus_i P_i\right) X$, and $\boldsymbol{\xi}^* = \left(\bigoplus_i P_i\right) \boldsymbol{\xi}$, where $\boldsymbol{A} \bigoplus \boldsymbol{B}$ denotes the direct sum of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\boldsymbol{A} \bigoplus \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}, \tag{4.25}$$

and $\bigoplus_i P_i$ denotes the direct sum over i of the matrices P_i. Then we can write the models in (4.2) and (4.5) as

$$\boldsymbol{y}_k = X\boldsymbol{\alpha}_k + \boldsymbol{\xi}\Lambda_k + \boldsymbol{\epsilon}_k, \tag{4.26}$$

$$\boldsymbol{\epsilon}_k \sim \mathrm{N}\left(\boldsymbol{0}, \sigma_k^2 I_n\right), \tag{4.27}$$

and

$$\boldsymbol{z} = X^*\boldsymbol{\beta} + \boldsymbol{\xi}^*\boldsymbol{\gamma} + \boldsymbol{\delta} \tag{4.28}$$

$$\boldsymbol{\delta} \sim \mathrm{N}\left(\boldsymbol{0}, I_n\right), \tag{4.29}$$

where $\mathbf{\Lambda}$ is the $K \times P$ matrix of factor loadings with rows Λ_k and $\boldsymbol{\gamma}$ is a p vector of regression coefficients. Thus for the regression coefficients $\boldsymbol{\alpha}_k$, the full conditional distributions are given by

$$\boldsymbol{\alpha}_k | \boldsymbol{y}_k, \Lambda, \boldsymbol{\xi}, \sigma_k^2 \sim \mathrm{N}_M \left(\mu^*_{\boldsymbol{\alpha}_k}, \Sigma^*_{\boldsymbol{\alpha}_k} \right), \tag{4.30}$$

$$\mu^*_{\boldsymbol{\alpha}_k} = \left(\frac{X^T X}{\sigma_k^2} + \Sigma_\alpha^{-1} \right)^{-1} \left(\frac{X^T \left(\boldsymbol{y}_k - \boldsymbol{\xi} \Lambda_k \right)}{\sigma_k^2} \right), \tag{4.31}$$

$$\Sigma^*_{\boldsymbol{\alpha}_k} = \left(\frac{X' X}{\sigma_k^2} + \Sigma_\alpha^{-1} \right)^{-1}, \tag{4.32}$$

where $\Sigma_\alpha = \mathrm{diag}(10^2, 1, \ldots, 1)$ is the prior covariance matrix for $\boldsymbol{\alpha}$. Similarly, the full conditionals for the rows Λ_k of the loadings matrix $\mathbf{\Lambda}$ are truncated multivariate normal if $k \leq P$ and multivariate normal if $k > P$,

$$\Lambda_k | \boldsymbol{y}_k, \boldsymbol{\alpha}_k, \boldsymbol{\xi}, \sigma_k^2 \sim \begin{cases} \mathrm{N}_P \left(\mu^*_{\Lambda_k}, \Sigma^*_{\Lambda_k} \right) \mathbf{1} \left\{ \lambda_{kk} > 0 \right\}, & \text{if } k \leq P \\ \mathrm{N}_P \left(\mu^*_{\Lambda_k}, \Sigma^*_{\Lambda_k} \right) & \text{if } k > P, \end{cases} \tag{4.33}$$

$$\mu^*_{\Lambda_k} = \left(\frac{\boldsymbol{\xi}^T \boldsymbol{\xi}}{\sigma_k^2} + I_P \right)^{-1} \left(\frac{\boldsymbol{\xi}^T \left(\boldsymbol{y}_k - X \boldsymbol{\alpha}_k \right)}{\sigma_k^2} \right), \tag{4.34}$$

$$\Sigma^*_{\Lambda_k} = \left(\frac{\boldsymbol{\xi}^T \boldsymbol{\xi}}{\sigma_k^2} + I_P \right)^{-1}. \tag{4.35}$$

