Nonparametric Bayes Analysis of Social Science Data

by

Tsuyoshi Kunihama

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Jerome P. Reiter

_____
Fan Li

_____
Allison Ashley-Koch

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2015

Abstract

Nonparametric Bayes Analysis of Social Science Data

by

Tsuyoshi Kunihama

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Jerome P. Reiter

_____
Fan Li

_____
Allison Ashley-Koch

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2015

Abstract

Social science data often contain complex characteristics that standard statistical methods fail to capture. Social surveys assign many questions to respondents, which often consist of mixed-scale variables. Each of the variables can follow a complex distribution outside parametric families and associations among variables may have more complicated structures than standard linear dependence. Therefore, it is not straightforward to develop a statistical model which can approximate structures well in the social science data. In addition, many social surveys have collected data over time and therefore we need to incorporate dynamic dependence into the models. Also, it is standard to observe massive number of missing values in the social science data. To address these challenging problems, this thesis develops flexible nonparametric Bayesian methods for the analysis of social science data.

Chapter 1 briefly explains backgrounds and motivations of the projects in the following chapters. Chapter 2 develops a nonparametric Bayesian modeling of temporal dependence in large sparse contingency tables, relying on a probabilistic factorization of the joint pmf. Chapter 3 proposes nonparametric Bayes inference on conditional independence with conditional mutual information used as a measure of the strength of conditional dependence. Chapter 4 proposes a novel Bayesian density estimation method in social surveys with complex designs where there is a gap between sample and population. We correct for the bias by adjusting mixture weights in Bayesian mixture models. Chapter 5 develops a nonparametric model for mixed-scale longi-

tudinal surveys, in which various types of variables can be induced through latent continuous variables and dynamic latent factors lead to flexibly time-varying associations among variables.

To my family

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

Be	Beta
AR	Autoregressive
DP	Dirichlet Process
DX	Dunson and Xing (2009).
Ga	Gamma
IG	Inverse-Gamma
MCMC	Markov chain Monte Carlo

Acknowledgements

To begin with, I would like to thank my advisor, David Dunson, for his insightful guidance, invaluable advice and outstanding support throughout my time at Duke. I have been impressed by his great enthusiasm to new problems and he provided me a perfect role model as a leading researcher with full of innovative and inspiring ideas. I would also like to thank Jerome Reiter, Fan Li and Allison Ashley-Koch for serving on my committee and for providing valuable comments. Thanks also to all the department faculty and staff members for giving me the great environment for my study. I am also grateful to Nakajima foundation for its generous financial support of my graduate studies.

I would like to thank Yasuhiro Omori for introducing me into the wonderful world of statistics and providing me constant support in my graduate studies. I also would like to thank Jouchi Nakajima for his enormous support and guidance in my time at Duke. Huge thanks to my best classmates for their friendship: Daniel, Maria, Mary Beth, Nick, Thais, Tim and Tommy. I am also very grateful to all other Duke friends for their support and companionship over the years. They have enriched my life at Duke with lots of fun.

I would like to thank my family for their constant support throughout my life. Finally, my deepest thanks to my wife Ayako for her invaluable support and encouragement all the time. Without her, I would have never accomplished my PhD program.

1

Introduction

To understand structures and development of societies, many social surveys have been conducted with various research interests. These surveys contain demographic, behavioural and attitudinal questions, which include critical information for studying social phenomena in population or sub-populations of special interest. However, social science data often holds complex characteristics, which make it challenging to build realistic statistical models for extracting the information from the data. For example, General Social Survey consists of categorical questions such as race, education level and political party affiliations, producing a large contingency table with enormous number of cells compared to a sample size. The large sparse table requires a massive number of parameters in standard log-linear models, leading to computational intractability.

Also, various private and government research institutions have collected data over a period of time to investigate trends of structures of societies. To achieve the research goal, it is crucial to incorporate time-dependence appropriately into statistical models but dynamic structures depend on survey designs. For longitudinal studies, same respondents repeatedly answer questions, inducing subject-specific

time-dependence. On the other hand, in other surveys such as General Social Survey, different individuals are randomly collected from population at each time point, resulting in no subject-specific dependence, but we still need to express dynamic structures of the population. Hence, it is important to develop statistical models taking fully into account the survey purpose and designs.

Multivariate response variables in a large-scale survey are usually mixed-scale. For example, The National Longitudinal Study of Adolescent to Adult Health has collected sexual behavior variables, including attraction to same sex (binary variable), number of sex partners (count variable) and sexual self-definition (unordered categorical variable). However, it is not straightforward to construct a flexible statistical method for the analysis of mixed-scale data. Also, although one of research interests is often in estimating associations among response variables, it is not clear even how to measure interactions for mixed-scale data. In addition, existing methods for conditional associations between a response variable and covariates given other variables are usually based on a linear relationship, which can be too restrictive for social science studies with complex structures.

In addition, social surveys are often collected via complex sampling designs such as stratified sampling and oversampling, leading to discrepancies between sample and population. On the other hand, standard statistical models are based on the assumption that the collected respondents are representative of population. Therefore, one needs to adjust for the bias to obtain an estimation result for target population. Although there are several existing adjustment methods in the literature, they often rely on restrictive associations between a response and survey weights or complicated joint modeling of them. Hence, it is important to develop a flexible but not complicated approach with easy computation.

To address these challenging problems in the analysis of social science data, this thesis develops novel Bayesian methods based on joint modeling of multivariate sur-

vey data. The class of joint models is comprehensive that it can be applied for a wide variety of studies with various research interests. For example, the joint models can flexibly induce conditional models given covariates because a conditional density can be expressed as a ratio of two joint densities in general. Therefore, conditional associations between a response and predictors given other variables can be analysed in this framework. Also, missing values can be easily imputed from the joint model assuming missing at random. To build a flexible joint model, we utilize nonparametric Bayes density estimation methods, relying on Dirichlet process mixtures (West et al. (1994); Escobar and West (1995); Müller et al. (1996); Hannah et al. (2011)). For the DP mixtures, various efficient posterior computation methods are developed such as the Polya urn scheme (Bush and MacEachern (1996)) and the blocked Gibbs sampler (Ishwaran and James (2001)). Also, the slice sampler (Walker (2007); Kalli et al. (2011)) and retrospective MCMC methods (Papaspiliopoulos and Roberts (2008)) can avoid truncation approximations in the computation. In addition, there is an emerging literature providing an asymptotic frequentist justification for these models (Ghosal et al. (1999); Ghosh and Ramamoorthi (2003); Tokdar (2006); Wu and Ghosal (2008)).

Chapter 2 addresses problems in social surveys with categorical responses. In many applications in social science, it is of interest to study trends over time in relationships among categorical variables, such as age group, ethnicity, religious affiliation, political party and preference for particular policies. At each time point, a sample of individuals provide responses to a set of questions, with different individuals sampled at each time. In such settings, there tends to be abundant missing data and the variables being measured may change over time. At each time point, one obtains a large sparse contingency table, with the number of cells often much larger than the number of individuals being surveyed. To borrow information across time in modeling large sparse contingency tables, we propose a Bayesian autore-

gressive tensor factorization approach. The proposed model relies on a probabilistic Parafac factorization of the joint pmf characterizing the categorical data distribution at each time point, with autocorrelation included across times. Efficient computational methods are developed relying on MCMC. The methods are evaluated through simulation examples and applied to social survey data. The result in this chapter was published in Kunihama and Dunson (2013).

Chapter 3 tackles the problem of conditional associations among mixed-scale variables. In many social science case studies, a primary focus is on assessing evidence in the data refuting the assumption of independence of Y and X conditionally on Z, with Y response variables, X predictors of interest, and Z covariates. Ideally, one would have methods available that avoid parametric assumptions, allow Y, X, Z to be random variables on arbitrary spaces with arbitrary dimension, and accommodate rapid consideration of different candidate predictors. As a formal decision-theoretic approach has clear disadvantages in this context, we instead rely on an encompassing nonparametric Bayes model for the joint distribution of Y, X and Z, with conditional mutual information used as a summary of the strength of conditional dependence. The implementation relies on a single Markov chain Monte Carlo run under the encompassing model, with conditional mutual informations for candidate models calculated as a byproduct. We provide asymptotic theory supporting the approach, and apply the method to variable selection. The methods are illustrated through simulations and criminology applications. This chapter corresponds to the technical report of Kunihama and Dunson (2014).

Chapter 4 deals with complex survey design with non-representative sample. In population studies, it is standard to sample data via designs in which the population is divided into strata, with the different strata assigned different probabilities of inclusion. Although there have been some proposals for including sample survey weights into Bayesian analyses, existing methods require complex models or ignore

the stratified design underlying the survey weights. We propose a simple approach based on modeling the distribution of the selected sample as a mixture, with the mixture weights appropriately adjusted, while accounting for uncertainty in the adjustment. We focus for simplicity on Dirichlet process mixtures but the proposed approach can be applied more broadly. We sketch a simple Markov chain Monte Carlo algorithm for computation, and assess the approach via simulations and an application of sexual behaviour analysis. This chapter comes from the paper of Kunihama et al. (2014).

Chapter 5 develops flexible joint modeling of mixed-scale longitudinal surveys. In many social science surveys, participants repeatedly respond to questions over time. Modeling and computation for multivariate longitudinal data has proven challenging, particular when data are not all continuous and Gaussian but contain discrete measurements. Also, study participants are often selected via stratified random sampling, leading to discrepancies between sample and population. To adjust for this gap, survey weights are constructed but it is not clear how to include them in hierarchical models. Motivated by an application to sexual preference data, we propose a novel nonparametric approach for mixed-scale longitudinal data in surveys. The proposed approach relies on an underlying variable mixture model, with time-varying latent factors. Bias from survey design is adjusted for in posterior computation relying on a Markov chain Monte Carlo algorithm. The approach is applied to the analysis of the sexual behaviour data from the National Longitudinal Study of Adolescent to Adult Health. This chapter is a joint project with Amy Herring, Caloryn Halpern and David Dunson.

2

Bayesian modeling of temporal dependence in large sparse contingency tables

2.1 Introduction

Time-indexed multivariate categorical data are collected in many areas, with partially-overlapping categorical variables measured for different subjects at the different time points. As a motivating application, we consider social science surveys that are conducted at regular time intervals, containing many categorical questions such as gender, race, age group, ethnicity, religious affiliation, political party and preference for particular policies. For such surveys and other types of time-indexed multivariate categorical data, it is common for the variables measured (questions asked) to vary somewhat over time while a subset of the variables will be measured at all times. In addition, the number of variables measured can be moderate to large leading to a contingency table with an *enormous* number of cells, the vast majority of which are empty. Given the fact that social science data often contain complex interactions, it becomes extremely challenging to build realistic and computationally tractable models that allow ultra-sparse data. We define ultra-sparse contingency tables as

having exponentially or super-exponentially more cells than the sample size.

Let $\mathbf{x}_{ti} = (x_{ti1}, \ldots, x_{tip})'$ denote the multivariate response for the ith subject in the survey at time t, with the jth categorical question having d_j elements, $x_{tij} \in \{1, \ldots, d_j\}, j = 1, \ldots, p$. We accommodate the case in which the specific variables measured can vary across time by introducing missingness indicators, $m_{ti} = (m_{ti1}, \ldots, m_{tip})'$, with $m_{tij} = 1$ if variable j is missing for subject i at time t; we allow design-based missingness in which certain variables are not measured for any subjects at a particular time and for individual-specific missingness in which certain individuals fail to answer all the questions posed to them. In both cases we assume missing at random.

There is a rich literature on the analysis of contingency tables (Agresti (2002); Fienberg and Rinaldo (2007)). Log linear models are perhaps the most commonly used modeling framework. Routine implementations rely on maximum likelihood estimation, though there is also a rich Bayesian literature. For large, sparse contingency tables, maximum likelihood estimates do not exist in many cases except for overly-simplistic log-linear models and richer classes of models become challenging to implement computationally. There is a rich literature on graphical modeling approaches to estimating conditional independence structures in categorical variables, with Dobra and Lenkoski (2011) proposing a recent Bayesian approach. Although their method is computationally efficient, except for very small tables, the number of possible graphical models is so enormous that it becomes infeasible to visit more than a vanishingly small fraction of the models making accurate model selection or averaging difficult. To facilitate scaling to large tables, Dunson and Xing (2009) and Bhattacharya and Dunson (2012) recently proposed Bayesian probabilistic tensor factorizations. These methods express the probability tensor corresponding to the joint probability mass function of the categorical variables as a convex combination of independent components. Such methods have not yet been developed for

time-indexed contingency tables.

There is a rich literature on categorical time series and longitudinal data analysis in which the same categorical variable is repeatedly measured for each subject over time. For example, Markov models, state space models and random effects models are routinely applied in such settings. However, these models are not relevant to the problem of incorporating dependence over time in modeling of large sparse contingency tables. As subjects can be different over time, we do not focus on the problem of incorporating within-subject dependence in repeated observations; instead our goal is to include dependence in the parameters characterizing the time-dependent joint pmfs for the categorical variables. To our knowledge, this problem has not yet been addressed in the literature. Although one can potentially adapt log-linear or graphical models developed for contingency tables at one time in a somewhat straightforward manner, the hurdles mentioned above for the static case are compounded in the dynamic setting.

To facilitate routine implementations in ultra sparse cases, we propose to adapt the Dunson and Xing (DX) (2009) probabilistic Parafac factorization to the dynamic setting. The DX model induces a tensor factorization through a Dirichlet process (DP) mixture of product multinomial distributions for the categorical observations. There is an increasingly rich literature proposing nonparametric Bayes dynamic models, which allow time-indexed dependent random probability measures. Perhaps the most common approach relies on a dependent DP (MacEachern (1999, 2000)), which incorporates time dependence in the weights and/or atoms in a stick-breaking representation (Griffin and Steel (2006); Rodriguez and Horst (2008); Chung and Dunson (2011)). Most applications of dependent DPs fix the weights and allow the atoms to vary, as varying weights can lead to computational complexities. For dynamic modeling of contingency tables, it is more parsimonious to allow varying weights and varying atoms can lead to a substantial computational burden. An alternative

approach, which allows varying weights in a computationally convenient and flexible manner, relies of dynamic mixtures of DPs (Dunson (2006); Ren et al. (2010)). Recently, a class of probit stick-breaking processes was proposed (Chung and Dunson (2009)), which has the appealing feature of allowing one to induce time dependence in random probability measures through Gaussian time series models (Rodriguez and Dunson (2011)).

We propose a new nonparametric state space model for time-indexed ultra sparse contingency tables. Relying on a DX-type probabilistic Parafac factorization, we place a dynamic model on the weights, which relies on transformed normal random variables in a similar manner to probit stick-breaking. The model is nonparametric in the sense that the induced prior for each time-indexed joint pmf assigns positive probability in arbitrarily small neighborhoods of any "true" data-generating pmf. Hence, our model can allow higher-order interactions and complex dependences, while shrinking towards a low-dimensional structure and borrowing information across time to address the curse of dimensionality. In addition, and crucially for the approach to be useful in the motivating applications, posterior computation can be implemented via a highly efficient Markov chain Monte Carlo (MCMC) algorithm relying on a slice sampler related to Kalli et al. (2011). Finally, the factorization produces a low-dimensional representation of the joint pmf, which is otherwise characterized by a daunting number of parameters in many cases, as the number of cells of the tables can be truly massive.

2.2 Model specification

2.2.1 Modeling of multivariate categorical data

We review the nonparametric Bayes approach of Dunson and Xing (2009) for a static large sparse contingency table. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ be multivariate categorical

data for the ith subject, with $x_{ij} \in \{1, \ldots, d_j\}$, $j = 1, \ldots, p$. Let

$$\boldsymbol{\pi} = \left\{\pi_{c_1 \cdots c_p},\, c_j = 1, \ldots, d_j,\, j = 1, \ldots, p\right\} \in \Pi_{d_1 \cdots d_p}$$

be a probability tensor where $\pi_{c_1 \cdots c_p} = P(x_{i1} = c_1, \ldots, x_{ip} = c_p)$ is a cell probability and $\Pi_{d_1 \cdots d_p}$ is the set of all probability tensors of size $d_1 \times \cdots \times d_p$. Dunson and Xing (2009) show that any $\boldsymbol{\pi} \in \Pi_{d_1 \cdots d_p}$ can be decomposed as

$$\boldsymbol{\pi} = \sum_{h=1}^{k} \nu_h \Psi_h, \qquad \Psi_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)} \tag{2.1}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)'$ is a probability vector, $\Psi_h \in \Pi_{d_1 \cdots d_p}$ and $\boldsymbol{\psi}_h^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)})'$ is a $d_j \times 1$ probability vector for $h = 1, \ldots, k$ and $j = 1, \ldots, p$. This expression relies on a Parafac tensor factorization (Harshman (1970) and Kolda (2001)). It follows that any multivariate categorical data distribution can be expressed as a mixture of product multinomials,

$$P(x_{i1} = c_1, \ldots, x_{ip} = c_p) = \pi_{c_1 \cdots c_p} = \sum_{h=1}^{k} \nu_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)}.$$

By introducing a latent class index $s_i \in \{1, \ldots, k\}$ for the ith subject, the multivariate responses $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ are conditionally independent given s_i. Instead of conditioning on a fixed k, Dunson and Xing (2009) developed a nonparametric Bayes approach that lets

$$\boldsymbol{\pi} = \sum_{h=1}^{\infty} \nu_h \Psi_h, \qquad \Psi_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \tag{2.2}$$

$$\boldsymbol{\psi}_h^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}), \text{ independently for } j = 1, \ldots, p, \quad h = 1, \ldots, \infty,$$

$$\nu_h = V_h \prod_{l<h} (1 - V_l),$$

$$V_h \sim \text{beta}(1, \alpha), \text{ independently for } h = 1, \ldots, \infty,$$

where $a_{jl} > 0$ for $l = 1, \ldots, d_j$ and $\alpha > 0$. Although (2) allows infinitely many components, the number k_n occupied by the n subjects in the sample will tend to be $k_n << n$, so few components will be occupied. The model corresponds to a Dirichlet process mixture of product multinomial distributions relying on a stick-breaking representation (Sethuraman (1994)). A prior is induced on the joint pmf which has large support in the sense of assigning positive probability to L_1 neighborhoods of any true joint pmf.

2.2.2 Modeling of time-indexed multivariate categorical data

Relying on the DX type probabilistic Parafac factorization, we propose a new nonparametric Bayes approach for time-indexed large sparse contingency tables. In a dynamic setting, we obtain the time-indexed multivariate response $\mathbf{x}_{ti} = (x_{ti1}, \ldots, x_{tip})'$, $x_{tij} \in \{1, \ldots, d_j\}$, for the ith subject at time t for $i = 1, \ldots, n_t$, $t = 1, \ldots, T$ and $j = 1, \ldots, p$. At time t we have a probability tensor π_t for the multivariate categorical response given by

$$\boldsymbol{\pi}_t = \left\{\pi_{tc_1 \cdots c_p},\, c_j = 1, \ldots, d_j,\, j = 1, \ldots, p\right\} \in \Pi_{d_1 \cdots d_p}$$

where $\pi_{tc_1 \cdots c_p} = P(x_{ti1} = c_1, \ldots, x_{tip} = c_p)$ is a cell probability at time t. Relying on the probabilistic Parafac factorization, each probability tensor $\boldsymbol{\pi}_t$ can be expressed as a mixture of product multinomials

$$\boldsymbol{\pi}_t = \sum_{h=1}^{k_t} \nu_{th} \Psi_{th}, \qquad \Psi_{th} = \boldsymbol{\psi}_{th}^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_{th}^{(p)} \tag{2.3}$$

where $k_t \in \mathbb{N}$, $\boldsymbol{\nu}_t = (\nu_{t1}, \ldots, \nu_{tk_t})'$ is a probability vector, $\Psi_{th} \in \Pi_{d_1 \cdots d_p}$ and $\boldsymbol{\psi}_{th}^{(j)} = (\psi_{th1}^{(j)}, \ldots, \psi_{thd_j}^{(j)})'$ is a $d_j \times 1$ probability vector for $h = 1, \ldots, k_t$. Letting $s_{ti} \in \{1, \ldots, k_t\}$ denote a latent class index for the ith subject at time t, the observations $\mathbf{x}_{ti}$ are conditionally independent given s_{ti}.

To borrow information across time, we place a dynamic structure on the probability tensor $\boldsymbol{\pi}_t$ in (2.3) assuming time varying weights ν_{th} and static atoms $\psi_{th}^{(j)} = \psi_h^{(j)}$.

Time dependence is induced in the weights through a state space model, which assumes that stick-breaking increments on ν_{th} arise through transforming Gaussian autoregressive processes using a monotone differentiable link function $g : \Re \to (0, 1)$. This characterization is motivated by the probit stick-breaking process (Chung and Dunson (2009); Rodriguez and Dunson (2011)), and leads to a parsimonious but flexible characterization of time-dependence in joint pmfs underlying large, sparse contingence tables.

Similarly to expression (2), we develop a nonparametric Bayes approach that sets the number of components to $k_t = \infty$, though the number of occupied components will tend to be much less than the sample size and can vary across time. The specific model is

$$\boldsymbol{\pi}_t = \sum_{h=1}^{\infty} \nu_{th} \Psi_h, \qquad \Psi_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \tag{2.4}$$

$$\boldsymbol{\psi}_h^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}), \text{ independently for } j = 1, \ldots, p, \quad h = 1, \ldots, \infty, \tag{2.5}$$

$$\nu_{th} = g(W_{th}) \prod_{l<h} \{1 - g(W_{tl})\}, \tag{2.6}$$

$$W_{th} = \alpha_{th} + \varepsilon_{th}, \quad \varepsilon_{th} \sim N(0, \sigma_\varepsilon^2), \tag{2.7}$$

$$\alpha_{th} = \mu + \phi \alpha_{t-1h} + \eta_{th}, \quad \eta_{th} \sim N(0, \sigma_\eta^2), \tag{2.8}$$

where $|\phi| < 1$, $\{\varepsilon_{th}\}$ and $\{\eta_{th}\}$ are sequences of independently normally distributed random variables with mean 0 and variance σ_ε^2 and σ_η^2 respectively. The parameter ϕ controls the autocorrelation in the weights ν_{th} on the different components over time. For sake of parsimony and simplicity in modeling and computation, we include a single time-stationary correlation parameter ϕ instead of allowing dependence to be time or element specific. In the limiting case in which $\phi = 0$, the weights ν_{th} will be modeled as independent. This does not mean that independent priors are placed

on the unknown joint pmfs at each time, as the incorporation of common atoms automatically induces some degree of *a priori* dependence. However, in applications one typically expects that the joint pmfs will be quite similar over time, and by using varying weights one does not rule out arbitrarily large changes in the pmfs over time. When ϕ is close to one, there will be very high time dependence in the weights, leading to effective collapsing on a model that assumes a single time stationary joint pmf. For the initial state variables, we assume the stationary distributions, $\alpha_{1h} \sim N(\mu/(1-\phi), \sigma_\eta^2/(1-\phi^2))$ independently for $h = 1, \ldots, \infty$. Also, we choose priors $\mu \sim N(\mu_0, \sigma_0^2)$, $\phi \sim U(-1, 1)$, $\sigma_\varepsilon^2 \sim IG(m_\varepsilon/2, S_\varepsilon/2)$ and $\sigma_\eta^2 \sim IG(m_\eta/2, S_\eta/2)$ respectively.

Due to the Parafac factorization leading to a massive reduction in the number of parameters, the proposed method can efficiently estimate all the cell probabilities using cells with both positive and zero observed counts; the cells having zero counts can vary over time and are not assumed to be structural zeros. The marginal posterior distributions for the cell probabilities will not be concentrated at zero even if the observed counts are zero.

Expressions (2.4)-(2.8) induce a prior on the time-dependent joint pmfs, but it is not immediately obvious how the chosen hyperpriors in the hierarchical specification impact the properties of the prior for $\{\boldsymbol{\pi}_t\}$. In particular, it is important to obtain characterizations of the moments of the induced prior for the cell probabilities, as well as the prior covariance between different elements and across time. Such expressions are provided in Lemma 1, with the proof in Appendix A. Lemma 2 shows that the prior is well defined in the sense that $\sum_{h=1}^{\infty} \nu_{th}$ converges to one almost surely.

Lemma 1. *The expectation, variance and covariance of the joint prior on the ele-*

ments of $\{\boldsymbol{\pi}_t\}$ *induced through (2.4)-(2.8) are*

$$E\{\pi_{tc_1\cdots c_p}\} = \prod_{j=1}^{p} \frac{a_{jc_j}}{\hat{a}_j}, \quad V\{\pi_{tc_1\cdots c_p}\} = \left(\prod_{j=1}^{p} \frac{a_{jc_j}(a_{jc_j}+1)}{\hat{a}_j(\hat{a}_j+1)} - \prod_{j=1}^{p} \frac{a_{jc_j}^2}{\hat{a}_j^2} \right) \left(\frac{\beta_2}{2\beta_1 - \beta_2} \right),$$

$$Cov\{\pi_{tc_1\cdots c_p}, \pi_{t+kc_1'\cdots c_p'}\} = \left(\prod_{j=1}^{p} \frac{a_{jc_j}\{a_{jc_j'} + 1(c_j = c_j')\}}{\hat{a}_j(\hat{a}_j+1)} - \prod_{j=1}^{p} \frac{a_{jc_j}a_{jc_j'}}{\hat{a}_j^2} \right) \left(\frac{\gamma_k}{2\beta_1 - \gamma_k} \right),$$

where $\beta_1 = E\{g(W_{th})\}$, $\beta_2 = E\{g^2(W_{th})\}$, $\gamma_k = E\{g(W_{th})g(W_{t+kh})\}$, $\hat{a}_j = \sum_{l=1}^{d_j} a_{jl}$ *and* $1(\cdot)$ *is an indicator function.*

The expectation of cell probabilities can be expressed as the product of expectations of Dirichlet priors for atoms. The variance and covariance are expressed as the product of two terms, the first one is related to atoms and the second one comes from time varying weights. As $\mu \to \infty$, then $\beta_2/(2\beta_1 - \beta_2) \to 1$ and $\gamma_k/(2\beta_1 - \gamma_k) \to 1$, and the variance and covariance will be influenced only by atoms. In such a case, the measure corresponding to the stick-breaking process will become a point mass at a random atom almost surely. In addition, β_1, β_2 and γ_k do not depend on time t, hence all expectation, variance and covariance are independent of t though the covariance depends on the time difference k. Also, the covariance between cell probabilities with $c_j = c_j'$ for all j is always positive and, on the other hand, those with $c_j \neq c_j'$ for all j have negative covariance. In a special case in which the hyperparameters in the Dirichlet prior are $a_{j1} = \cdots = a_{jd_j} = a$ the variance and covariance is zero in the limit as $a \to \infty$.

Lemma 2. $\sum_{h=1}^{\infty} \nu_{th} = 1$ *almost surely.*

Lemma 2 is important in showing that the prior is well defined. The proof is in Appendix A.

Our proposed prior setting is parsimonious but highly flexible in the sense that the induced prior assigns positive probability in arbitrarily small neighborhoods of

any true data-generating pmf. Let Π denote the space having elements of the form $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t \in \Pi_{d_1 \cdots d_p},\, t \in \{1, \ldots, T\}\}$. We show in Theorem 3 that the proposed prior has large support on Π.

Theorem 3. *Let $\mathcal{Q}$ denote the prior on Π through the proposed model and $\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$ denote an L_1 neighborhood around an arbitrary $\boldsymbol{\pi}^0 \in \Pi$. Then for any $\boldsymbol{\pi}^0 \in \Pi$ and $\epsilon > 0$, the prior assigns positive probability in the ϵ-neighborhood, $\mathcal{Q}\{\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)\} > 0$.*

Since the proposed prior is defined on a space with finitely many components, a straightforward extension of theorem 4.3.1 in Ghosh and Ramamoorthi (2003) ensures that the posterior concentrates in arbitrary small neighborhoods of any true data-generating distribution as the sample size increases.

2.2.3 Interpretability, prior elicitation and structural zeros

This subsection discusses how to interpret interactions, induce prior information and accommodate structural zero conditions, relying on different expressions of cell probabilities.

In categorical data analysis, detecting interactions of various order is one of the main interests. To interpret our model, we propose a novel approach where cell probabilities are expressed relying on generalized linear modeling. Let x_t^A, x_t^B, x_t^C, x_t^D be categorical variables with $x_t^* \in \{1, \ldots, d_*\}$ where $*$ is one of $\{A, B, C, D\}$. We express the cell probability as

$$P(x_t^A = a, x_t^B = b, x_t^C = c, x_t^D = d) = \pi_{tabcd} = \frac{\mu_{tabcd}}{\sum_{\hat{a}=1}^{d_A} \sum_{\hat{b}=1}^{d_B} \sum_{\hat{c}=1}^{d_C} \sum_{\hat{d}=1}^{d_D} \mu_{t\hat{a}\hat{b}\hat{c}\hat{d}}}, \quad (2.9)$$

where μ_{tabcd} is defined through its logarithm form as

$$\log \mu_{tabcd} = \lambda_t + \lambda_{ta}^A + \lambda_{tb}^B + \lambda_{tc}^C + \lambda_{td}^D \quad (2.10)$$

$$+ \lambda_{tab}^{AB} + \lambda_{tac}^{AC} + \lambda_{tad}^{AD} + \lambda_{tbc}^{BC} + \lambda_{tbd}^{BD} + \lambda_{tcd}^{CD} \quad (2.11)$$

$$+ \lambda_{tabc}^{ABC} + \lambda_{tabd}^{ABD} + \lambda_{tacd}^{ACD} + \lambda_{tbcd}^{BCD} + \lambda_{tabcd}^{ABCD}. \quad (2.12)$$

This model corresponds to the multinomial model for contingency tables (Agresti (2002)). The right term in (2.10) represents constant term and main effects, (2.11) shows 2-way interactions and (2.12) corresponds to 3-way and 4-way interactions. For identification, we assume $\mu_{td_Ad_Bd_Cd_D} = 1$ and main effects and interactions are zero if at least one of $a = d_A$, $b = d_B$, $c = d_C$ and $d = d_D$ is satisfied. The constraints imply $\lambda_t = 0$. Given the cell probabilities, the interactions can be deterministically computed. This approach can be applied to a general case since the proposed prior has Kolmogorov consistency properties so that if one starts with a prior for P variables the same marginal is obtained for any subset of variables as if one started with a prior for the joint pmf of the subset.

In addition, we propose an approach for inducing an informative prior on time varying main effects and interactions. We start with eliciting a prior sample size n as well as an informative prior $\pi^*(\theta^*)$ for the parameters θ^* in a parametric model, such as (2.9)-(2.12). To update our initial default prior to include this information, we follow a data augmentation approach in which we add data y^n generated from the prior predictive under the parametric model to our observed data, with y^n structured to have the same variables and sample size n at each time. That is, $y^n = \{\boldsymbol{y}_{ti}, t = 1, \ldots, T, i = 1, \ldots, n\}$ where $\boldsymbol{y}_{ti}$ is a vector of categorical variables with the same structure as $\mathbf{x}_{ti}$ in section 2.2. To marginalize over the prior predictive in inducing an informative prior for the parameters in our nonparametric model, we generate a new value of y^n at each MCMC iteration.

In many contingency table applications, certain combinations of categories are known to have probability zero *a priori*. For example, males cannot get pregnant. There are two ways in which structural zeros can be easily accommodated by the proposed method. Manrique-Vallier and Reiter (2012) recently proposed a Bayesian approach for latent structure models with structural zeros. Under their approach, the observed cell counts equal latent cell counts multiplied by an indicator function,

which gives value zero for structural zeros. Our model could be used for the latent cell counts, with the missing data in the cells with structural zeros imputed within an MCMC sampler.

Another approach is to combine variables which include combinations of structural zeros. For example, consider a survey which contains indicator variables of gender and pregnancy where $x_{ti1} = 1$ if subject i at time t is female and $x_{ti2} = 1$ if pregnant, leading to a 2×2 sub-table with a structural zero. Then, we define a new variable $\tilde{x}_{ti}$ by vectorizing all cell components in the table except those that are impossible to occur. In this example, $\tilde{x}_{ti}$ has 3 categories such that $\tilde{x}_{ti} = 1$ corresponds to $(x_{ti1}, x_{ti2}) = (0, 0)$, $\tilde{x}_{ti} = 2$ to $(x_{ti1}, x_{ti2}) = (1, 0)$ and $\tilde{x}_{ti} = 3$ to $(x_{ti1}, x_{ti2}) = (1, 1)$. Replacing x_{ti1} and x_{ti2} by $\tilde{x}_{ti}$, the corresponding contingency tables have no structural zeros, and the proposed method can be applied. The cell probabilities of x_{ti1} and x_{ti2} can be computed from those of $\tilde{x}_{ti}$, so that inferences are based on relationships among the original variables.

2.3 MCMC algorithm for posterior computation

For posterior computation in DP mixtures, one common approach is marginalizing out the random probability measure with the Polya urn scheme (Bush and MacEachern (1996)). Avoiding marginalization, Ishwaran and James (2001) developed the blocked Gibbs sampler relying on truncation approximation of the stick-breaking representation. Without truncation, Walker (2007) and Papaspiliopoulos and Roberts (2008) proposed the slice sampler and retrospective MCMC methods respectively. Though the slice sampler is simpler to implement, conditional constraints on sticks can cause slow mixing of the chain. Kalli et al. (2011) proposed a more efficient slice sampler avoiding such a mixing problem.

Relying on a slice sampler related to Kalli et al. (2011), we developed a simple and efficient MCMC algorithm for the proposed model. In the motivating application,

we have two types of missing data, design-based missingness and individual-specific missingness. We assume missing at random for both cases and handle the missing data using missingness indicators, $m_{ti} = (m_{ti1}, \ldots, m_{tip})'$, with $m_{tij} = 1$ if variable j is missing for subject i at time t. In addition, we introduce latent variables $u_t = (u_{t1}, \ldots, u_{tn_t})'$ for the slice sampler. The likelihood of $\{u_t\}$ and $\{\mathbf{x}_t\}$ given $\{m_{ti}\}$, $\{\boldsymbol{\nu}_t\}$ and $\left\{\boldsymbol{\psi}_h^{(j)}\right\}$ can be expressed as

$$\prod_{t=1}^{T}\prod_{i=1}^{n_t}\left\{\sum_{h=1}^{\infty} 1(u_{ti} < \nu_{th}) \prod_{j:\, m_{tij}=0}\prod_{l=1}^{d_j}\left(\psi_{hl}^{(j)}\right)^{1(x_{tij}=l)}\right\}.$$

This representation is consistent with the original model setting if latent variables $\{u_t\}$ are marginalized out. In a special case in which g is a probit link function, the data augmentation approach in Albert and Chib (2001) can improve efficiency of the posterior sampling by introducing independent normal latent variables $\{z_{tih}\}$ with mean W_{th} and variance 1 satisfying

$$P(z_{tih} > 0,\, z_{til} \le 0,\, l < h) = \Phi(W_{th}) \prod_{l<h}\{1 - \Phi(W_{th})\} = \nu_{th} = P(s_{ti} = h).$$

We propose the following MCMC sampling steps:

1. For $h = 1, \ldots, k^*$, with $k^* = \max\{s_{ti}\}$, update $\boldsymbol{\psi}_h^{(j)}$ from the following Dirichlet full conditional posterior distribution,

$$\text{Dirichlet}\left(a_{j1} + \sum_{(t,i)\in A_{jh}} 1(x_{tij} = 1),\, \ldots,\, a_{jd_j} + \sum_{(t,i)\in A_{jh}} 1(x_{tij} = d_j)\right).$$

   where $A_{jh} = \{(t,i) : m_{tij} = 0,\, s_{ti} = h\}$.

2. Update z_{tih} from the marginal (w.r.t. u_{ti}) conditional posterior distribution,

$$z_{tih} \mid \cdots \sim \begin{cases} N_-(W_{th}, 1) & h < s_{ti}, \\ N_+(W_{th}, 1) & h = s_{ti}, \end{cases}$$

where $N_-(W_{th}, 1)$ and $N_+(W_{th}, 1)$ denote the normal distributions with mean W_{th} and variance 1 truncated on $(-\infty, 0]$ and $(0, \infty)$ respectively.

3. Update W_{th} from the normal marginal (w.r.t. u_{ti}) conditional posterior distribution, $N(\hat{W}_{th}, \sigma^2_{W_{th}})$ where

$$\hat{W}_{th} = \sigma^2_{W_{th}} \left(\sum_{i:s_{ti} \geq h}^{n_t} z_{tih} + \sigma_\varepsilon^{-2} \alpha_{th} \right), \quad \sigma^2_{W_{th}} = \frac{1}{\sum_{i=1}^{n_t} 1(s_{ti} \geq h) + \sigma_\varepsilon^{-2}}.$$

4. Update u_{ti} from the full conditional distribution, Uniform$(0, \nu_{ts_{ti}})$.

5. Update s_{ti} from the multinomial full conditional distribution,

$$Pr(s_{ti} = h \,|\, \cdots) = \frac{1(h \in B_{ti}) \prod_{j:m_{tij}=0} \psi^{(j)}_{hx_{tij}}}{\sum_{l \in B_{ti}} \prod_{j:m_{tij}=0} \psi^{(j)}_{lx_{tij}}},$$

where $B_{ti} = \{h : \nu_{th} > u_{ti}\}$. To identify the elements in $\{B_{ti}\}$, we first update α_{th} and W_{th} for $t = 1, \ldots, T$ and $h = 1, \ldots, \tilde{k}$ where $\tilde{k}$ is the smallest number with $\sum_{h=1}^{\tilde{k}} \nu_{th} > 1 - \min\{s_{ti}\}$ for all t.

6. For $h = 1, \ldots, k^*$, update α_{th} using the forward filtering backward sampling algorithm by Früwirth-Schnatter (1994) and Carter and Kohn (1994), or Kalman filter and the simulation smoother by de Jong and Shephard (1995) and Durbin and Koopman (2002).