For the regression coefficients $\boldsymbol{\beta}$, the full conditionals are given by

$$\boldsymbol{\beta} | \boldsymbol{y}_k, \boldsymbol{\gamma}, \boldsymbol{\xi} \sim \mathrm{N}_M \left(\mu^*_{\boldsymbol{\beta}}, \Sigma^*_{\boldsymbol{\beta}} \right), \tag{4.36}$$

$$\mu^*_{\boldsymbol{\beta}} = \left(X^{*T} X^* + \Sigma_\beta^{-1} \right)^{-1} \left(X^{*T} \left(\boldsymbol{y}_k - \boldsymbol{\xi}^* \boldsymbol{\gamma} \right) \right), \tag{4.37}$$

$$\Sigma^*_{\boldsymbol{\beta}} = \left(X^{*T} X^* + \Sigma_\beta^{-1} \right)^{-1}, \tag{4.38}$$

where $\Sigma_\beta = \mathrm{diag}(10^2, 1, \ldots, 1)$ is the prior covariance matrix for $\boldsymbol{\beta}$. For the regression coefficients $\boldsymbol{\gamma}$,

$$\boldsymbol{\gamma} | \boldsymbol{y}_k, \boldsymbol{\gamma}, \boldsymbol{\xi} \sim \mathrm{N}_M \left(\mu^*_{\boldsymbol{\gamma}}, \Sigma^*_{\boldsymbol{\gamma}} \right), \tag{4.39}$$

$$\mu^*_{\boldsymbol{\gamma}} = \left(\boldsymbol{\xi}^{*T} \boldsymbol{\xi}^* + I_P \right)^{-1} \left(\boldsymbol{\xi}^{*T} \left(\boldsymbol{y}_k - X^* \boldsymbol{\beta} \right) \right), \tag{4.40}$$

$$\Sigma^*_{\boldsymbol{\beta}} = \left(\boldsymbol{\xi}^{*T} \boldsymbol{\xi}^* + I_P \right)^{-1}, \tag{4.41}$$

To derive the full conditionals for the longitudinal factors $\boldsymbol{\tau}_{ij}$, we first write the density for $\boldsymbol{y}_{ij}$ as

$$f(\boldsymbol{y}_{ij}|\boldsymbol{\alpha},\Lambda,\boldsymbol{\tau}_{ij},\boldsymbol{\psi}_{s(i,j)},\Sigma_y) \propto |\boldsymbol{\Sigma}_y|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{y}_{ij}-\boldsymbol{\alpha}x_i-\Lambda\boldsymbol{\psi}_{s(i,j)}\right)-\Lambda\boldsymbol{\tau}_{ij}\right]^T\Sigma_y^{-1}\right.$$
$$\left.\left[\left(\boldsymbol{y}_{ij}-\boldsymbol{\alpha}x_i-\Lambda\boldsymbol{\psi}_{s(i,j)}\right)-\Lambda\boldsymbol{\tau}_{ij}\right]\right\}., \quad (4.42)$$

where $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\ldots,\boldsymbol{\alpha}_k)^T$ is a $K\times M$ matrix of regression coefficients. Then letting $\boldsymbol{\tau}_i$ be an $n_iP\times 1$ vector of longitudinal factors formed by stacking the n_i vectors $\boldsymbol{\tau}_{ij}$, we can take advantage of the conditional independence of the $\boldsymbol{y}_{ij}$ given the factors and write the log likelihood for $\boldsymbol{y}$ as

$$\log L(\boldsymbol{\tau}_i|\boldsymbol{y}) \propto \boldsymbol{\tau}_i^T\left(I_{n_i}\otimes\Lambda^T\Sigma_y^{-1}\Lambda\right)\boldsymbol{\tau}_i-$$
$$2\boldsymbol{\tau}_i^T\left[I_{n_i}\otimes\Lambda^T\Sigma_y^{-1}\right]\left(\left[\boldsymbol{y}_{i1}^T,\ldots,y_{in_i}^T\right]^T-\mathbf{1}_{n_i}\otimes(\boldsymbol{\alpha}x_i)-(I_{n_i}\otimes\Lambda)\,\boldsymbol{\psi}_{s(i,.)}\right), \quad (4.43)$$

where $\otimes$ denotes the Kronecker product. Similarly, for seroconversion, the density for $\boldsymbol{z}_i$ is given by

$$f(\boldsymbol{z}_i|\boldsymbol{\beta},\gamma,\boldsymbol{\tau}_{ij},\boldsymbol{\psi}_{s(i,.)}) \propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{z}_i-\left(\boldsymbol{\beta}^T\otimes\mathbf{1}_{n_i}\right)x_i-P_i\boldsymbol{\psi}_{s(i,.)}\boldsymbol{\gamma}\right)-P_i\boldsymbol{\tau}_i\boldsymbol{\gamma}\right]^T\right.$$
$$\left.\left[\left(\boldsymbol{z}_i-\left(\boldsymbol{\beta}^T\otimes\mathbf{1}_{n_i}\right)x_i-P_i\boldsymbol{\psi}_{s(i,.)}\boldsymbol{\gamma}\right)-P_i\boldsymbol{\tau}_i\boldsymbol{\gamma}\right]\right\}, \quad (4.44)$$

We can write $P_i\boldsymbol{\tau}_i\boldsymbol{\gamma}=(\boldsymbol{\gamma}^T\otimes P_i)\boldsymbol{\tau}_i$, which allows us to write

$$\log L(\boldsymbol{\tau}_i|\boldsymbol{z}_i) \propto \boldsymbol{\tau}_i^T\left[\left(\boldsymbol{\gamma}^T\otimes P_i\right)^T\left(\boldsymbol{\gamma}^T\otimes P_i\right)\right]\boldsymbol{\tau}_i-$$
$$2\boldsymbol{\tau}_i^T\left(\boldsymbol{\gamma}^T\otimes P_i\right)\left(\boldsymbol{z}_i-\left(\boldsymbol{\beta}^T\otimes\mathbf{1}_{n_i}\right)x_i-P_i\boldsymbol{\psi}_{s(i,.)}\boldsymbol{\gamma}\right). \quad (4.45)$$

Therefore, the full conditional for $\boldsymbol{\tau}_i$ is multivariate normal,

$$f(\boldsymbol{\tau}_i|\boldsymbol{y},\boldsymbol{\alpha},\lambda,\Sigma_y) \sim \mathrm{N}\left(\mu^*_{\boldsymbol{\tau}_i},\Sigma^*_{\boldsymbol{\tau}_i}\right) \quad (4.46)$$

$$\Sigma^*_{\boldsymbol{\tau}_i} = \left(I_{n_i}\otimes\Lambda^T\Sigma_y^{-1}\Lambda+\left(\boldsymbol{\gamma}^T\otimes P_i\right)^T\left(\boldsymbol{\gamma}^T\otimes P_i\right)+\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\tau}_i}\right)^{-1} \quad (4.47)$$

$$\mu^*_{\boldsymbol{\tau}_i} = \Sigma^*_{\boldsymbol{\tau}_i}\left[\left(I_{n_i}\otimes\Lambda^T\Sigma_y^{-1}\right)(\boldsymbol{y}-\mathbf{1}_{n_i}\otimes\boldsymbol{\alpha}x_i)+\right.$$
$$\left.\left(\boldsymbol{\gamma}^T\otimes P_i\right)\left(\boldsymbol{z}_i-\left(\boldsymbol{\beta}^T\otimes\mathbf{1}_{n_i}\right)x_i\right)\right]. \quad (4.48)$$

Finally, for the spatial factors, we again take advantage of the conditional independence of the $\boldsymbol{y}_{ij}$ and z_{ij} given all the factors, and write for $s \in 1, \ldots, S$,

$$\log L(\psi_s|\boldsymbol{y}) \propto \psi_s^T(n_s\Lambda^T\Sigma_y^{-1}\Lambda)\psi_s - 2\psi_s^T\Lambda^T\Sigma_y^{-1}\left(\sum_{i,j:s(i,j)=s}(\boldsymbol{y}_{ij} - \boldsymbol{\alpha}x_i - \Lambda\boldsymbol{\tau}_{ij})\right), \tag{4.49}$$

and

$$\log L(\psi_s|\boldsymbol{z}) \propto \psi_s^T(n_{s-1}\boldsymbol{\gamma}\boldsymbol{\gamma}^T)\psi_s - 2\psi_s^T\boldsymbol{\gamma}\left(\sum_{i,j:s(i,(j-1))=s}(z_{ij} - \boldsymbol{\alpha}x_i - \boldsymbol{\gamma}^T\tau_{i(j-1)})\right), \tag{4.50}$$