7. Update μ from the conditional posterior, $N(\mu_*, \sigma^2_\mu)$ where $\mu_* = \sigma^2_\mu(\hat{\sigma}^{-2}\hat{\mu} + \sigma_0^{-2}\mu_0)$, $\sigma^2_\mu = (\hat{\sigma}^{-2} + \sigma_0^{-2})^{-1}$ and

$$\hat{\mu} = \frac{\sum_{h=1}^{k^*} \sum_{t=2}^{T} (\alpha_{th} - \phi \alpha_{t-1h}) + (1+\phi) \sum_{h=1}^{k^*} \alpha_{1h}}{k^* \{T - 1 + (1+\phi)/(1-\phi)\}},$$

$$\hat{\sigma}^2 = \frac{\sigma^2_\eta}{k^* \{T - 1 + (1+\phi)/(1-\phi)\}}.$$

8. Update ϕ using the independence MH algorithm in which the proposal distribution is constructed relying on the mode and Hessian of the logarithm of the conditional posterior densities $\pi(\phi|\cdots)$. First, we compute $\hat{\phi}$ which maximizes (or approximately maximizes) the conditional posterior density. Then, we generated a candidate from a truncated normal distribution $TN_{(-1,1)}(\phi_*, \sigma^2_\phi)$, where

$$\phi_* = \hat{\phi} + \sigma^2_\phi \left. \frac{\partial \log \pi(\phi|\cdots)}{\partial \phi} \right|_{\phi=\hat{\phi}}, \quad \sigma^2_\phi = \left\{ - \left. \frac{\partial \log \pi(\phi|\cdots)}{\partial^2 \phi} \right|_{\phi=\hat{\phi}} \right\}^{-1}.$$

9. Update σ^2_ε from the conditional distribution, $IG(\hat{m}_\varepsilon/2, \hat{S}_\varepsilon/2)$ where $\hat{m}_\varepsilon = Tk^* + m_\varepsilon$ and $\hat{S}_\varepsilon = \sum_{t=1}^{T} \sum_{h=1}^{k^*} (W_{th} - \alpha_{th})^2 + S_\varepsilon$.

10. Update σ^2_η from the conditional distribution, $IG(\hat{m}_\eta/2, \hat{S}_\eta/2)$ where $\hat{m}_\eta = Tk^* + m_\eta$ and $\hat{S}_\eta = \sum_{h=1}^{k^*} \sum_{t=2}^{T} (\alpha_{th} - \mu - \phi\alpha_{t-1h})^2 + (1-\phi^2) \sum_{h=1}^{k^*} \{\alpha_{1h} - \mu/(1-\phi)\}^2 + S_\eta$.

If g is an arbitrary link function, we update W_{th} using the independent MH algorithm, instead of step 2 and 3 above. We generate a candidate from a normal distribution relying on the mode and Hessian of the logarithm of the conditional posterior densities of W_{th}. For logistic and t-links we can instead change the variance of z_{tih} from one to an observation-specific random variable having an appropriate mixture distribution; by treating the mixture distribution as unknown using a Dirichlet process, one can estimate the link nonparametrically.

2.4 Simulation study

In this section, we assess the impact of borrowing of information over time by comparing our proposed method to static approaches, such as Dunson and Xing (DX) (2009). The static models are applied independently at each time point with no time-dependence included. First, we simulate time-indexed contingency tables from the

model shown in expressions (2.4)-(2.8) with $T = 10$, $P = 20$, $d_j = 4$ for all j, $\mu = 0$, $\phi = 0.8$, $\sigma_\varepsilon = 0.1$ and $\sigma_\eta = 0.8$. At the respective time points we generated 120, 110, 150, 80, 100, 120, 100, 140, 110 and 150 observations, tiny sample sizes compared with the number of cells. For prior distributions, we assumed $\boldsymbol{\psi}_h^{(j)} \sim \text{Dirichlet}(1, \ldots, 1)$, $\mu \sim N(0, 1)$, $\phi \sim U(-1, 1)$, $\sigma_\varepsilon^2 \sim IG(2.5, 0.025)$, $\sigma_\eta^2 \sim IG(2.5, 0.025)$. We draw 60,000 MCMC samples after the initial 20,000 samples are discarded as a burn-in period and every fifth sample is saved. We observed that the sample paths were stable and the sample autocorrelations dropped smoothly. Therefore, the chains apparently converged and mixed rapidly.

We first assess performance in estimation of cell probabilities. We randomly picked several cell probabilities and tracked their movements over time. We report true values, posterior means and 95% credible intervals in Figure 2.1 (the proposed method) and Figure 2.2 (DX method). The proposed approach covers all true values in 95% intervals and interval widths are much narrower than for the DX approach consistently across time.

We additionally investigate performance in estimating associations among the categorical variables using the following measure of dependence from Dunson and Xing (2009)

$$\rho_{tjj'}^2 = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{\left(\pi_{tc_jc_{j'}} - \bar{\psi}_{tc_j}^{(j)} \bar{\psi}_{tc_{j'}}^{(j')}\right)^2}{\bar{\psi}_{tc_j}^{(j)} \bar{\psi}_{tc_{j'}}^{(j')}}, \tag{2.13}$$

where $\bar{\psi}_{tl}^{(j)} \equiv P(x_{tij} = l) \approx \sum_{h=1}^{k^*} \nu_{th} \psi_{hl}^{(j)}$. The first row of Figure 2.3 reports plots of all pairs of true values (y-axis) and posterior means (x-axis) of $\rho_{tjj'}$ at time $t = 2$ and 7. Since all cell probabilities are given in the simulation study, the true $\rho_{tjj'}$ can be computed through (2.13). At each time point, coordinate points by our approach locate closely to the $y = x$ line, compared to widely scattered points by the DX method. In addition, Table 2.1 shows correlations between true values and posterior

means of $\rho_{tjj'}$. Although correlations by the DX method are high, the proposed method consistently produces higher correlations.

Log linear models provide a standard choice for the analysis of contingency tables. However, one issue is that flexible log-linear models that accommodate arbitrary interactions among the variables and allow time dependence cannot be applied directly to large, sparse tables. Certainly, maximum likelihood estimates typically do not exist and Bayesian methods that allow an unknown dependence structure do not scale beyond small tables. Dahinden et al. (2010) proposed an approach for high-dimensional log-linear models with interactions, which relies on solving several low-dimensional subproblems that are then combined. An earlier approach by Dahinden et al. (2007) instead relied on L1 penalized log-linear models allowing sparsity of tables. Also, Dahinden et al. (2007) proposed an efficient estimation algorithm for model selection for two level categorical variables.

As a second alternative to our proposed approach, we implemented the method of Dahinden et al. (DH) (2007) in a second simulation example with $T = 8$, $P = 13$ and $d_j = 2$ for all j. Other settings are the same as in the first simulation case. As DH did not consider time-indexed contingency tables, we applied their approach separately at each time point using the logilasso R package, with 5-way cross validation used to choose penalty parameters. The second row of Figure 2.3 and Table 2 summarize the resulting dependence measures $\rho_{tjj'}$ at time $t = 2$ and 7 for each method. For the proposed method, the posterior means are close to true values and correlations between estimates and true values are uniformly high. The DH method has a tendency to underestimate dependence, particularly when true values are low, and has the lowest correlation between the estimates and truth.

To gauge robustness we also simulated data from a time-dependent log-linear model in which all main effects and two-way interactions independently follow random walk processes, $\xi_t \sim N(\xi_{t-1}, 1)$ with $\xi_0 = 0$ where ξ_t is a main effect or 2-way

interaction at time t, and other higher interactions are zero. The third row of Figure 2.3 and Table 2.3 report the estimation results. Although we find less difference among them in this case, the proposed method still shows the best performance.

In addition, for the interactions discussed in subsection 2.3, we generated contingency tables from the model (2.9)-(2.12) with $N = 1,200$, $T = 30$, $P = 4$ and $d_j = 2$ for all j. For the first 10 time points, we assumed two-way interactions (x_t^A, x_t^B), (x_t^A, x_t^C), (x_t^B, x_t^C), (x_t^B, x_t^D), (x_t^C, x_t^D) where we express the structure of these interactions as

$$M_1 = \{AB, AC, BC, BD, CD\}.$$

For the next 10 time points, a three-way interaction involving variables x_t^A, x_t^B and x_t^C also occurs:

$$M_2 = M_1 \cap \{ABC\}.$$

For the last 10 time points, another three-way interaction involving variables x_t^B, x_t^C and x_t^D occurs:

$$M_3 = M_2 \cap \{BCD\}$$

We assumed no 4-way interaction for simplicity. Using the same choice of prior as in our other analyses, we generated $20,000$ MCMC samples after a $10,000$ burn-in and every fifth sample was saved. Figure 2.4 reports the estimation results of interactions. The proposed method clearly has good performance in estimating the interactions and detecting the time changes, with 95% credible intervals covering the true parameters 97.5% of the time for the constant interactions and 96.6% for the time-varying interactions.

2.5 Analysis of social survey data

In this section, we apply the proposed method to data from the General Social Survey (GSS, http://www3.norc.org/GSS+Website). Our focus is on studying associations

among demographic and preference variables over time. We select $p = 29$ categorical variables from 1994 to 2010, including gender, ethnicity, preference for particular policies and many more listed in the supplemental materials. The GSS was conducted every two years across this time period. The numbers of observations are 2,992 (1994), 2,904 (1996), 2,832 (1998), 2,817 (2000), 2,765 (2002), 2,812 (2004), 4,510 (2006), 2,023 (2008) and 2,044 (2010) respectively. There are abundant missing data in which only a subset of the variables were recorded for an individual, and compared to the number of cells, the sample size is quite small at each time point.

We first compared our proposed approach to log-linear models. Unfortunately, current methodology for fitting log-linear models that allow flexible dependence structures cannot accommodate these data due to the large sparse structure, time variation and abundant missing data. Hence, in order to provide a comparison, we initially focused on a bivariate subset of the data consisting of religious preference ($i = 1, \ldots, 5$) and attitude towards abortion ($j = 1, 2$) from 1994 to 2010. We consider the following Log-Linear Poisson (LLP) models.

$$\text{LLP-Model 1:} \qquad N_{tij} \sim \text{Poisson}(N_t\, \mu_{ij}), \quad \log \mu_{ij} = \lambda + \lambda_i^R + \lambda_j^A + \lambda_{ij}^{RA},$$

where N_{tij} is count of the cell ij at time t, $N_t = \sum_i \sum_j N_{tij}$, λ_i^R is an effect of the first variable (religious preference), λ_j^A is an effect of the second variable (view of abortion) and λ_{ij}^{RA} is an association term. For identifiability, we assume constraints $\lambda_5^R = \lambda_2^A = \lambda_{5j}^{RA} = \lambda_{i2}^{RA} = 0$. LLP-Model 1 is a standard choice in the contingency table literature (Agresti (2002)), and the cell probabilities $\mu_{ij} / \sum_{i'} \sum_{j'} \mu_{i'j'}$ do not depend on time. Next, we extended LLP-Model 1 to incorporate time-varying effects on cell probabilities.

$$\begin{aligned}
\text{LLP-Model 2:} \qquad & N_{tij} \sim \text{Poisson}(N_t\, \mu_{tij}), \quad \log \mu_{tij} = \lambda_t + \lambda_{ti}^R + \lambda_{tj}^A + \lambda_{tij}^{RA}, \\
& \boldsymbol{\beta}_t = (\lambda_t, \lambda_{t1}^R, \ldots, \lambda_{t4}^R, \lambda_{t1}^A, \lambda_{t11}^{RA}, \ldots, \lambda_{t41}^{RA})', \\
& \beta_{tl} = \mu_l + \phi_l \beta_{t-1l} + \varepsilon_{tl}, \quad \varepsilon_{tl} \sim N(0, \sigma_l^2), \quad l = 1, \ldots, 10,
\end{aligned}$$

where λ_{ti}^{R}, λ_{tj}^{A} and λ_{tij}^{RA} are effects of the first variable, the second variable and interactions at time t respectively. We assume $\lambda_{t5}^{R} = \lambda_{t2}^{A} = \lambda_{t5j}^{RA} = \lambda_{ti2}^{RA} = 0$ at each time point and $\boldsymbol{\beta}_0 = \mathbf{0}$ for the initial values. LLP-Model 2 is a time dependent hierarchical model in which all parameters in the log-linear model follow first order autoregressive processes independently.

We first estimate all models using the data from 1994 to 2008. Then, relying on the estimated parameters, we predict the contingency table in 2010 (Table 2.4). For the proposed model, we used the same MCMC settings as in the simulation study. For log-linear models, we estimated parameters using an MCMC algorithm where missing values are imputed from conditional probabilities given observed data at each iteration. For example, we generate the religious preference i given the view of abortion j with probability $\mu_{tij}/\sum_{i'}\mu_{ti'j}$. For priors, we assumed $\boldsymbol{\beta} = (\lambda, \lambda_1^R, \ldots, \lambda_4^R, \lambda_1^A, \lambda_{11}^{RA}, \ldots, \lambda_{41}^{RA})' \sim N(\mathbf{0}, I)$ for LLP-Model 1, $\mu_l \sim N(0,1)$, $\phi_l \sim U(-1,1)$ and $\sigma_l^2 \sim IG(2.5, 0.025)$ for all l for LLP-Model 2. Using Gibbs sampling, we generated posterior samples of μ_l and σ_l^2 from normal and Inverse-Gamma distributions respectively. For $\boldsymbol{\beta}$, ϕ_l, $\boldsymbol{\beta}_t$, we used a MH algorithm in which candidates were generated from normal distributions relying on the mode and Hessian of the logarithm of the conditional posterior densities. We generated 10,000 MCMC samples after a 1,000 burn-in for LLP-Model 1 and 20,000 MCMC samples after the 2,000 burn-in for LLP-Model 2 and, for both cases, every fifth sample was saved.

We generated replications at every fifth MCMC iteration and computed averages of the following predictive criteria,

$$\text{Absolute deviation (AD):} \quad \sum_{i=1}^{5}\sum_{j=1}^{2}\left|N_{ij}^{rep} - N_{ij}^{obs}\right|,$$

$$\text{Mean absolute percentage error (MAPE):} \quad \frac{1}{10}\sum_{i=1}^{5}\sum_{j=1}^{2}\left|\frac{N_{ij}^{rep} - N_{ij}^{obs}}{N_{ij}^{obs}}\right|,$$

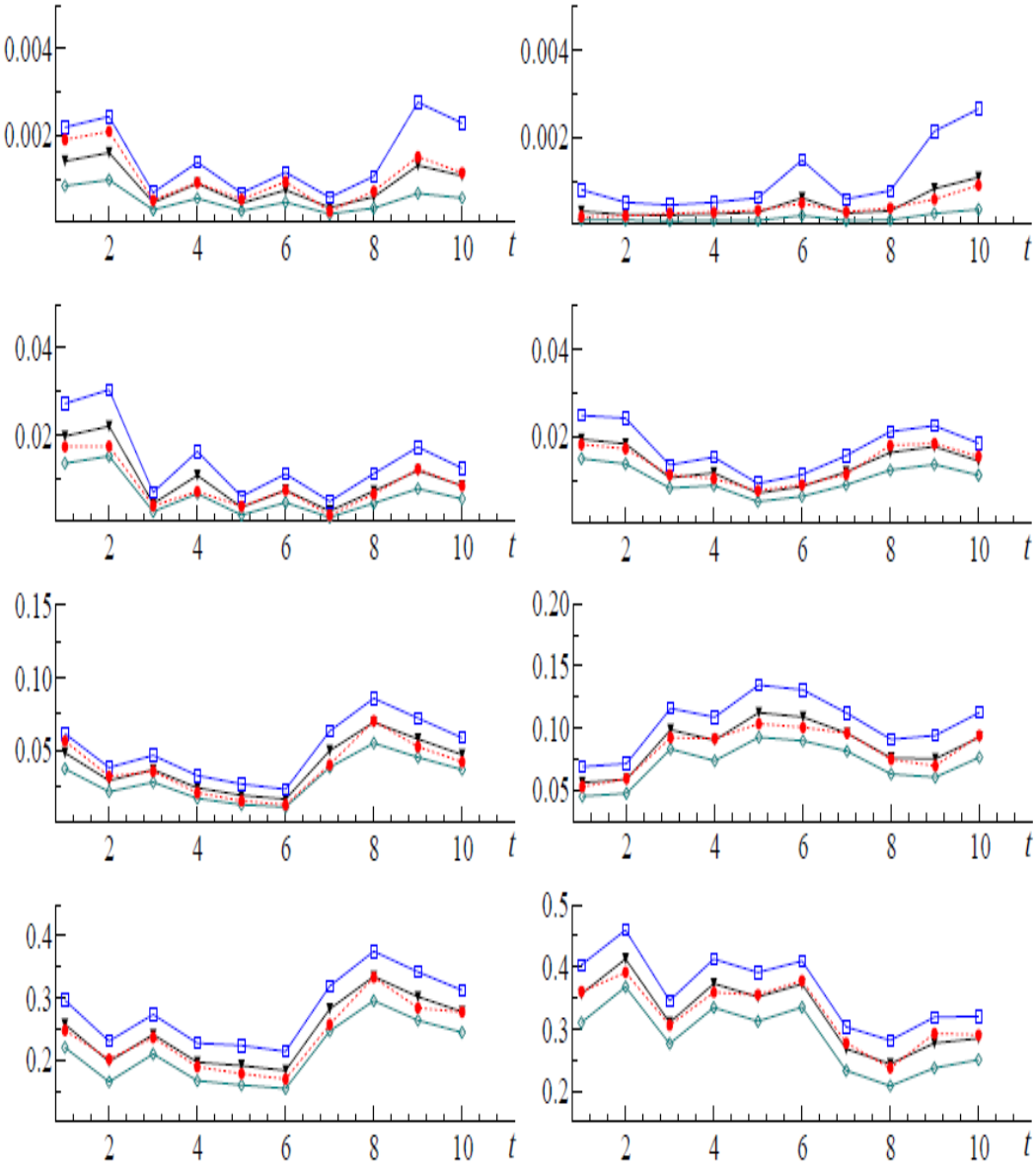
where N_{ij}^{rep} and N_{ij}^{obs} are the replication and observation of count of the cell ij respectively. To keep the same total number of replications among all methods, predictions are generated from cell probabilities $\mu_{ij}/\sum_{i'}\sum_{j'}\mu_{i'j'}$ for LLP-Model 1 and $\mu_{2010ij}/\sum_{i'}\sum_{j'}\mu_{2010i'j'}$ for LLP-Model 2. Table 2.5 reports the prediction results. Although LLP-Model 2 produces better performance than LLP-Model 1 by incorporating time-dependence, our proposed method clearly outperforms log-linear models in terms of both predictive criteria. In addition, we compared 95% predictive intervals by the proposed method and LLP-Model 2 and found that our method could capture all actual observations, while LLP-Model 2 had poor coverage. The plot is included in the supplemental materials.

Next, we apply the proposed method to all 29 categorical variables. We generated 30,000 MCMC samples after the initial 10,000 samples are discarded as the burn-in and every fifth sample are saved. We observed the sample paths are stable and the sample autocorrelations are small. Table 2.6 shows the estimation result of parameters in the time dependent stick-breaking processes. Concerning the measure of time dependence ϕ, the posterior mean is close to 1 and the 95% credible interval locates near 1, which means the weights of the stick-breaking processes have strong time dependence over time.