where n_s is the sum over i and j of records on patients living in ZIP code s, and n_{s-1} is the sum over i and $j-1$ of records on patients living in ZIP code s. Thus, the full conditional for ψ is also multivariate normal

$$\psi|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda, \Sigma_y \sim \mathrm{N}\left(\mu_\psi^*, \Sigma_\psi^*\right), \tag{4.51}$$

where

$$\Sigma_\psi^* = \left(\mathrm{diag}(n_s) \otimes \Lambda^T\Sigma_y^{-1}\Lambda + \mathrm{diag}\left(n_{s-1}\right) \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}^T + \boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\right)^{-1} \tag{4.52}$$

$$\mu_\psi^* = \Sigma_\psi^*\left[\left(I_S \otimes \Lambda^T\Sigma_y^{-1}\right)\mu_{\psi y} + \left(I_S \otimes \boldsymbol{\gamma}\right)\mu_{\psi z}\right], \tag{4.53}$$

where

$$\mu_{\psi y} = \begin{pmatrix} \sum_{i,j:s(i,j)=1}(\boldsymbol{y}_{ij} - \boldsymbol{\alpha}x_i - \Lambda\boldsymbol{\tau}_{ij}) \\ \vdots \\ \sum_{i,j:s(i,j)=S}(\boldsymbol{y}_{ij} - \boldsymbol{\alpha}x_i - \Lambda\boldsymbol{\tau}_{ij}) \end{pmatrix} \tag{4.54}$$

and

$$\mu_{\psi z} = \begin{pmatrix} \sum_{i,j:s(i,(j-1))=1}(z_{ij} - \boldsymbol{\alpha}x_i - \boldsymbol{\gamma}^T\tau_{i(j-1)}) \\ \vdots \\ \sum_{i,j:s(i,(j-1))=S}(z_{ij} - \boldsymbol{\alpha}x_i - \boldsymbol{\gamma}^T\tau_{i(j-1)}) \end{pmatrix}. \tag{4.55}$$

CHAPTER 5

Conclusions and Future Work

In Chapter 2, we presented a multivariate logistic regression model with spatially correlated random effects to jointly model history of STIs, history of illicit drug use, and HIV seroconversion by the end of the study. We then developed a statistic called the GMR to assess the extent to which the spatial heterogeneity in HIV risk is explained by the STIs and drug use random effects. In Chapter 3, we jointly modeled a patient's time to HIV seroconversion and frequency of clinic visits, and showed that if seroconversion is correlated with visit frequency that the censoring is informative. Finally in Chapter 4, we used a spatio-temporal factor analysis model to construct a set of measures that capture aspects of a patient's propensity for risky behavior. We then treated the factor scores as predictors in a model for predicting HIV seroconversion by the time of their next visit. In the rest of this chapter, we discuss some potential areas for further developing our methods.

5.1 Extensions to the GMR

To construct the GMR in Chapter 2, we started with a model for the random effects that had the form

$$\boldsymbol{Y} = \sum_{k=1}^{K} a_k \boldsymbol{X}_k + \boldsymbol{\epsilon}, \tag{5.1}$$

where $\boldsymbol{Y}$ is an $S \times 1$ multivariate normal random variable, the $\boldsymbol{X}_k$ are correlated S-dimensional Gaussian processes, and $\boldsymbol{\epsilon}$ is an S-dimensional Gaussian process independent of the $\boldsymbol{X}_k$. We then defined $\text{GMR} = \left\{\det\left(\text{Var}\left(\boldsymbol{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k\right)\right) / \det\left(\text{Var}\left(\boldsymbol{Y}\right)\right)\right\}^{1/S}$ and showed that the GMR is bounded in the interval $(0, 1)$ and equals 1 if and only if all the a_k are zero.