Then, we investigate cross interactions among the variables over time. Figure 2.5 show the posterior means of $\rho_{tjj'}$ for all pairs in 2002 and 2010. We find the structure of interactions is complex at each time point. Also, though each interaction gradually changes over time, all tables look similar to one another, implying they have close dependence. This is consistent with the result of the strong dependent weights in the stick-breaking processes. Some categorical variables such as Race [$j = 3$], Attitude toward abortion [6], Political party affiliation [9] and Think of self as liberal or conservative [14] intricately correlate with many other variables. On the other hand, zodiac [11] shows little interactions with all other variables. Among all pairs

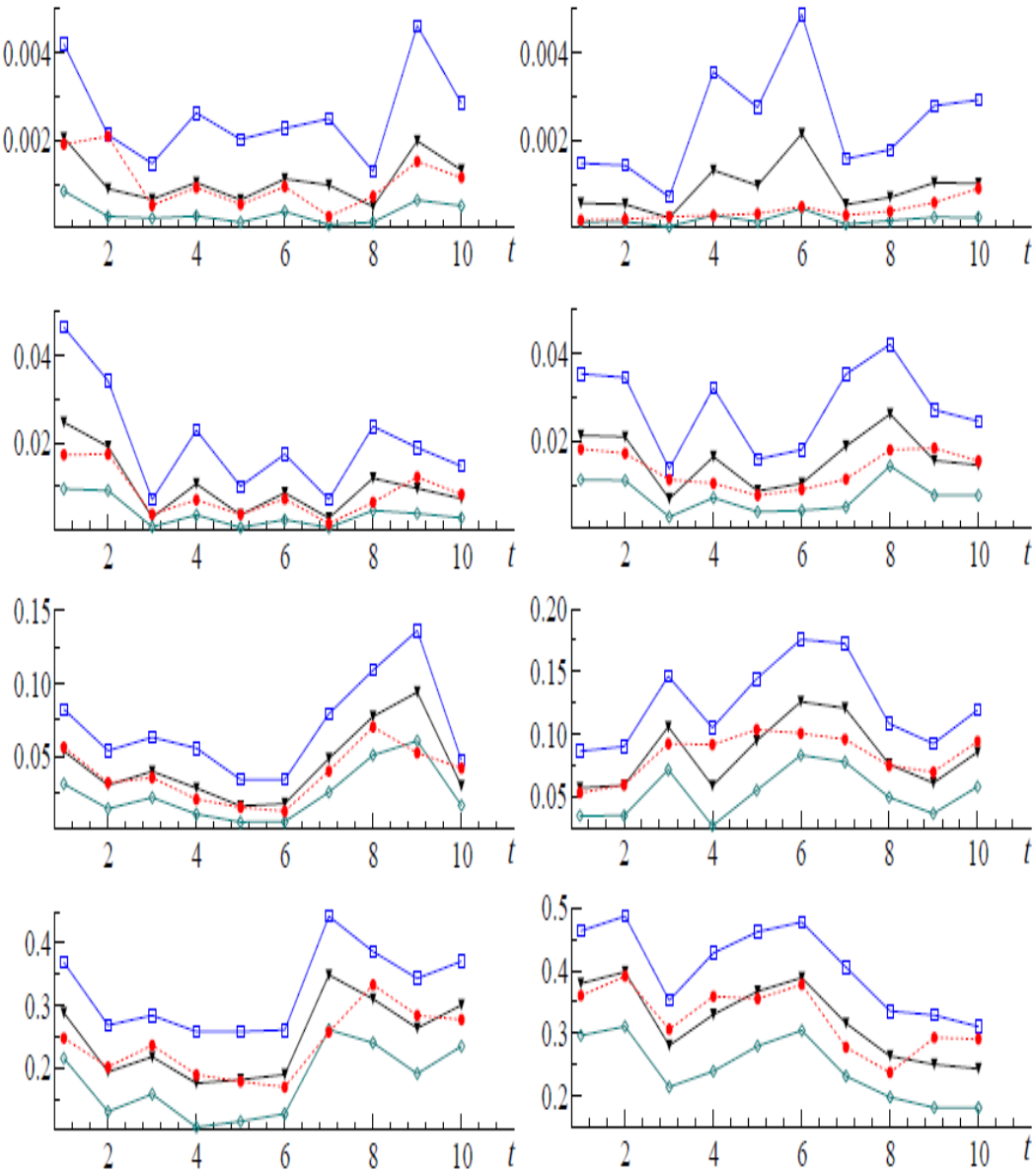
of variables, {Age [1], Marital status [10]}, {Attitude toward abortion [6], Attitude toward homosexual [16]} and {Attitude toward homosexual [16], Attitude toward Marijuana [19]} show strong interactions in the whole period. Also, we observed several pairs of variables showing relatively close interactions over time, such as {Attitude toward abortion [6], Think of self as liberal or conservative [14]}, {Race [3], Political party affiliation [9]} and {Marital status [10], Having gun [17]}. In addition, the views of government expense show moderate interactions, especially to the environment [23], nation's health [24], halting the rising crime [25], dealing with drug addiction [26] and education system [27].

Next, we study trends of dependence between categorical variables. Figure 2.6 reports the posterior means and 95% credible intervals of $\rho_{tjj'}$ for pairs with close interactions. We observed various patterns of time paths. For {Age, Marital status}, the interaction increased around 2000 then declined sharply to a lower level. {Race, Political party affiliation} and {Race, Having gun} have peaks in 2006 and the interactions have steeply decreased after that. In addition, we can see similar trends in {Attitude toward abortion, Think of self as lib or con}, {Attitude toward abortion, Attitude toward homosexual}, {Attitude toward homosexual, Attitude toward Marijuana}, {Religion, Attitude toward abortion} and {Religion, Attitude toward Marijuana}. The interactions have roughly increased over time, especially in the 2000s. On the other hand, the dependence in {Race, Death penalty for murder} decreased at first and kept stable in the middle of the period then declined again. {Having gun, Family income} gradually increased over the period but the difference is small. For {Marital status, Having gun}, the interaction dropped in the middle of the period but recovered recently at the same level as the beginning.



Red lines with circle symbols represent true values, black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals and blue-green lines with diamonds lower bounds of 95% intervals.
The first row: $P(x_{ti4}=1, x_{ti6}=2, x_{ti10}=3, x_{ti15}=4)$ and $P(x_{ti7}=3, x_{ti9}=1, x_{ti13}=4, x_{ti19}=2)$.
The second row: $P(x_{ti1}=3, x_{ti7}=2, x_{ti20}=4)$ and $P(x_{ti3}=4, x_{ti12}=2, x_{ti18}=1)$.
The third row: $P(x_{ti11}=2, x_{ti17}=2)$ and $P(x_{ti5}=3, x_{ti19}=2)$.
The fourth row: $P(x_{ti8}=1)$ and $P(x_{ti20}=4)$.

FIGURE 2.1: Estimation results of cell probabilities by the proposed method.



Red lines with circle symbols represent true values, black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals and bule-green lines with diamonds lower bounds of 95% intervals.
The first row: $P(x_{ti4}=1, x_{ti6}=2, x_{ti10}=3, x_{ti15}=4)$ and $P(x_{ti7}=3, x_{ti9}=1, x_{ti13}=4, x_{ti19}=2)$.
The second row: $P(x_{ti1}=3, x_{ti7}=2, x_{ti20}=4)$ and $P(x_{ti3}=4, x_{ti12}=2, x_{ti18}=1)$.
The third row: $P(x_{ti11}=2, x_{ti17}=2)$ and $P(x_{ti5}=3, x_{ti19}=2)$.
The fourth row: $P(x_{ti8}=1)$ and $P(x_{ti20}=4)$.

FIGURE 2.2: Estimation results of cell probabilities by DX method.

Table 2.1: Correlations between true values and posterior means of $\rho_{tjj'}$ using the first simulation data.

	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10	Total
Proposed	0.948	0.977	0.990	0.977	0.983	0.986	0.985	0.965	0.969	0.968	0.974
DX	0.837	0.794	0.880	0.761	0.766	0.921	0.846	0.817	0.831	0.793	0.841

Table 2.2: Correlations between true values and posterior means of $\rho_{tjj'}$ using the second simulation data.

	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	Total
Proposed	0.951	0.978	0.979	0.984	0.986	0.969	0.981	0.944	0.965
DX	0.872	0.803	0.838	0.599	0.807	0.884	0.932	0.827	0.696
DH	0.705	0.557	0.733	0.466	0.725	0.506	0.763	0.487	0.562

Table 2.3: Correlations between true values and posterior means of $\rho_{tjj'}$ using the third simulation data.

	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	Total
Proposed	0.725	0.827	0.768	0.798	0.818	0.916	0.791	0.807	0.817
DX	0.642	0.640	0.726	0.664	0.611	0.864	0.769	0.713	0.724
DH	0.371	0.716	0.821	0.491	0.611	0.877	0.764	0.715	0.624

Table 2.4: Contingency table of the religious preference and view of abortion in 2010.

	Protestant	Catholic	Jewish	None	Other	Total
Agree	216	103	21	137	60	537
Disagree	372	182	7	81	47	689
Total	588	285	28	218	107	1226

Table 2.5: Prediction results.

	Proposed	Model 1	Model 2
AD	194.4	208.6	204.5
MAPE	0.216	0.232	0.227

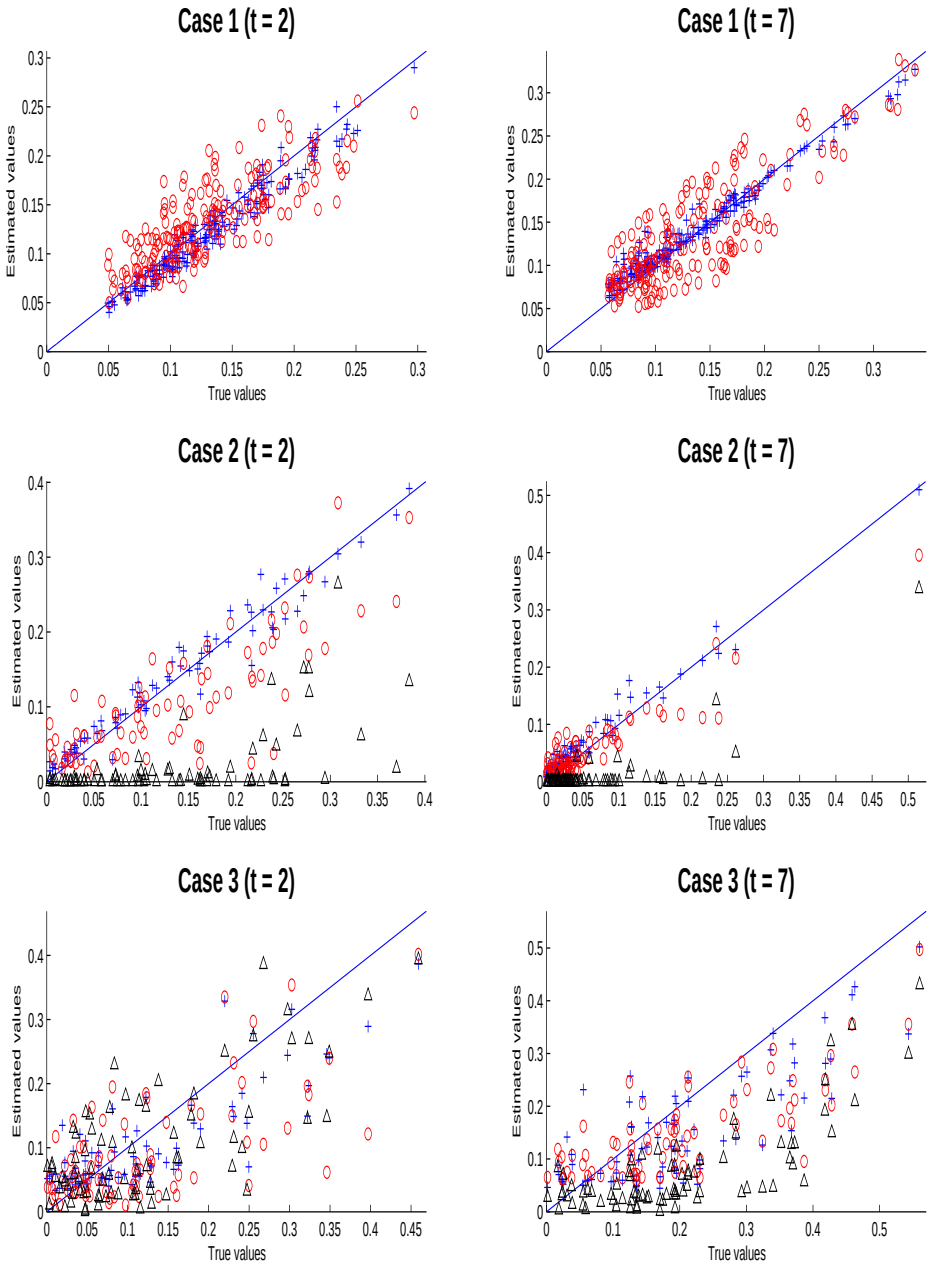


FIGURE 2.3: Plots of true and estimated values of $\rho_{tjj'}$ using the simulation data. y axis represents estimated values and x axis true values. Cross-shaped dots represent the proposed method, circles DX method and triangles DH method. The first, second and third rows show the results at time $t = 2$ and 7 using the first (case 1), second (case 2), third (case 3) simulation data sets.

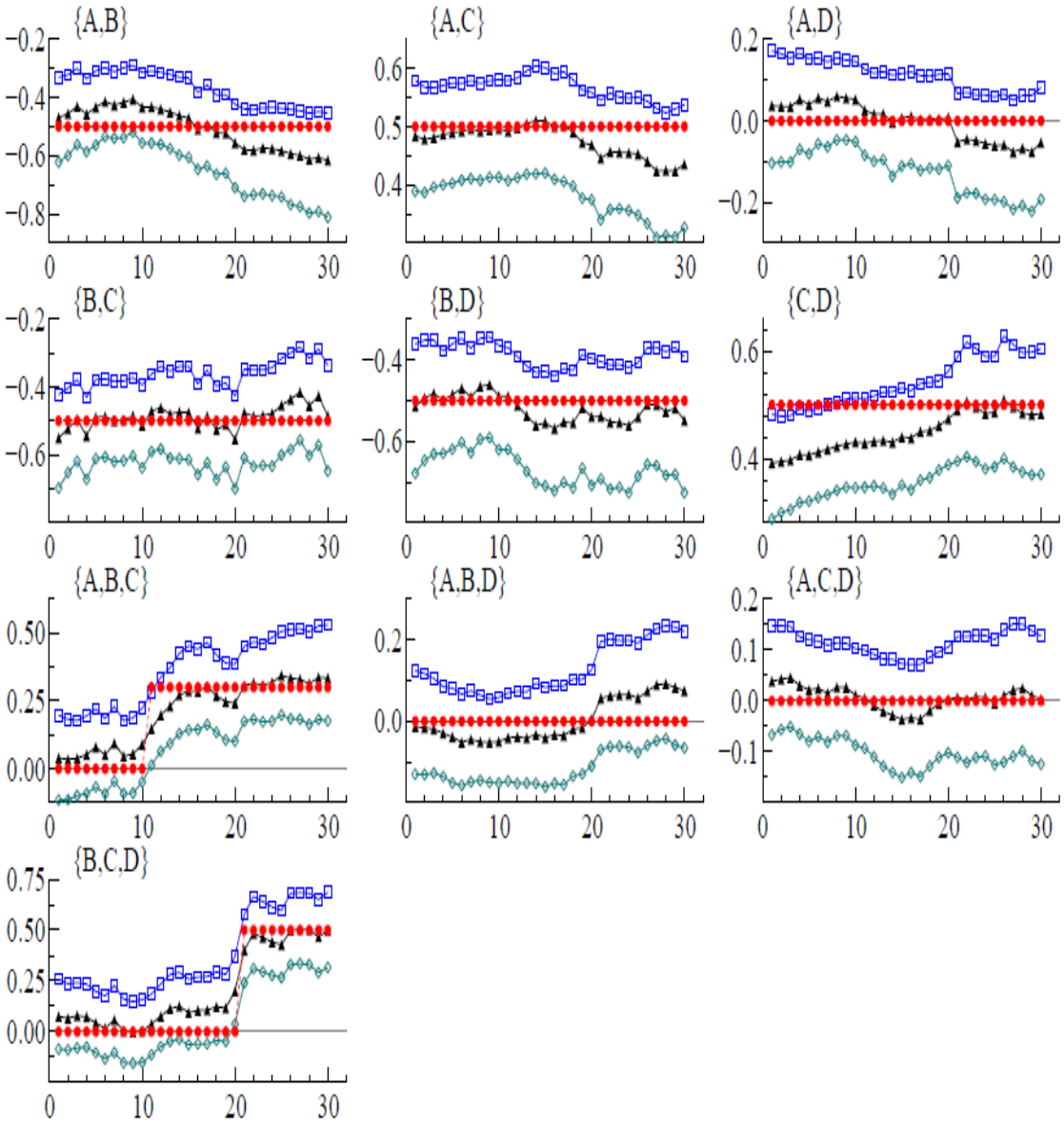


FIGURE 2.4: Estimation results of interactions in the parametric modeling. y axis represents estimated values and x axis time. Red lines with circle symbols represent true values, black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals and blue-green lines with diamonds lower bounds of 95% intervals.

Table 2.6: Estimation result of parameters in the proposed stick-breaking process.

Parameter	Mean	Stdev.	95% interval
μ	-0.012	0.004	[-0.023, -0.005]
ϕ	0.988	0.004	[0.978, 0.994]
σ_ε	0.062	0.009	[0.046, 0.082]
σ_η	0.126	0.011	[0.104, 0.149]

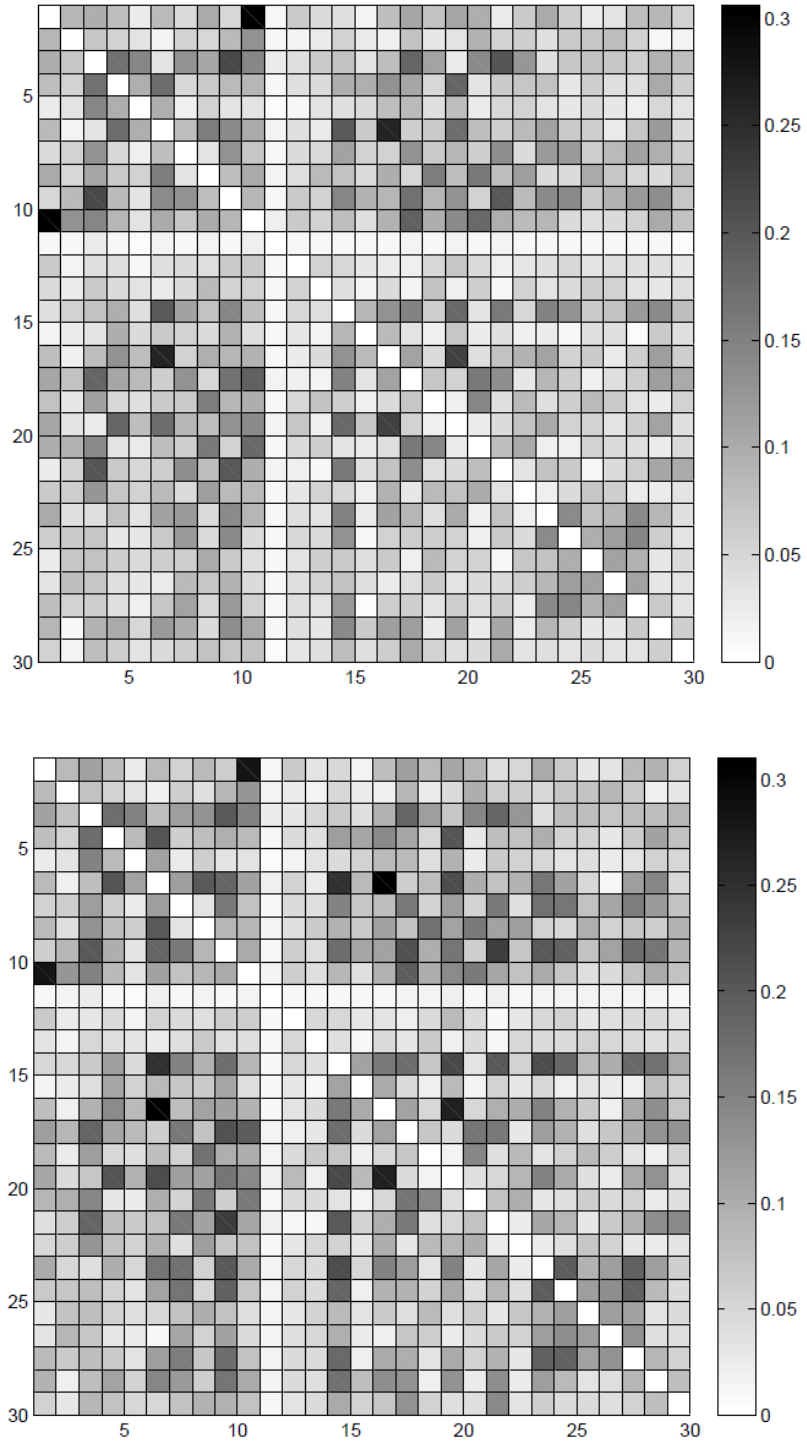
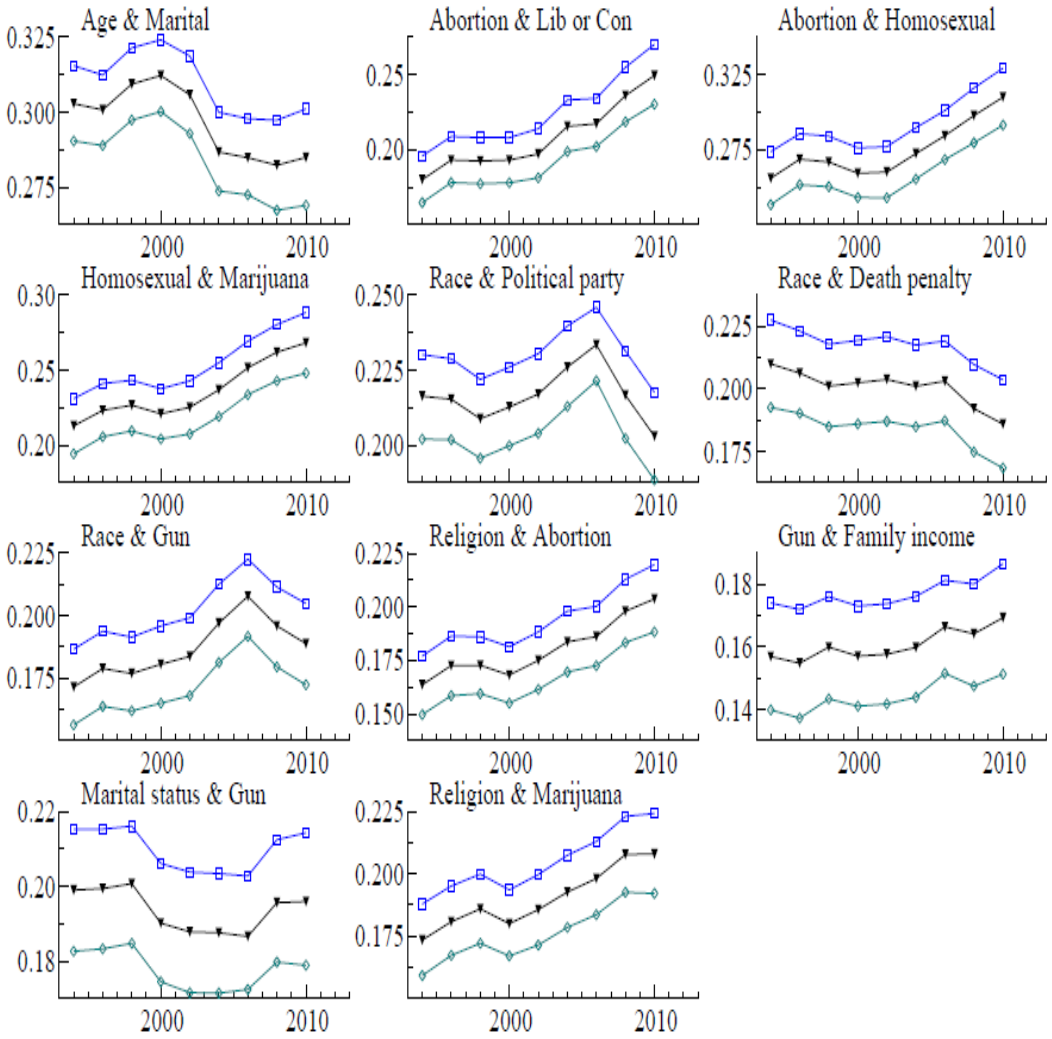


FIGURE 2.5: Posterior means of $\rho_{tjj'}$ in 2002 (above) and 2010 (below).



Black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals and bule-green lines with diamonds lower bounds of 95% intervals.
The first row: (Age group, Current marital status), (Attitude toward abortion, Think of self as liberal or conservative) and (Attitude toward abortion, Attitude toward homosexual sex relations).
The second row: (Attitude toward homosexual sex relations, Should Marijuana be made legal), (Race, Political party affiliation) and (Race, Favor or oppose death penalty for murder).
The third row: (Race, Have gun in home), (Religious preference, Attitude toward abortion) and (Have gun in home, Total family income).
The fourth row: (Current marital status, Have gun in home) and (Religious preference, Should Marijuana be made legal).

FIGURE 2.6: Estimation results of $\rho_{tjj'}$ for several pairs.

3

Nonparametric Bayes inference on conditional independence

3.1 Introduction

One of the canonical problems in statistics is to assess whether or not Y is conditionally independent of X given Z, expressed as $Y \perp X \mid Z$. In general, $Y \in \mathcal{Y}$ is a response, $X \in \mathcal{X}$ are predictors of interest, $Z \in \mathcal{Z}$ are adjustment variables or covariates, and the variables can be multivariate and have a variety of measurement scales and domains. There is a rich literature on testing of conditional independence in parametric models; often this corresponds to testing whether a vector of regression coefficients for the X variables are equal to zero. However, much less consideration has been given to this problem from a nonparametric perspective, particularly from a model-based Bayesian perspective.

In the frequentist literature, various nonparametric methods of testing conditional independence have been proposed, relying on different expressions of conditional independence with characteristic functions (Su and White (2007)), probability density functions (Su and White (2008); Pérez-Cruz (2008)), distribution functions (Seth and

Principe (2010); Györfi and Walk (2012)), copula densities (Bouezmarni et al. (2012)) and kernel methods (Fukumizu et al. (2008)). Seth and Príncipe (2012a) develop an asymmetric measure of conditional independence based on cumulative distribution functions. Also, Song (2009) constructs a test using Rosenblatt-transforms of random variables. However, these approaches do not work well in the case where the dimension of data is not small and the performance can be heavily affected by the choice of free parameters (Seth and Príncipe (2012b)).

A rich variety of Bayesian nonparametric models have been proposed for joint and conditional distributions, ranging from Dirichlet process mixtures (Lo (1984); West et al. (1994); Escobar and West (1995); Müller et al. (1996)) to kernel stick-breaking processes (Dunson and Park (2008); An et al. (2008)). However, such models do not allow testing of conditional independence relationships. A Bayesian decision-theoretic approach to the problem would (i) define a list of possible conditional independence relationships *a priori*, (ii) specify a nonparametric Bayes model for each relationship, (iii) calculate marginal likelihoods, and (iv) choose the relationship having minimal expected loss. However, a number of major practical problems arise. It is in general not straightforward to define a nonparametric Bayes model, which has full support on the space of distributions satisfying a particular conditional independence relationship, making (ii) problematic. Even if one could define appropriate models, (iii) is an issue due to the intractability of accurately approximating marginal likelihoods in infinite-dimensional Bayesian models. Also, even if (ii)-(iii) could be achieved, the behavior of marginal likelihoods in infinite-dimensional models is poorly understood, and misleading results are possible as mentioned in a 2012 Ohio State University PhD thesis by L. Pingbo.

There is a small literature on Bayesian nonparametric methods for variable selection (Chung and Dunson (2009); Ma (2013); Reich et al. (2012)), attempting to follow the above strategy in specialized settings. However, there has been essen-

tially no theoretic justification for these methods, and the practical implementation is limited to low-dimensional settings. In this article, we propose a substantially different approach. In particular, instead of attempting to select between different *exact* conditional independence relationships, we define an encompassing Bayesian nonparametric model, which is sufficiently flexible to approximate any relationship. We then use conditional mutual information as a scalar summary of the strength of departure from a particular conditional independence relationship. We estimate the conditional mutual information relying on a functional of the encompassing model and the empirical measure. The proposed framework is useful for rapid screening of variables that add significantly to prediction, and can be implemented easily leveraging on Markov chain Monte Carlo algorithms for the encompassing model. Based on empirical process theory, we show that the proposed method consistently selects conditionally dependent predictors under appropriate conditions.

3.2 Inference on conditional independence

3.2.1 Conditional mutual information

Let Y, X and Z be univariate or multivariate random variables where each element can have any type of scale and domain. We also let $f(y, x, z)$ denote the joint density of Y, X and Z with respect to a product measure ξ. The marginal densities we use below are denoted by $f(y, z)$, $f(x, z)$ and $f(z)$. Suppose the primary interest is in assessing if Y and X are conditionally independent given Z. Relying on the joint density, $Y \perp X \mid Z$ can be equivalently expressed as

$$f(y, x, z)f(z) = f(y, z)f(x, z),$$

for all (y, x, z) in the support of f.

In information theory, conditional mutual information measures the strength of functional relationship between Y and X given Z (Wyner (1978); Joe (1989); MacKay (2003); Cover and Thomas (2006)),

$$\zeta = \int f(y,x,z) \log \frac{f(y,x,z)f(z)}{f(y,z)f(x,z)} d\xi.$$

Letting $KL(p,q) = \int p \log(p/q)$ denote the Kullback-Leibler divergence, $\zeta = KL\{f(y,x,z), f(y,z)f(x,z)/f(z)\}$, which is always non-negative. In general, $\zeta = 0$ if and only if $Y \perp X \mid Z$, while large values of ζ indicate substantial violations of conditional independence with an approximate functional relationship between Y and X given Z.

3.2.2 Empirical Bayes estimation of conditional mutual information

Let P_0 denote a *true* data-generating probability having density $f_0 \in L_\xi$, with L_ξ the set of all probability densities with respect to a measure ξ. Let Π denote a prior probability on L_ξ with $\Pi(\mathcal{F}) = 1$ for $\mathcal{F} \subset L_\xi$. Data D_n consist of independently identically distributed observations (y_i, x_i, z_i) from P_0 with $i = 1, \ldots, n$. Let ζ_0 be the conditional mutual information induced by the true data-generating distribution,

$$\zeta_0 = \int \log \frac{f_0(y,x,z)f_0(z)}{f_0(y,z)f_0(x,z)} dP_0 = \int f_0(y,x,z) \log \frac{f_0(y,x,z)f_0(z)}{f_0(y,z)f_0(x,z)} d\xi.$$

As noted above, $Y \perp X \,|\, Z$ if and only if $\zeta_0 = 0$. To estimate ζ_0, we rely on an encompassing nonparametric Bayes model for the joint density f_0. First, we define a function $\zeta(\cdot,\cdot)$ of a joint density $f \in L_\xi$ and a probability measure P on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ as

$$\zeta(f,P) = \int \log \frac{f(y,x,z)f(z)}{f(y,z)f(x,z)} dP. \tag{3.1}$$

Using this function, ζ_0 can be expressed as $\zeta(f_0, P_0)$. Intuitively, if f and P are close to f_0 and P_0 in some sense, $\zeta(f,P)$ can approximate ζ_0 well. In general, a probability measure P having a density leads to a computationally intractable $\zeta(f,P)$ because of the difficulty in evaluating its integral. Therefore, we utilize the empirical measure

as an estimate of P_0,

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(y_i,x_i,z_i)},$$

where $\delta_{(y,x,z)}$ is the Dirac measure concentrated at (y, x, z). The empirical measure P_n is a consistent estimate of P_0 in that $P_n(A) \to P_0(A)$ almost surely for any A by the strong law of large numbers. Then, we let

$$\zeta(f, P_n) = \int \log \frac{f(y,x,z)f(z)}{f(y,z)f(x,z)} dP_n = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(y_i,x_i,z_i)f(z_i)}{f(y_i,z_i)f(x_i,z_i)}, \quad f \in \mathcal{F}, \qquad (3.2)$$

where $\zeta(f, P_n) \in \Re$ and, for any fixed $f \in \mathcal{F}$, $\zeta(f, P_n) \to \zeta(f, P_0)$ almost surely P_0^∞ by the law of large numbers. By using the empirical measure P_n for P_0 while defining a nonparametric Bayes encompassing prior for the joint density f, we define an empirical Bayes approach that induces a posterior on ζ accounting for uncertainty. In finite samples this posterior assigns non-zero probability to $\zeta < 0$, which results because P_n does not exactly correspond to the measure induced from the density f.

Plugging in the empirical measure P_n, expression (3.2) for the conditional mutual information depends on the unknown joint density f and corresponding marginals. Updating prior $f \sim \Pi$ with data $(y_i, x_i, z_i), i = 1, \ldots, n$, we obtain a posterior quantifying our current state of knowledge about the density f. We can obtain samples from this posterior by running Markov chain Monte Carlo for the encompassing model ignoring any conditional independence structure. Then, to marginalize f out of expression (3.2) and obtain an empirical Bayes estimate of ζ_0, we simply use Monte Carlo integration. In particular, for each draw from the posterior, we compute and save $\zeta(f, P_n)$. The resulting draws of ζ are from the induced empirical Bayes posterior of the conditional mutual information; we use this posterior as the basis for our inferences.

Under our asymptotic theory below, as n increases the posterior of $\zeta(f, P_n)$ will be increasingly concentrated around the true conditional mutual information ζ_0. There-

fore, if ζ_0 is not close to zero, zero should locate in the left tail of the distribution of $\zeta(f, P_n)$. We consider the posterior probability of $\zeta(f, P_n)$ being positive as a weight of evidence of violations of conditional independence. The posterior probability can be estimated by $(1/R)\sum_{r=1}^{R} 1\{\zeta(f^{(r)}, P_n) > 0\}$ where R is the number of Markov chain Monte Carlo iterations after the burn-in period, $1\{\cdot\}$ is an indicator function and $f^{(r)}$ is the joint density under the encompassing model at the rth iteration.

3.2.3 Theoretic support

The next theorem provides sufficient conditions under which the posterior of $\zeta(f, P_n)$ concentrates on arbitrarily small neighborhoods of ζ_0 as the sample size increases.

Theorem 4. *Suppose for any $\epsilon > 0$,*

$$\Pi\left[KL\{f_0(y, x, z), f(y, x, z)\} < \epsilon\right] > 0 \tag{3.3}$$

and the following classes of functions

$$\left\{\log \frac{f_0(y, x, z)}{f(y, x, z)}\right\}, \left\{\log \frac{f_0(y, z)}{f(y, z)}\right\}, \left\{\log \frac{f_0(x, z)}{f(x, z)}\right\}, \left\{\log \frac{f_0(z)}{f(z)}\right\},$$

are P_0-Glivenko-Cantelli. Then, for any $\epsilon' > 0$

$$\Pi\left\{|\zeta(f, P_n) - \zeta_0| < \epsilon' \mid D_n\right\} \to 1, \quad \textit{almost surely } P_0^{\infty}.$$

The proof is in Appendix B. The condition (3.3) means the true data-generating density is in the Kullback-Leibler support of the prior. Such support conditions are standard for Bayesian nonparametric models, and are routinely employed in theorems of posterior asymptotics (Ghosal et al. (1999); Ghosh and Ramamoorthi (2003); Tokdar (2006)). Wu and Ghosal (2008) discuss the Kullback-Leibler property for various types of kernels in Dirichlet process mixture models. As for the Glivenko-Cantelli class, theoretical properties of the class have been studied in empirical process theory (van der Vaart and Wellner (1996); Kosorok (2008)). It is a wide class of functions such that the law of large numbers holds uniformly over the space.

3.2.4 Variable selection

Suppose we have a univariate response $Y \in \mathcal{Y}$ and vector of predictors $X = (X_1, \ldots, X_p)'$. Conditional mutual information provides a measure of how much information a particular predictor X_j adds when included in a model already containing the predictors in $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)'$. We can potentially use our method for predictive variable selection, conducting a search for the smallest subset of variables $\gamma \subset \{1, \ldots, p\}$ such that there is no evidence of departure from $Y \perp X_{-\gamma} \mid X_\gamma$, with $X_\gamma = \{X_j : j \in \gamma\}$ and $X_{-\gamma} = \{X_j : j \notin \gamma\}$. However, instead of identifying parsimonious models for predicting Y, we focus here on selecting predictors that add significantly to models containing all other predictors. This reduces the search from 2^p to p, while still producing results of inferential interest. The computational savings come at the potential expense of excluding a set of important predictors containing redundant information about Y.

Let $\zeta_{0,j}$ be the true conditional mutual information for $Y \perp X_j \mid X_{-j}$. Let $\zeta_j(f, P_n)$ denote the value of $\zeta(f, P_n)$ in expression (2) with x the jth predictor and z the other predictors. Posterior computation proceeds as in subsection 3.2.2. We use the posterior probability of $\zeta_j(f, P_n) > 0$ as evidence of violating $Y \perp X_j \,|\, X_{-j}$ for $j = 1, \ldots, p$, selecting predictors having large probabilities. This method is justified by the next theorem, which indicates zero should be in the left tail of the posterior distribution of $\zeta_j(f, P_n)$ under conditional dependence.

We show posterior consistency of $\zeta_j(f, P_n)$ to $\zeta_{0,j}$ under appropriate conditions. Theorem 5 modifies Theorem 4 to the case of measuring dependence between each predictor and the response, adjusting for all other predictors as covariates. The difference from Theorem 4 is the Glivenko-Cantelli class condition depends on j. Also, Theorem 5 states the posterior of $\zeta_j(f, P_n)$ will concentrate on $\zeta_{0,j}$ uniformly over j as the sample size increases, allowing us to avoid multiple separate pairwise

comparisons. The proof is in the Appendix B.