A more general way to construct the GMR is to start with a multivariate normal vector

$(\boldsymbol{X}_1^T, \boldsymbol{X}_2^T)$ distributed as

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right) \tag{5.2}$$

so that

$$\boldsymbol{X}_2 | \boldsymbol{X}_1 \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right), \tag{5.3}$$

where $\boldsymbol{X}_1$ has dimension S_1 and $\boldsymbol{X}_2$ has dimension S_2. We can then define the GMR as

$$\mathrm{GMR} = \left[\frac{\det\left(\mathrm{Var}\left(\boldsymbol{X}_2 | \boldsymbol{X}_1\right)\right)}{\det\left(\mathrm{Var}\left(\boldsymbol{X}_2\right)\right)} \right]^{1/S_2}, \tag{5.4}$$

$$= \left[\frac{\det\left(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)}{\det\left(\boldsymbol{\Sigma}_{22}\right)} \right]^{1/S_2}. \tag{5.5}$$

Under the construction in 5.5, the GMR is still bounded within the interval (0,1) but now GMR = 1 if and only if $\boldsymbol{\Sigma}_{21}$ is the zero matrix.

In addition to the potential theoretical developments for the GMR, we can also examine its behavior under a variety of models. For example, we made the claim that strictly speaking, the GMR is not the fraction of the spatial variability in the stage two random effects that is explained by the stage 1 random effects, because for any two positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\frac{|\boldsymbol{A}|}{|\boldsymbol{A}+\boldsymbol{B}|} + \frac{|\boldsymbol{B}|}{|\boldsymbol{A}+\boldsymbol{B}|} \leq 1. \tag{5.6}$$

In (5.5), $\boldsymbol{A} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, and $\boldsymbol{B} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Our simulations showed that under the construction in (5.1), the left hand side of (5.6) is approximately equal to 1 under a wide range of values for the Gaussian process variances and decay parameters, and we would like to investigate whether this is the case under the more general construction in (5.5).

5.2 Survival Model for HIV Seroconversion Times

In Chapter 3, we modeled patients' HIV seroconversion times as lognormally distributed and modeled their total number of clinic visits by approximating a zero-truncated Poisson

process with a non-truncated Poisson process. We model correlation between seroconversion times and clinic visit frequency with correlated frailties. The main reason for approximating the zero-truncated Poisson distribution with a non-truncated Poisson distribution was to analytically evaluate the marginal correlation between survival and clinic visits, but if we instead model clinic visits as an exact zero-truncated Poisson process, we can still approximately calculate the marginal correlation between survival and clinic visit frequency with the MCMC samples. This would also make the simulations match exactly the model that we are using to fit the data.

We also want to extend the survival portion of the model by considering other parametric distributions such as Weibull or log-logistic. The interval-censoring will still be informative if the survival times are correlated with clinic visits, and we can compare different distributions to determine which one fits the data best, and under which distribution or distributions the interval censoring is most informative.

5.3 Generalizing the Factor Model

In the factor model in Chapter 4, we modeled all outcomes, including HIV serostatus, as functions of latent normal random variables. One of the outcomes we considered was the log of a patient's number of partners in the last month. Because this is a count outcome, it may be more appropriate to model it as a Poisson or Negative Binomial random variable.

More generally, instead of treating y_{ijk} as a function of a latent normal random variable, where y_{ijk} is outcome k for patient i at visit j, we would like to consider more general models of the form

$$g_k(\mathrm{E}\left[y_{ijk}\right]) = x_{ij}^T \boldsymbol{\alpha}_k + \Lambda_k^T \boldsymbol{\xi}_{ij}, \tag{5.7}$$

where x_{ij} is a covariate vector allows for time-varying covariates, g_k is a link function, $\boldsymbol{\alpha}_k$ are regression coefficients, $\boldsymbol{\Lambda}$ is a loadings matrix with rows Λ_k^T, and $\boldsymbol{\xi}_{ij}$ are factors. This more general model structure for the factors is nice because it allows the y_{ijk} to come from any member of exponential family distributions. It may be the case that this negatively impacts

the autocorrelation of the factors and loadings. Ghosh and Dunson (2009) used parameter expansion to improve the mixing of a factor model with all normal outcomes, and we could potentially extend their methods to the more general model in (5.7).

Bibliography

Abellan, J. J., Richardson, S., and Best, N. (2008). Use of space – time models to investigate the stability of patterns of disease. *Environmental Health Perspectives* **116,** 1111 – 1119.

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics* **18,** 338 – 357.

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76,** 643 – 653.

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics* **29,** 156 – 163.

Ansari, A. and Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika* **65,** 475 – 496.

Arminger, G. and Küsters, U. (1988). Latent Trait Models with Indicators of Mixed Measurement Level. In *Latent Trait and Latent Class Models*, pages 51 – 73. Springer.

Bachireddy, C., Soule, M. C., Izenberg, J. M., Dvoryak, S., Dumchev, K., and Altice, F. L. (2014). Integration of health services improves multiple healthcare outcomes among HIV-infected people who inject drugs in Ukraine. *Drug and Alcohol Dependence* **134,** 106 – 114.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, 2nd edition.

Banerjee, S. and Gelfand, A. E. (2002). Prediction, interpolation and regression for spatially misaligned data. *Sankhya A* **64,** 227 – 245.

Belot, A., Rondeau, V., Remontet, L., Giorgi, R., and CENSUR working survival group (2014). A joint frailty model to estimate the recurrence process and the disease-specific

mortality process without needing the cause of death. *Statistics in Medicine* **33,** 3147 – 3166.

Bhandary, M. (1996). Test for generalized variance in signal processing. *Statistics and Probability Letters* **27,** 155 – 162.

Bhandary, M. (2006). Test for generalized variance in factor analysis model. *Communications in Statistics-Simulation and Computation* **35,** 969 – 973.

Bompotas, P., Kimber, A., and Biedermann, S. (2017). Sensitivity analysis for informative censoring in parametric survival models: an evaluation of the method.

Buchacz, K., McFarland, W., Kellogg, T. A., Loeb, L., Holmberg, S. D., Dilley, J., and Klausner, J. D. (2005). Amphetamine use is associated with increased HIV incidence among men who have sex with men in San Francisco. *AIDS* **19,** 1423 – 4438.

Cai, N., Lu, W., and Zhang, H. H. (2012). Time-varying latent effect model for longitudinal data with informative observation times. *Biometrics* **68,** 1093 – 1102.

Campigotto, F. and Weller, E. (2014). Impact of informative censoring on the Kaplan-Meier estimate of progression-free survival in phase II clinical trials. *Journal of Clinical Oncology* **32,** 3068.

Cheng, W., Gill, G. S., Zhang, Y., and Cao, Z. (2018). Bayesian spatiotemporal crash frequency models with mixture components for space-time interactions. *Accident Analysis & Prevention* **112,** 84 – 93.

Chib, S. and Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics* **19,** 428 – 435.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., and Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics* **183,** 31 – 57.

Cowling, B., Hutton, J., and Shaw, J. (2006). Joint modelling of event counts and survival times. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **55,** 31 – 39.

Cramb, S. M., Baade, P. D., White, N. M., Ryan, L. M., and Mengersen, K. L. (2015). Inferring lung cancer risk factor patterns through joint Bayesian spatio-temporal analysis. *Cancer Epidemiology* **39,** 430 – 439.

Crow, E. L. and Shimizu, K. (1987). *Lognormal Distributions*. Marcel Dekker New York.

Crowther, M. J. (2017). Extended multivariate generalised linear and non-linear mixed effects models. arXiv:1710.02223.

Daniels, M. J. and Zhao, Y. D. (2003). Modelling the random effects covariance matrix in longitudinal data. *Statistics in Medicine* **22,** 1631 – 1647.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics* **42,** 204 – 223.

Dickey, J. M. and Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov Chain. *The Annals of Mathematical Statistics* **41,** 215 – 226.

Dunson, D. B. (2007). Bayesian methods for latent trait modelling of longitudinal data. *Statistical Methods in Medical Research* **16,** 399 – 415.

Fisher, D. G., Reynolds, G. L., Ware, M. R., and Napper, L. E. (2011). Methamphetamine and Viagra use: relationship to sexual risk behaviors. *Archives of Sexual Behavior* **40,** 273 – 279.

Fleming, D. T. and Wasserheit, J. N. (1999). From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection. *Sexually Transmitted Infections* **75,** 3 – 17.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7,** 457 – 511.

Geweke (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, volume 4, pages 169 – 193. Oxford University Press.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies* **9,** 557 – 587.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics* **18,** 306 – 320.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association* **105,** 1167 – 1177.