Theorem 5. *Suppose for any* $\epsilon > 0$,

$$\Pi\left[KL\{f_0(y,x), f(y,x)\} < \epsilon\right] > 0 \tag{3.4}$$

and the following classes of functions

$$\left\{\log \frac{f_0(y,x)}{f(y,x)}\right\}, \left\{\log \frac{f_0(x)}{f(x)}\right\}, \left\{\log \frac{f_0(y,x_{-j})}{f(y,x_{-j})}\right\}, \left\{\log \frac{f_0(x_{-j})}{f(x_{-j})}\right\},$$

are P_0*-Glivenko-Cantelli with* $j = 1, \ldots, p$. *Then, for any* $\epsilon' > 0$

$$\Pi\left\{\max_{1\leq j\leq p} |\zeta_j(f, P_n) - \zeta_{0,j}| < \epsilon' \mid D_n\right\} \to 1, \quad \textit{almost surely } P_0^\infty.$$

We illustrate a simple but non-trivial encompassing model which satisfies the sufficient conditions. Let $y \in \Re$, $x \in \Re^p$ and ϕ_σ be the univariate normal density with mean 0 and standard deviation σ. Then, we consider location mixtures of normals in which the kernel is the product of a regression density for the response and independent normal densities for the predictors,

$$f(y,x) = \int \phi_\sigma(y - \tilde{x}'\beta) \prod_{j=1}^{p} \phi_{\tau_j}(x_j - \mu_j) Q(d\beta, d\mu), \tag{3.5}$$

where $\tilde{x} = (1, x')'$, $\beta = (\beta_0, \ldots, \beta_p)'$, $\tau = (\tau_1, \ldots, \tau_p)'$ and $\mu = (\mu_1, \ldots, \mu_p)'$. Dirichlet process mixture models of this type have been widely studied (West et al. (1994); Escobar and West (1995); Müller et al. (1996); Hannah et al. (2011)). We assume the mixing measure Q can be expressed as

$$Q = \sum_{h=1}^{\infty} \pi_h \delta_{(\beta_h, \mu_h)}, \quad \pi_h \geq 0, \quad \sum_{h=1}^{\infty} \pi_h = 1, \quad (\beta_h, \mu_h) \sim G, \tag{3.6}$$

where $\beta_h = (\beta_{0,h}, \ldots, \beta_{p,h})'$, $\mu_h = (\mu_{1,h}, \ldots, \mu_{p,h})'$ and G is a distribution on $\Re^{p+1} \times \Re^p$. This class of functions (3.5) and (3.6) includes Dirichlet process mixtures with $\pi_h = V_h \prod_{l<h}(1 - V_l)$, $V_h \sim \text{Be}(1, \alpha_0)$ for $h = 1, \ldots, \infty$ (Sethuraman (1994)). The

prior distribution for the joint densities is induced through $\Pi = \Pi^{Q} \times \Pi^{(\sigma,\tau)}$ where Π^{Q} and $\Pi^{(\sigma,\tau)}$ are the prior distributions for Q and (σ, τ). Under some conditions on f_0 and Π, the next lemma illustrates the encompassing model (3.5) and (3.6) assures consistency.

Lemma 6. *Suppose the true density can be expressed in the form* $f_0(y,x) = \int \phi_{\sigma_0}(y - \tilde{x}'\beta) \prod_{j=1}^{p} \phi_{\tau_{0,j}}(x_j - \mu_j) Q_0(d\beta, d\mu)$. *If* G *has compact support,* $\Pi^{(\sigma,\tau)}$ *has compact support excluding zero,* Q_0 *belongs to the support of* Π^{Q} *and* $(\sigma_0,\ \tau_0)$ *are in the support of* $\Pi^{(\sigma,\tau)}$, *then* $\Pi\left\{\max_{1\leq j\leq p} |\zeta_j(f, P_n) - \zeta_{0,j}| < \epsilon' \mid D_n\right\} \to 1$ *almost surely* P_0^{∞}.

The proof relies on Theorem 3 in Ghosal et al. (1999) and is in the Appendix B. As Remark 1 in Ghosal et al. (1999) mentions, the result can be extended to a wider class of location-scale mixture of normals. The condition of compact support is sufficient but not necessary.

3.3 Simulation study

In this section, we assess performance of the proposed method compared to frequentist nonparametric alternatives. As competitors, we employ a method based on cumulative distribution functions with Cramér-von-Mises type statistics from an unpublished 1996 technical report by O. Linton and P. Gozalo, the kernel measure method based on normalized cross-covariance operators on reproducing kernel Hilbert spaces (Fukumizu et al. (2008)) and the asymmetric quadratic measure (Seth and Príncipe (2012a)). Matlab code for these methods is available at http://www.sohanseth.com/Home/codes and we use the default settings recommended in Seth and Príncipe (2012a) with a Gaussian kernel for Fukumizu et al. (2008) and a Laplacian function for the asymmetric quadratic measure. Also, for these methods, we reject the hypothesis $Y \perp X_j \mid X_{-j}$ if $B^{-1} \sum_{b=1}^{B} 1(d_b^* > d) < 0.1$ where d

and d_b^* are the estimated conditional dependences using the observation and the bth randomly rearranged observation which mimics the case of conditional independence (Diks and DeGoede (2001)) with $b = 1, \ldots, B$ and $B = 100$. In addition, we apply the lasso function in Matlab using 5-fold cross validation for penalty coefficient selection and other default settings. We evaluate performance based on the following measures: type 1 error (false positive/(false positive+true negative)), type 2 error (false negative/(true positive+false negative)), positive predictive value (true positive/positive), negative predictive value (true negative/negative) and accuracy ((true positive+true negative)/(positive+negative)).

As an encompassing model, we employ the following Dirichlet process location-scale mixture,

$$f(y, x) = \int \phi_\sigma(y - \tilde{x}'\beta) \prod_{j=1}^{p} \phi_{\tau_j}(x_j - \mu_j) Q(d\beta, d\mu, d\sigma, d\tau), \tag{3.7}$$

$$= \sum_{h=1}^{H} \pi_h \phi_{\sigma_h}(y - \tilde{x}'\beta_h) \prod_{j=1}^{p} \phi_{\tau_{j,h}}(x_j - \mu_{j,h}), \tag{3.8}$$

where $\pi_h = V_h \prod_{l<h}(1 - V_l)$, $V_h \sim \text{Be}(1, \alpha_0)$ for $h = 1, \ldots, H-1$ with $V_H = 1$, $\beta = (\beta_0, \ldots, \beta_p)'$, $\tilde{x} = (1, x')'$, $\mu = (\mu_1, \ldots, \mu_p)'$ and $\tau = (\tau_1, \ldots, \tau_p)'$. As discussed in subsection 3.2.4, if the base measure of the Dirichlet process has compact support, we obtain consistent estimators of the conditional mutual information for each predictor. Compact support is a simplifying assumption for the theory, which can be relaxed, and we avoid this restriction in the computation letting $\sigma^2 \sim \text{Inverse-Gamma}(1.5, 0.5)$, $\mu_{j,h} \sim N(0, 1)$, $\tau_{j,h}^2 \sim \text{Inverse-Gamma}(1.5, 0.5)$ and $\alpha_0 \sim \text{Ga}(0.25, 0.25)$. To allow a sparse regression structure, we use a point mass mixture prior: $\beta_j \sim p_0\delta_0 + (1 - p_0)N(0, \lambda_j^2)$, $\lambda_j^2 \sim \text{Inverse-Gamma}(0.5, 0.5)$, λ_j are mutually independent over j and $p_0 \sim \text{Be}(4.75, 0.25)$. By integrating out λ_j^2, this prior corresponds to a mixture of a degenerate distribution concentrated at zero

and a Cauchy distribution. The prior for exclusion probability p_0 assumes 5% of regression coefficients out of $H(p+1)$ components are non-zero but allows substantial uncertainty since the prior sample size is set to be 4.75+0.25=5. Also, we set $H = 20$. Before posterior computation, we normalize data to have mean zero and standard deviation one. We draw 10,000 samples after the initial 5,000 samples are discarded as a burn-in period and every 10th sample is saved. Rates of convergence and mixing were adequate. We conclude there is substantial evidence of violations of $Y \perp X_j \mid X_{-j}$ if $\Pi\{\zeta_j(f, P_n) > 0 \mid D_n\} > 0.95$ with $j = 1, \ldots, p$.

We consider three different data-generating functions from which we simulate 100 data sets with $n = 100$ and $p = 10$. First, we generate data from a linear regression model with strong dependence among predictors.

$$\begin{aligned} \text{Case 1}: \quad & y_i = -x_{i,1} + x_{i,4} - x_{i,7} + \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \\ & x_i = (x_{i,1}, \ldots, x_{i,10}) \sim N(0, \Sigma_x), \\ & \Sigma_x = \{\sigma_{j,j'}\}, \ \sigma_{j,j'} = \mathrm{cov}(x_{i,j}, x_{i,j'}) = 0.7^{|j-j'|}, \end{aligned}$$

where $\{y_i\}$ are independent over i. The left panel in Figure 3.1 and last column in Table 3.1 show the receiver operating characteristic curves and area under the curve averaged over 100 data sets in Case 1. For the proposed method, we obtain the curve by shifting the threshold a in $\Pi\{\zeta_j(f, P_n) > a \mid D_n\} > 0.95$. For the lasso, we shift the threshold for absolute values of regression coefficients. We set the thresholds as $2.5k\%$ quantile points of all estimated measures of conditional dependence over 100 data sets for each method with $k = 0, \ldots, 40$. Although the area under the curve for the proposed method is slightly smaller than that for the lasso and the asymmetric quadratic measure, it is large and close to one. The top of Table 3.1 reports averaged measures of the test performance over 100 data sets in Case 1. For the lasso, its high type 1 error and low positive predictive value indicate it incorrectly rejects many hypotheses. Though the data are generated from the

linear model, the strong dependence among predictors can cause poor performance. On the other hand, high type 2 errors and low negative predictive values in the Cramér-von-Mises type statistic and asymmetric quadratic measure imply that they often fail to detect dependent relations. The normalized cross-covariance operator also faces the same problem of missing dependent predictors but the performance is much better. The proposed method works quite well, reporting small type 1 and 2 errors and high positive and negative predictive values. Compared to the normalized cross-covariance operator, there is not a big difference in measures with false positives but the proposed method less often produces false negatives since the new approach shows a lower type 2 error and a higher negative predictive value.

Next, we generate data from a model in which the strong dependence among predictors remains but the relation between the response and predictors is non-linear.

$$\begin{aligned} \text{Case 2}: \quad & y_i = -x_{i,1} + \exp(x_{i,4}) - x_{i,7}^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \\ & x_i = (x_{i,1}, \ldots, x_{i,10}) \sim N(0, \Sigma_x), \\ & \Sigma_x = \{\sigma_{j,j'}\}, \ \sigma_{j,j'} = \text{cov}(x_{i,j}, x_{i,j'}) = 0.7^{|j-j'|}. \end{aligned}$$

The receiver operating characteristic curves and area under the curve in Case 2 are in the middle of Figure 3.1 and Table 3.1. Though the competitors' curves are away from the random guess line $y = x$, the proposed method shows largest area under the curve. The middle of Table 3.1 summarizes the test performance measures. The proposed method reports small type 1 and 2 errors and high positive and negative predictive values and accuracy. From the high type 1 error and small positive predictive value, the lasso tends to wrongly pick up conditionally independent predictors. The high type 2 error and small negative predictive value indicate the Cramér-von-Mises type statistic and asymmetric quadratic measure have difficulty in finding dependent structures. The normalized cross-covariance operator performs better than the Cramér-von-Mises type statistic and asymmetric quadratic measure

but still reports a high type 2 error and a low negative predictive value compared to the proposed method.

We also simulate data from a different non-linear model where the dependence comes from division of the sample into subgroups and non-linear regressions.

$$\text{Case 3}: \quad y_i = \begin{cases} 0.8x_{i,1}^2 - x_{i,4} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 0.7^2), & \text{if } s_i = 0, \\ -x_{i,1} + 1.2\exp(x_{i,7}) + \varepsilon_i, \quad \varepsilon_i \sim N(0,1), & \text{if } s_i = 1. \end{cases}$$

$$s_i \sim \text{Bernoulli}(0.5), \quad x_{i,j} \sim N(\mu_{j,s_i}, \sigma_{j,s_i}^2), \ j = 1, \ldots, 10,$$

$$\mu_{j,s} \sim N(0,1), \ \sigma_{j,s}^2 \sim \text{Inverse-Gamma}(2,\, 0.5), \ s \in \{0,1\},$$

$$\mu_{j,0} = \mu_{j,1}, \quad \sigma_{j,0}^2 = \sigma_{j,1}^2, \quad j \notin \{1,4,7\}.$$

The right plot in Figure 3.1 and last column in Table 3.1 correspond to the receiver operating characteristic curves and area under the curve in Case 3. The Cramér-von-Mises type statistic works poorly with the curve close to the random guess line. The area under the curve by the proposed method is smaller than that for the asymmetric quadratic measure but the curve is still far away from the $y = x$ line. The bottom in Table 3.1 reports measures of the test performance. The lasso is likely to reject correct hypotheses and the Cramér-von-Mises type statistic produces the worst results in all measures except the type 1 error. The proposed method, the normalized cross-covariance operator and asymmetric quadratic measure show small type 1 errors and high positive predictive values, indicating they less likely produce false positives. As for the false negatives, the differences in the type 2 errors and negative predictive values between the proposed method and the normalized cross-covariance operator are small with the asymmetric quadratic measure slightly worse. Also, the proposed method leads to the highest accuracy among them. Overall these simulation results are promising that the proposed method has relatively good performance.

Table 3.1: Averages of type 1 and 2 errors, positive and negative predictive values, accuracy and area under the curve in Case 1 (top), Case 2 (middle) and Case 3 (bottom). Proposed, proposed method; CM, Cramér-von-Mises type statistic; NCCO, normalized cross-covariance operator; AQM, asymmetric quadratic measure; PPV, positive predictive value; NPV, negative predictive value; ACC, accuracy; AUC, area under the curve.

Case 1	Type 1	Type 2	PPV	NPV	ACC	AUC
Proposed	2.2	12.6	95.5	95.6	94.6	98.4
LASSO	49.7	0.0	50.6	100.0	65.2	99.9
CM	0.2	80.3	97.1	74.6	75.7	80.6
NCCO	0.1	24.3	99.6	91.3	92.6	92.8
AQM	0.0	67.6	100.0	77.9	79.7	98.6
Case 2	Type 1	Type 2	PPV	NPV	ACC	AUC
Proposed	4.0	12.0	92.8	95.5	93.6	98.9
LASSO	32.0	20.0	58.5	89.1	71.6	84.8
CM	1.7	90.6	71.9	71.7	71.6	64.3
NCCO	0.2	37.0	99.4	87.4	88.7	87.8
AQM	0.0	76.0	100.0	75.9	77.2	97.3
Case 3	Type 1	Type 2	PPV	NPV	ACC	AUC
Proposed	2.8	27.0	94.3	90.4	89.9	89.6
LASSO	27.2	27.6	64.7	88.8	72.6	78.4
CM	15.5	78.0	43.2	72.1	65.7	47.6
NCCO	3.5	27.0	94.0	90.2	89.4	82.4
AQM	0.2	41.3	99.4	85.5	87.4	94.7

3.4 Application to criminology data

In this section, we apply the proposed method to communities and crime data from the University of California Irvine machine learning repository. Details of the data are in the Appendix B. The data set is culled from 1990 United States census, 1995 United States Federal Bureau of Investigation uniform crime report and 1990 United States law enforcement management and administrative statistics survey. Data include various types of crime and demographic information for $n = 2,215$ communities in the United States. We use 10 count variables as responses: numbers of murders, rapes, robberies, assaults, burglaries, larcenies, auto thefts, arsons, violent crimes (sum of murders, rapes, robberies and assaults) and non-violent crimes (sum of bur-

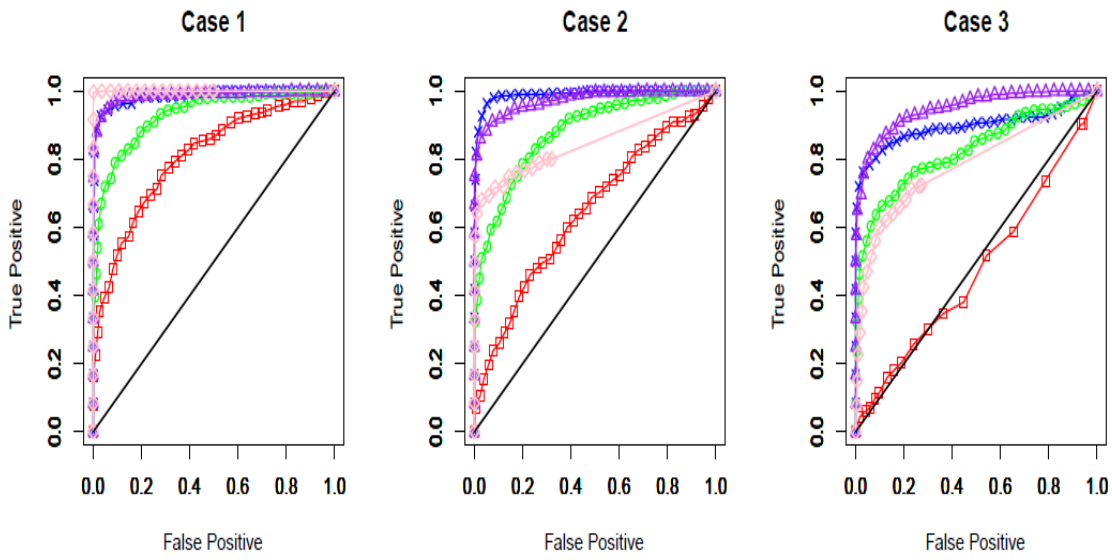


FIGURE 3.1: Receiver operating characteristic curves and area under the curve curves in Case 1 (left), Case 2 (middle) and Case 3 (right). y axis represents the true positive rate and x axis the false positive rate. Blue crosses, pink diamonds, red square, green circles and purple triangles indicate the averages of the true and false positive rates over 100 data sets for the proposed method, lasso, Cramér-von-Mises type statistic, normalized cross-covariance operator and asymmetric quadratic measure.

glaries, larcenies, auto thefts and arsons). As predictors, we select $p = 68$ variables, such as per capita income and population density, which indicate demographic characteristics of the communities. The list is in the Appendix B. The data set consists of count, percentage and positive continuous variables. We observe the count variables have right-skewed distributions and the percentage variables can inflate at 0% and 100%. Also, the data set includes missing values in the response.