Gómez, G., Calle, M. L., Oller, R., and Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling* **9,** 259 – 297.

Grinsztejn, B., Hosseinipour, M. C., Ribaudo, H. J., Swindells, S., Eron, J., Chen, Y. Q., et al. (2014). Effects of early versus delayed initiation of antiretroviral treatment on clinical outcomes of HIV-1 infection: results from the phase 3 HPTN 052 randomised controlled trial. *The Lancet Infectious Diseases* **14,** 281 – 290.

Grover, G., Swain, P. K., Deo, V., and Varshney, M. K. (2015). A joint modeling approach to assess the impact of CD4 cell count on the risk of loss to follow up in HIV/AIDS patients on antiretroviral therapy. *International Journal of Statistics and Applications* **5,** 99 – 108.

Hao, J. and Krishnamoorthy, K. (2001). Inferences on a normal covariance matrix and generalized variance with monotone missing data. *Journal of Multivariate Analysis* **78,** 62 – 82.

Hedeker, D., Mermelstein, R. J., and Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics* **64,** 627 – 634.

Hu, Z., Wong, C., Thach, T., Lam, T., and Hedley, A. (2004). Binary latent variable modelling and its applicationin the study of air pollution in hong kong. *Statistics in Medicine* **23,** 667 – 684.

Huang, X. and Liu, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics* **63,** 389 – 397.

Iliopoulos, G. and Kourouklis, S. (1999). Improving on the best affine equivariant estimator of the ratio of generalized variances. *Journal of Multivariate Analysis* **68,** 176 – 192.

Konomi, B., Karagiannis, G., and Lin, G. (2015). On the Bayesian treed multivariate Gaussian process with linear model of coregionalization. *Journal of Statistical Planning and Inference* **157,** 1 – 15.

Król, A., Ferrer, L., Pignon, J.-P., Proust-Lima, C., Ducreux, M., Bouché, O., Michiels, S., and Rondeau, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000 – 05 trial. *Biometrics* **72,** 907 – 916.

Liang, Y., Lu, W., and Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* **65,** 377 – 384.

Liu, L., Huang, X., and O'Quigley, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* **64,** 950 – 958.

Lopes, H. F., Gamerman, D., and Salazar, E. (2011). Generalized spatial dynamic factor models. *Computational Statistics & Data Analysis* **55,** 1319 – 1330.

Lopes, H. F., Salazar, E., and Gamerman, D. (2008). Spatial dynamic factor analysis. *Bayesian Analysis* **3,** 759 – 792.

Luttinen, J. and Ilin, A. (2009). Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Advances in Neural Information Processing Systems*, pages 1177 – 1185.

Ma, Z. and Krings, A. W. (2008). Multivariate survival analysis (i): shared frailty approaches to reliability and dependence modeling. In *Aerospace Conference, 2008 IEEE*, pages 1 – 21. IEEE.

Marcus, M. and Minc, H. (1992). *A Survey of Matrix Theory and Matrix Inequalities*, volume 14. Courier Corporation.

Martins, R., Silva, G. L., and Andreozzi, V. (2016). Bayesian joint modeling of longitudinal and spatial survival AIDS data. *Statistics in Medicine* **35,** 3368 – 3384.

Mathai, A. (1972). The exact distributions of three multivariate statistics associated with wilks' concept of generalized variance. *Sankhyā: The Indian Journal of Statistics, Series A* pages 161 – 170.

Matheron, G. (1982). Pour une analyse krigeante des données régionalisées. *Centre de Géostatistique, Report N-732, Fontainebleau* page 22.

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research* **17,** 53 – 73.

Mostafa, M. and Mahmoud, M. (1964). On the problem of estimation for the bivariate lognormal distribution. *Biometrika* **51,** 522 – 527.

Oller, R., Gómez, G., and Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics* **32,** 315 – 326.

Orton, T., Pringle, M., Page, K., Dalal, R., and Bishop, T. (2014). Spatial prediction of soil organic carbon stock using a linear model of coregionalisation. *Geoderma* **230,** 119 – 130.