To incorporate mixed-scale measurements, we develop a joint model which relies on the rounded kernel method of Canale and Dunson (2011). Let $y^* \in \Re$ and $x^* = (x_1^*, \ldots, x_p^*)' \in \Re^p$ be latent continuous variables for the response y and predictors $x = (x_1, \ldots, x_p)'$. We induce a flexible nonparametric model on y and x through a Dirichlet process mixture of normals for the latent variables. If x_j is a count variable, it can be expressed as $x_j = l$ if $a_l < x_j^* \leq a_{l+1}$ with $l = 0, 1, 2, \ldots$ where

$-\infty = a_0 < a_1 < a_2 < \cdots$ with $a_l = \log(l)$ for $l \geq 1$. This expression corresponds to $x_j = [\exp(x_j^*)]$ where $[x]$ denotes the maximum integer smaller than x. Since the log function shrinks large values, a distribution with positive skewness can be efficiently approximated by mixtures of normals with the log cut-points. Percentage variables with inflation at 0 and 100% can be induced by

$$x_j = \begin{cases} 0 & \text{if } x_j^* \leq 0, \\ x_j^* & \text{if } 0 < x_j^* < 100, \\ 100 & \text{if } 100 \leq x_j^*. \end{cases}$$

As for a positive continuous variable, we apply the log transformation to the original data and treat it as a continuous variable with $x_j = x_j^*$. For the latent variables, we utilize the Dirichlet process mixture of normals (3.7) and (3.8) except we use the observed predictors for the regression on y^*. Then, we obtain the following joint model of y and x by integrating out the latent variables.

$$f(y, x) = \sum_{h=1}^{H} \pi_h f(y \mid x, \theta_h) \prod_{j=1}^{p} f(x_j \mid \theta_h), \tag{3.9}$$

where $\pi_h = V_h \prod_{l<h}(1 - V_l)$, $V_h \sim \text{Be}(1, \alpha_0)$ for $h = 1, \ldots, H-1$ with $V_H = 1$, θ is a parameter set in the model and

$$f(y \mid x, \theta) = \int_{a_y}^{a_{y+1}} \phi_\sigma(y^* - \tilde{x}'\beta) dy^* = \Phi(a_{y+1} \mid \tilde{x}'\beta, \sigma) - \Phi(a_y \mid \tilde{x}'\beta, \sigma), \tag{3.10}$$

and

$$f(x_j \mid \theta) = \begin{cases} \text{count: } \Phi(a_{x_j+1} \mid \mu_j, \tau_j) - \Phi(a_{x_j} \mid \mu_j, \tau_j), \\ \text{percentage: } 1(x_j = 0)\Phi(0 \mid \mu_j, \tau_j) + 1(x_j = 100)\{1 - \Phi(100 \mid \mu_j, \tau_j)\} \\ \quad +1(0 < x_j < 100)\phi_{\tau_j}(x_j - \mu_j), \\ \text{continuous: } \phi_{\tau_j}(x_j - \mu_j), \end{cases}$$

where $1(\cdot)$ is an indicator function and $\Phi(\cdot \mid a, b)$ is the cumulative density function of normal with mean a and standard deviation b. We constructed priors relying on empirical information, $\sigma^2 \sim \text{Inverse-Gamma}(1.5, s_y^2/2)$ where s_y^2 is the sample variance

of $\log(y_i + 0.5)$ since $y_i = 0$ for certain subjects. Also, we use $\mu_j \sim N(\bar{\mu}_j, s_j^2)$ and $\tau_j^2 \sim$ Inverse-Gamma$(1.5, s_j^2/2)$ where $\bar{\mu}_j$ and s_j^2 are the sample mean and variance of $\log(x_{i,j} + 0.5)$ for a count and of $x_{i,j}$ for a percentage and a continuous variable. The priors for α_0 and β are the same as in Section 3.3. We standardize the predictors in (3.10) so that each variable has mean zero and standard deviation one. Assuming missing at random, we impute missing values at each Markov chain Monte Carlo iteration from the conditional distributions given observed data. The details of the Markov chain Monte Carlo algorithm are in the Appendix B. We apply the proposed method with $H = 20$ separately to each response. We draw 80,000 samples from the posterior after the initial 5,000 samples are discarded as a burn-in period and every 20th sample is saved. We observe that the sample paths were stable and the sample autocorrelations dropped smoothly; hence we concluded the chains converged. In the computation of $\zeta_j(f, P_n)$, we need to evaluate $f(y_i, x_{i,-j})$ but it is not straightforward to integrate x_j out from the joint density (3.9). Hence, we apply a Monte Carlo approximation based on 500 random samples from $f(x_{i,j} \mid \theta_h)$ for each h.

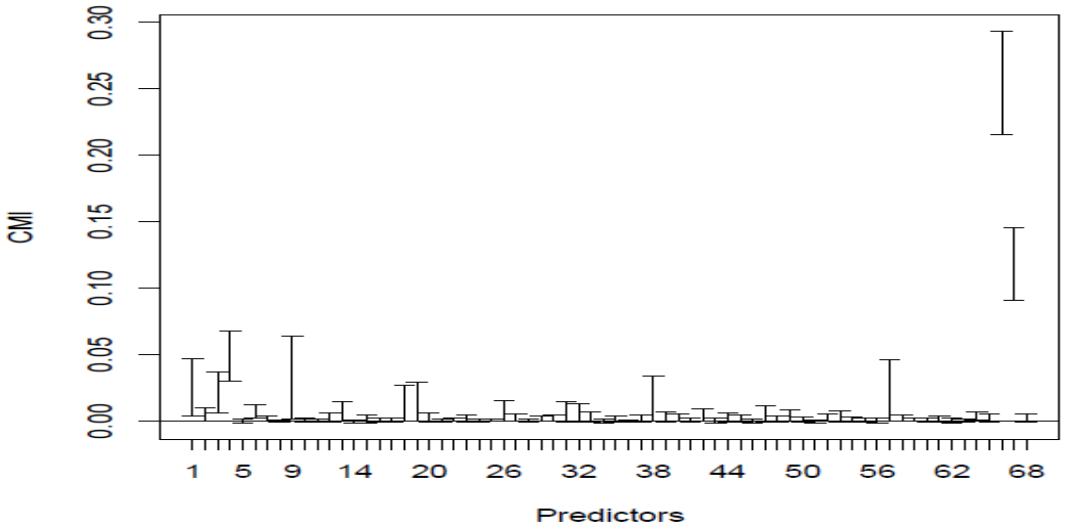


FIGURE 3.2: 90% credible intervals of the estimated conditional mutual information with murder as the response for each of the 68 demographic predictors adjusting for the other predictors.

Figure 3.2 shows 90% credible intervals of $\zeta_j(f, P_n)$ for all j and Table 3.2 reports the top 10 selected predictors in descending order of the posterior mean of conditional mutual information for murders. Full lists of the selected predictors for all responses are in the Appendix B. Certain predictors are selected for many different crime-related response variables. For all crimes, land area and population density show the first and second largest conditional dependence adjusting for other factors. Also, their posterior means of the conditional mutual information are much larger than those of other predictors especially in burglaries, larcenies, auto thefts and non-violent crimes. In addition, population in urban areas is selected 8 times, population, the percentage of kids with two parents and the percentage of persons in dense housing are picked up 7 times, and the percentage of Caucasian, the percentage of households with investment and rent income, the percentage of housing occupied and the percentage of families with two parents are conditionally dependent with 6 types of crimes. On the other hand, 12 predictors such as the percentage of housing units with less than 3 bedrooms and the percentage of moms of kids under 18 in labor force are not selected for any crimes.

Also, we can find similarities in the top 10 selected predictors among all crimes. We observe that certain types of variables obtain high ranks for many responses. For example, all crimes except larcenies and auto thefts share at least one of population in the community and population in urban areas in their lists. In addition, the percentage of families with parents and the percentage of kids with parents show relatively strong conditional dependence with all crimes other than murders, auto thefts and arsons. The posterior means of conditional mutual information of race variables are large for murders, robberies, assaults and violent crimes. Also, the top 10 lists of rapes, burglaries, arsons and non-violent crimes include more than one predictor related to divorce.

Table 3.2: Top 10 selected predictors in descending order of the posterior means of conditional mutual information with murders as the response. j, j-th predictor; Mean, posterior mean; 90% CI corresponds to a 90% credible interval.

j	Mean	90%CI	Predictor
66	0.2587	[0.2157, 0.2936]	land area in square miles
67	0.1188	[0.0905, 0.1454]	population density in persons per square mile
4	0.0507	[0.0302, 0.0678]	% of population that is caucasian
9	0.0250	[0.0043, 0.0636]	# of people living in areas classified as urban
1	0.0250	[0.0015, 0.0469]	population for community
3	0.0192	[0.0058, 0.0374]	% of population that is african american
57	0.0177	[0.00007, 0.0463]	rental housing: lower quartile rent
13	0.0075	[0.0004, 0.0149]	% of households with investment / rent income in 1989
6	0.0067	[0.0021, 0.0125]	% of population that is of hispanic heritage
64	0.0039	[0.0005, 0.0067]	% of people born in the same state as currently living

4

Nonparametric Bayes modeling with sample survey weights

4.1 Introduction

In sample surveys, it is routine to conduct stratified sampling designs to ensure that a broad variety of groups are adequately represented in the sample. In particular, the population is divided into mutually exclusive strata having different probabilities of inclusion. Analyzing data from such designs is challenging, since the collected sample is not representative of the overall population. To correct for discrepancies in the statistical analysis, survey weights are constructed. However, it is unclear how to appropriately include these weights, particularly in Bayesian analyses.

Little (2004) and Gelman (2007) clarify the importance of including survey weights into model-based analyses. Zheng and Little (2003, 2005) propose a nonparametric spline model and Chen et al. (2010) extend the framework for binary variables. Although these approaches can flexibly connect the survey weights with the response, they rely on the assumption of survey weights being known for all population units. Zangeneh and Little (2012) propose a modification to allow the number of non-

sampled units to be unknown. Si et al. (2014) instead propose a nonparametric model in which the survey weights are linked with a response through a Gaussian process regression. However, additional modeling of survey weights for non-sampled subjects in the population can lead to highly complex models.

In this article, we propose a simple approach in which we apply standard mixture models, such as Dirichlet process mixtures, for the selected sample, and then adjust the mixture weights based on the survey weights. We allow probabilistic uncertainty in this adjustment in a Bayesian manner. Posterior computation relies on a simple modification to add an additional step to Markov chain Monte Carlo algorithms for mixture models.

4.2 Mixture Models with Survey Weights

4.2.1 Adjusted density estimates

Let $y_1, \ldots, y_N$ denote independently and identically distributed observations from a superpopulation density f_0 with $y_i \in \mathcal{R}$ for $i \in D = \{1, \ldots, N\}$. From the finite population D, n subjects are sampled, with $w_i = c/\pi_i$ the survey weight for subject i, c a positive constant, and π_i the inclusion probability for $i \in D$. We assume D can be divided into mutually exclusive subpopulations $D_1, \ldots, D_M$, with $\{y_i, i \in D_m\}$ independently and identically distributed from density f_m, for $m = 1, \ldots, M$. Then, f_0 can be expressed as

$$f_0(y) = \sum_{m=1}^{M} \nu_m f_m(y), \tag{4.1}$$

where $\nu_m \geq 0$ and $\sum_{m=1}^{M} \nu_m = 1$. By applying kernel density estimation to each f_m in (4.1), Buskirk (1998) and Bellhouse and Stafford (1999) propose an adjusted density estimate,

$$\hat{f}_0(y) = \sum_{i \in S} \frac{\tilde{w}_i}{b} \mathcal{K}\left(\frac{y - y_i}{b}\right), \tag{4.2}$$

where $S \subset D$ are the selected subjects in the survey, $\tilde{w}_i = w_i / \sum_{j \in S} w_j$, $\mathcal{K}$ is a kernel function and $b > 0$. Estimator (4.2) adjusts for bias in the usual kernel estimator applied to sample S by modifying the weight for the ith subject from $1/n$ to $\tilde{w}_i$. This adjustment leads to consistency under some conditions (Buskirk and Lohr (2005)).

4.2.2 Bayesian adjustments with uncertainty

Section 2·1 focuses on univariate continuous variables, while our goal is to develop a general approach for adjusting posterior distributions to take into account sample survey weights. Let $y \in \mathcal{Y}$ denote a random variable, with $\mathcal{Y}$ a Polish space that may correspond to a p-dimensional Euclidean space, a discrete space, a mixed continuous and discrete space, a non-Euclidean Riemannian manifold, such as a sphere, and other cases. Extending (4.1) to general spaces, we let $f_0(\cdot)$ and $f_m(\cdot)$, for $m = 1, \ldots, M$, denote densities on $\mathcal{Y}$ with respect to a dominating measure μ. The density in the mth subpopulation is expressed as a mixture,

$$f_m(y) = \sum_{h=1}^{H} \nu_{mh} f(y \mid \theta_h), \tag{4.3}$$

where $\nu_{mh} \geq 0$, $\sum_{h=1}^{H} \nu_{mh} = 1$ and θ_h are parameters characterizing the hth mixture component. Then, f_0 can be approximately expressed as a mixture having the same kernels as in (4.3) but with adjusted weights as in (4.2).

Theorem 7. *Let $s_i \in \{1, \ldots, H\}$ denote the mixture index for subject i for $i \in S$. Let $S_h = \{i : s_i = h, i \in S\}$, for $h = 1, \ldots, H$. Then, for large N and n,*

$$f_0(y) \approx \sum_{h=1}^{H} \frac{\sum_{i \in S_h} w_i / c}{N} f(y \mid \theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y \mid \theta_{s_i}). \tag{4.4}$$

Proof. Letting N_m be the number of subjects in D_m, $N_m / N \to \nu_m$ as $N \to \infty$ by the law of large numbers. Letting w_m^* and π_m^* denote the survey weight and inclusion

probability for the mth subpopulation, $w_i = w_m^*$ and $\pi_i = \pi_m^*$ for $i \in D_m$. From (4.1) and (4.3), f_0 can be expressed as

$$f_0(y) = \sum_{m=1}^{M} \nu_m f_m(y) \approx \sum_{m=1}^{M} \frac{N_m}{N} f_m(y) = \sum_{h=1}^{H} \sum_{m=1}^{M} \frac{N_m \nu_{mh}}{N} f(y \mid \theta_h)$$

$$\approx \sum_{h=1}^{H} \frac{\sum_{i \in S_h} w_i/c}{N} f(y \mid \theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y \mid \theta_{s_i}). \tag{4.5}$$

The first approximation in (4.5) can be induced by $N_m \approx w_m^* n_m / c$ and

$$\nu_{mh} \approx \frac{\sum_{i \in S} 1(i \in D_{mh})}{n_m},$$

for large N_m and n_m, where $D_{mh} = \{i : s_i = h, i \in D_m\}$. The second approximation in (4.5) is based on $c \approx \sum_{i \in S} w_i / N$, which is derived by summation of $N_m \approx w_m^* n_m / c$ over m.

Under random designs with $w_i \propto c$, f_0 can be approximated by

$$f_R(y) = \sum_{h=1}^{H} \frac{\sum_{i \in D} 1(i \in S_h)}{n} f(y \,|\, \theta_h) = \sum_{i \in S} \frac{1}{n} f(y \mid \theta_{s_i}). \tag{4.6}$$

Comparing the last terms in (4.4) and (4.6), we can interpret that the bias can be adjusted by shifting the weight for the ith sampled subject from $1/n$ to $\tilde{w}_i$ as in (4.2).

We propose a simple Bayesian adjustment method using the second term in (4.4). We consider a standard Bayesian mixture model,

$$f_B(y) = \sum_{h=1}^{H} \lambda_h f(y \mid \theta_h), \quad \lambda \sim \pi(\lambda), \quad \theta_h \sim \pi(\theta_h), \tag{4.7}$$

where $\lambda = (\lambda_1, \ldots, \lambda_H)'$ with $\lambda_h \geq 0$ and $\sum_{h=1}^{H} \lambda_h = 1$, and $\pi(\lambda)$ and $\pi(\theta_h)$ are priors for λ and θ_h. For example, using a truncated stick-breaking process (Ishwaran and James (2001)), we let $\lambda_h = V_h \prod_{l<h}(1 - V_l)$, $V_h \sim \text{Beta}(1, \alpha)$ for $h = 1, \ldots, H-1$

with $V_H = 1$. However, our focus is not on the specific mixture model and prior but on the adjustment for sampling bias, and alternative priors can be used without complication.

Comparing the second terms in (4.4) and (4.6), the difference is in the mixture weights. The expression in (4.4) can be interpreted as implying that $\sum_{i \in S_h} w_i/c$ subjects are generated from the hth mixture component in the population. Updating the prior $\tilde{\lambda} \sim \text{Dirichlet}(a_1, \ldots, a_H)$ with this information, we obtain the following conditional posterior distribution for the adjusted weights $\tilde{\lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_H)'$,

$$\tilde{\lambda} \sim \text{Dirichlet}\left(a_1 + \frac{1}{\tilde{c}} \sum_{i:s_i=1} w_i, \ldots, a_H + \frac{1}{\tilde{c}} \sum_{i:s_i=H} w_i\right), \tag{4.8}$$

where $\tilde{c} = \sum_{i \in S} w_i/N \approx c$. In the equation (4.8), the sample size is enlarged from n to N based on the survey weights. Expression (4.8) takes into account uncertainty in the adjusted weights in mixture component allocation. Even as the population size N becomes large, there may be certain mixture components that are not represented in the selected sample, leading to substantial uncertainty in the adjustment. Posterior computation is straightforward: we simply apply any existing Markov chain Monte Carlo algorithm for mixture models to the selected sample, add sampling step (4.8) for generating the adjusted weights $\tilde{\lambda}$, and apply this adjustment to each step of the sampling algorithm to obtain samples from an adjusted posterior for the population density $f_0(y)$. As a default, we set $a_h = a$ for $h = 1, \ldots, H$, with prior sample size $Ha \sim 1 - 2\%$ of population size N.

4.3 Simulation Study

We illustrate performance of the proposed approach and compare to competitors. We consider three cases in which a population with $N = 1,000,000$ consists of three subpopulations having $N_1 = 650,000$, $N_2 = 300,000$ and $N_3 = 50,000$ with

$\nu_m = N_m/N$. From each stratum, we randomly generate $n_m = 500$ subjects and construct survey weights by $w_i = N_m/n_m$ for $i \in D_m$ for $m = 1, 2, 3$. As competitors, we employ three model-based Bayesian methods. First, we consider a model-based Horvitz-Thompson estimator (Horvitz and Thompson (1952); Little (2004)), $y_i = \beta\pi_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \pi_i^2\sigma^2)$ where $\pi_i = 1/w_i$. Second, we consider a polynomial regression with random effects, $y_i = \beta_0 + \beta_1\pi_i + \beta_2\pi_i^2 + \gamma_{[i]} + \varepsilon_i,\ \varepsilon_i \sim N(0, \pi_i^2\sigma^2)$, $\gamma_m \sim N(0, \tau^2)$ where $\gamma_{[i]}$ denotes a random effect for the subpopulation to which the ith subject belongs. This can be induced by the spline model of Zheng and Little (2003). Also, we apply the Gaussian process regression model from Si et al. (2014), $y_i = \mu(x_{[i]}) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),\ \mu(x) \sim \mathrm{GP}(\beta x, C),\ C\{\mu(x_m), \mu(x_{m'})\} = \mathrm{cov}\{\mu(x_m), \mu(x_{m'})\} = \tau^2 \exp(-\kappa|x_m - x_{m'}|)$ where $x_m = \log(w_m^*)$ and $x_{[i]}$ denotes the log weight for the stratum for the ith subject. For simplicity in modeling, additional information of the true size of non-sampled units for each stratum is given to the competitors while it is uncertain for the proposed method. We also apply Dirichlet process mixtures without weight adjustment.