Plankey, M. W., Ostrow, D. G., Stall, R., Cox, C., Li, X., Peck, J. A., and Jacobson, L. P. (2007). The relationship between methamphetamine and popper use and risk of HIV seroconversion in the multicenter AIDS cohort study. *Journal of Acquired Immune Deficiency Syndromes* **45,** 85 – 92.

Quinn, K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis* **12,** 338 – 353.

Raftery, A. E. and Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science* **7,** 493 – 497.

Richardson, S., Abellan, J. J., and Best, N. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Statistical Methods in Medical Research* **15,** 385 – 407.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59,** 731 – 792.

Schliep, E. M., Lany, N. K., Zarnetske, P. L., Schaeffer, R. N., Orians, C. M., Orwig, D. A., and Preisser, E. L. (2018). Joint species distribution modelling for spatio-temporal occurrence and ordinal abundance data. *Global Ecology and Biogeography* **27,** 142 – 155.

Schmidt, M. N. (2009). Function factorization using warped Gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 921 – 928. ACM.

Schmidt, M. N. and Laurberg, H. (2008). Nonnegative matrix factorization with Gaussian process priors. *Computational Intelligence and Neuroscience* **2008,** 3.

SenGupta, A. (1987). Generalizations of Barlett's and Hartley's tests of homogeneity using overall variability. *Communications in Statistics-Theory and Methods* **16,** 987 – 996.

SenGupta, A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis* **23,** 209 – 219.

Siannis, F., Copas, J., and Lu, G. (2005). Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics* **6,** 77 – 91.

Sinha, D., Chen, M.-H., and Ghosh, S. K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics* **55,** 585 – 590.

Specht, D. A. (1975). On the evaluation of causal models. *Social Science Research* **4,** 113 – 133.

Strickland, C., Simpson, D., Turner, I., Denham, R., and Mengersen, K. (2011). Fast Bayesian analysis of spatial dynamic factor models for multitemporal remotely sensed imagery. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60,** 109 – 124.

Sun, J., Sun, L., and Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association* **102,** 1397 – 1406.

Sun, L., Song, X., Zhou, J., and Liu, L. (2012). Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *Journal of the American Statistical Association* **107,** 688 – 700.

Sunethra, A. and Sooriyarachchi, M. (2016). Joint modeling of mortality incidence and survival.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., and Zipkin, E. F. (2016). Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography* **25,** 1144 – 1158.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey ratio. *Journal of the American Statistical Association* **90,** 614 – 618.

Vergeynst, L., Van Langenhove, H., and Demeestere, K. (2015). Balancing the false negative and positive rates in suspect screening with high-resolution orbitrap mass spectrometry using multivariate statistics. *Analytical Chemistry* **87,** 2170 – 2177.

Verity, C., Hosking, G., and Easter, D. (1995). A multicentre comparative trial of sodium valproate and carbamazepine in paediatric epilepsy. *Developmental Medicine & Child Neurology* **37,** 97 – 108.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika* **24,** 471 – 494.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53,** 330 – 339.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* **3,** 32 – 35.

Zeng, D., Ibrahim, J. G., Chen, M.-H., Hu, K., and Jia, C. (2014). Multivariate recurrent events in the presence of multivariate informative censoring with applications to bleeding and transfusion events in myelodysplastic syndrome. *Journal of Biopharmaceutical Statistics* **24,** 429 – 442.

Zhang, Z. and Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research* **19,** 53 – 70.

Zhang, Z., Sun, J., and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine* **24,** 1399 – 1407.

Zhang, Z., Sun, L., Sun, J., and Finkelstein, D. M. (2007). Regression analysis of failure time data with informative interval censoring. *Statistics in Medicine* **26,** 2533 – 2546.

Zhao, S., Hu, T., Ma, L., Wang, P., and Sun, J. (2015). Regression analysis of interval-censored failure time data with the additive hazards model in the presence of informative censoring. *Statistics and Its Interface* **8,** 367 – 377.

Zhu, Y. and Weiss, R. E. (2013). Modeling seroadaptation and sexual behavior among HIV+ study participants with a simultaneously multilevel and multivariate longitudinal count model. *Biometrics* **69,** 214 – 224.

Zuur, A. F., Fryer, R., Jolliffe, I., Dekker, R., and Beukema, J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics* **14,** 665 – 685.