In the first case, we assume $f_1(y) = f_N(y\,|2, 0.6),\ f_2(y) = f_N(y\,|0, 0.4)$ and $f_3(y) = f_N(y\,|-2, 0.3)$ in (4.1) where $f_N(y\,|\,a, b)$ denotes a normal density with mean a and standard deviation b. For the proposed method, we use the Dirichlet process mixture of normals, $f_B(y) = \sum_{h=1}^{H} \lambda_h f_N(y \mid \mu_h, \tau_h)$ where $\lambda_h = V_h(1 - V_l)$, $V_h \sim \mathrm{Be}(1, \alpha),\ V_H = 1$ with $H = 20,\ \alpha \sim \mathrm{Ga}(0.25, 0.25),\ \mu_h \sim N(\bar{y}, s_y^2),\ \tau_h^2 \sim$ Inverse-Gamma$(2, s_y^2/2)$ where $\bar{y}$ and s_y^2 are the sample mean and variance. As for the prior in the step (4.8), we set $a_h = 1,000$ for each h. For competitors, we assume the following priors: $\beta \sim N(0, s_y^2),\ \beta_j \sim N(0, s_y^2),\ \sigma^2 \sim$ Inverse-Gamma$(2, s_y^2/2)$, $\tau^2 \sim$ Inverse-Gamma$(2, s_y^2/2)$ and $\kappa \sim \mathrm{Ga}(1, 2)$. We draw 10,000 samples after the initial 5,000 samples are discarded as a burn-in period and every 10th sample is saved. Rates of convergence and mixing were adequate. Figure 4.1 shows the estimation results for case 1. The Horvitz-Thompson estimator fails to capture the

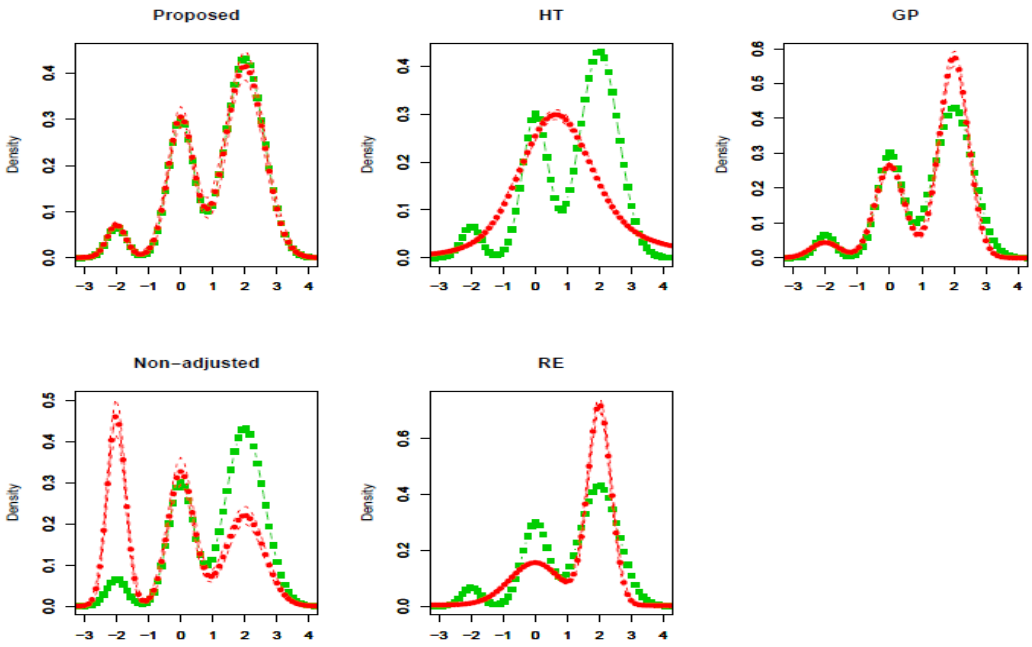


FIGURE 4.1: Estimated densities in case 1. Green lines with squares are the true density, red lines with circles the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Non-adjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

multimodality, while the non-adjusted estimator has considerable bias. The random effect model and Gaussian process have somewhat better performance, but clear bias remains. The proposed method accurately estimates the density, and 98% of true values are covered in the 95% credible intervals across 100 equally spaced grid points in [-6, 6].

We also considered a more complex density for each stratum, $f_1(y) = 0.2f_N(y\,|\,-2,1) + 0.8f_N(y\,|\,2,0.8)$, $f_2(y) = 0.4f_N(y\,|\,-2,1) + 0.6f_N(y\,|\,2,0.8)$ and $f_3(y) = 0.85f_N(y\,|\,-2,1) + 0.15f_N(y\,|\,2,0.8)$. The Markov chain Monte Carlo settings are the same as in case 1. Figure 4.2 reports the result for case 2. The Horvitz-Thompson estimator, random effect model and Gaussian process regression work poorly, missing the multimodal shape of the true density because they construct population densities

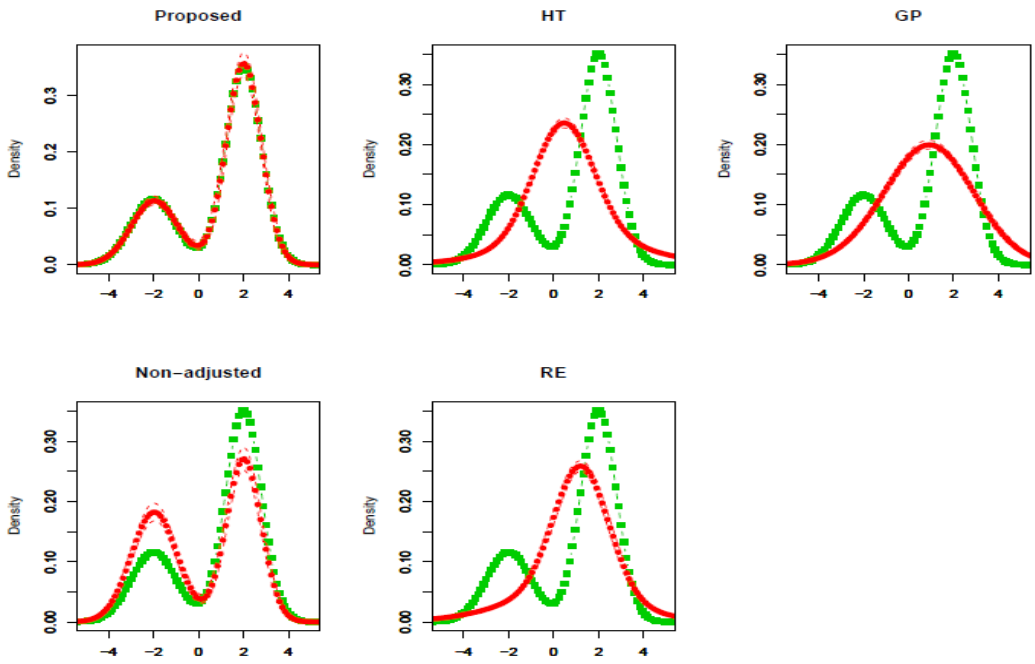


FIGURE 4.2: Estimated densities in case 2. Green lines with squares are the true density, red lines with circles the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Non-adjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

relying on unimodal densities for subpopulations. The non-adjusted method capture the bimodality but with substantial bias. The proposed method approximates the density well, while covering 100% of true values in the 95% intervals.

We also consider a mixture of Poisson distributions, $f_1(y) = 0.2\text{Poisson}(y\,|15) + 0.8\text{Poisson}(y\,|4)$, $f_2(y) = 0.4\text{Poisson}(y\,|15) + 0.6\text{Poisson}(y\,|4)$ and $f_3(y) = 0.85\text{Poisson}(y\,|15) + 0.15\text{Poisson}(y\,|4)$. For the Dirichlet process mixtures, we apply the rounded kernel method in Canale and Dunson (2011) where latent continuous variables are modeled by (4.7) with the same Markov chain Monte Carlo settings as in case 1. Also, we apply the competitors to log transformed observations $y_i^* = \log(y_i + 0.5)$ and estimate probabilities by $\text{pr}(y_i = y) = \text{pr}\{\log(y) < y_i^* \leq \log(y+1)\}$ for $y = 0, 1, \ldots, \infty$. Figure 4.3 shows the result. We observe the proposed

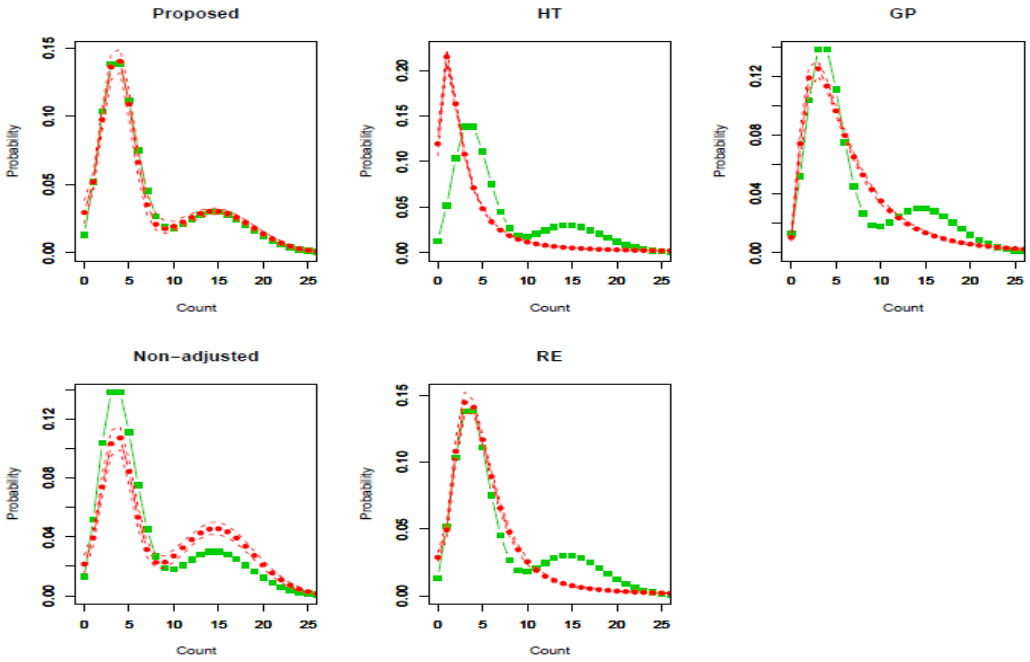


FIGURE 4.3: Estimated probabilities in case 3. Green lines with squares are the true density, red lines with circles the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Non-adjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

method obtains good approximation, while the competitors fail to capture the mode at 15. Also, 98% of the true values are covered in the 95% intervals in the support from 0 to 100.

To assess the impact of increasing the number of strata while decreasing within-strata sample size, we consider a case with $M = 100$ in which $N_m = 1000m$, $n_m = 20$ for $m = 1, \ldots, 100$ with $N = 5,050,000$ and $n = 2,000$ and $f_m(y) = f_N(y \mid -2, 0.3)$ for $m = 1, \ldots, 30$, $f_m(y) = f_N(y \mid 0, 0.4)$ for $m = 31, \ldots, 70$ and $f_m(y) = f_N(y \mid 2, 0.6)$ for $m = 71, \ldots, 100$. We obtain a similar result to case 1 with the proposed method dominating competitors.

4.4 Application to Adolescent Behaviour Analysis

We apply the proposed method to the National Longitudinal Study of Adolescent Health. Our focus is on studying the total number of sex partners in adolescence. The target population is adolescents in grades 7-12 in the United States during the 1994-95 school year with $N = 14,677,347$. The full study design is described by Harris et al. (2009). The study drew supplemental samples, oversampling groups of particular interest based on ethnicity, genetic relatedness to siblings, adoption status, disability, and black adolescents with highly educated parents. We use three waves of surveys in which participants are in grades 7-12 (1994-1995), young adults age 18-26 (2001-2002) and adults age 24-32 (2007-2008). In each wave, numbers of observations are 6447, 4812 and 4819, respectively. We use the rounded kernel method with Dirichlet process mixtures as in the simulation. Since we expect high right skew in these data, we use log cut-points instead of non-negative integers, so that the Dirichlet process mixtures can efficiently approximate such distributions. For the priors of the latent continuous variable, we use $\mu_h \sim N(\tilde{y}, \tilde{s}_y^2)$, $\tau_h^2 \sim$ Inverse-Gamma$(2, \tilde{s}_y^2/200)$ where $\tilde{y}$ and $\tilde{s}_y^2$ are the sample mean and variance of $\log(y_i + 0.5)$. Also, we set $a_h = 10,000$ for the Dirichlet prior in (4.8). We draw 20,000 samples after the initial 5,000 samples are discarded as a burn-in period and every 10th sample is saved. We observe that the sample paths were stable and the sample autocorrelations dropped smoothly.

Figure 4.4 shows the estimated probabilities for the three waves. 1994-1995 shows a high probability on zero with small values for positive counts. 2001-2002 expresses differences from 1994-1995 in that the probability on zero considerably decreases, while one shows the highest value and the tail gets heavy. The shape in 2007-2008 is similar to 2001-2002 in that both have highest probabilities at one and then steep declines. 2007-2008 shows a heavier tail with relatively high spikes at multiples of

five. This is probably because people with many partners do not remember the exact numbers.

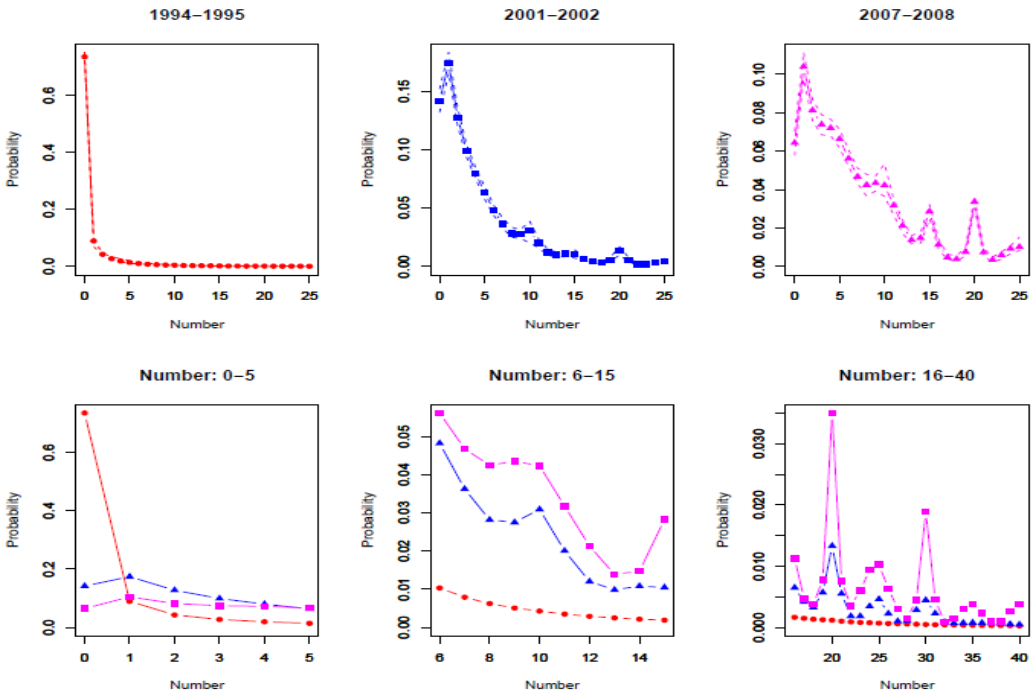


FIGURE 4.4: Estimated probabilities of total numbers of sex partners. The first row shows estimated probabilities in 1994-1995 (left), 2001-2002 (middle) and 2007-2008 (right). Lines with symbols show posterior means and dash lines 95% credible intervals. The second row shows posterior means of 0-5 partners (left), 6-15 (middle) and 15-40 (left). Red lines with circle represent posterior means for 1994-1995, blue lines with triangles for 2001-2002 and purple lines with squares for 2007-2008.

5

Nonparametric Bayes models for mixed-scale longitudinal surveys

5.1 Introduction

It is routine in social science that mixed-scale data are collected over time for longitudinal studies. As a motivating application, we consider social surveys that follow up individuals for long periods to study trends of their social behaviours and health conditions via diverse types of questions. The National Longitudinal Study of Adolescent to Adult Health (Add Health) has collected vast amounts of information of social, economic, and health-related environments and behaviours in adolescence over 20 years. Our primary interest is in studying associations among the adolescent sexual behaviours and investigating their trajectories from adolescence into adulthood. Also, it is of interest to compare trends of the interactions in the overall population with those in sub-populations characterized by biological backgrounds. However, the sexual behaviour data in Add Health have characteristics for which it is challenging to build a flexible but not too complicated statistical model.

First, the sexual behaviour data consist of mixed-scale variables. It is not straight-

forward to jointly model mixed-scale multivariate data, especially if they include both ordered variables and nominal variables. Second, individuals repeatedly answered a set of questions over time, leading to subject-specific time dependence. Also, the time interval can differ from person to person. Third, Add Health collects sample via a stratified sampling design in which the population is divided into mutually exclusive strata, each of which has a different probability of inclusion. Then, the resulting data is a non-representative sample of the population. Fourth, with respect to small area estimation, the number of subjects in certain sub-population can be relatively small, leading to an unstable estimation result. Hence, we need to borrow information over sub-populations instead of constructing a statistical model separately for each sub-population. Fifth, there are massive number of missing values in the data. We observe both design-based and individual-specific missingness in the data. Therefore, we need to conduct statistical analysis of the adolescent sexual behaviour, taking into account all of these challenging characteristics in the Add Health data.

There is a rich literature on modeling of mixed-scale data. One approach is to apply generalized linear models for each outcome in which dependence among responses is induced through shared latent factors (Sammel et al. (1997); Dunson (2000); Moustaki and Knott (2000)). However, the robustness of the approaches based on generalised linear mixed models can be weak due to the duel role of the random effects structure in controlling the dependence and shape of marginal distributions. Another approach is to use underlying continuous variables, specified by a Gaussian model (Muthén (1984)) or a Dirichlet process mixtures (Kottas (2005); Canale and Dunson (2011); Kim and Ratchford (2013)). In this approach, discrete variables can be expressed by thresholding the latent continuous variables. In addition, avoiding specification of marginal distributions, Hoff (2007) proposed a semiparametric Gaussian copula model in which the associations among variables

can be expressed as correlations among the latent Gaussian variables. Murray et al. (2013) and Gruhl et al. (2013) extend the approach in Hoff (2007) by incorporating factor structures but these copula models can incorporate only ordered variables. Also, Murray and Reiter (2014) propose a nonparametric Bayesian joint model for multiple imputation of missing values. McParland et al. (2014) develop a model-based clustering approach for mixed-scale data which combines item response theory models for ordered variables with factor analysis models for nominal variables in the framework of Bayesian mixtures.

Recently, a number of articles have worked on the analysis of multivariate longitudinal data (Bandyopadhyay et al. (2011), Verbeke et al. (2014)). It is routine to incorporate time effects into multivariate models through time-varying covariates such as polynomial functions of age with random coefficients (Gueorguieva and Sanacora (2006); Fieuws and Verbeke (2006); Luo and Wang (2014); Baghfalakia et al. (2014)). Dunson (2003) proposed dynamic latent trait models in which autoregressive Gaussian latent factors induce subject-specific time-dependence and generalized linear models describe mixed-scale outcomes. Ghosh and Hanson (2010) propose a semiparametric approach with a mixture of Polya trees prior for random effect distributions and B-splines for time effects. Liu et al. (2009) develop a joint model for longitudinal binary and continuous variables which consists of a marginal binary model and a conditional regression model relying on the Bartlett decomposition of a covariance matrix. However, it is not clear how to handle survey bias, missingness and small area estimation in these approaches. Therefore, none of these approaches can capture the exact nature of longitudinal surveys in Add Health.

In the literature on the survey data analysis, there are two major methods: design-based approach, which treats outcomes as fixed quantities and model-based approach, which models outcomes and predicts values for the non-sampled subjects in population (Little (2004); Levy and Lemeshow (2008); Rao (2011)). Little (2004) and

Gelman (2007) show difficulties with current methods in practice and point out the importance of including survey weights into model-based analyses. Zheng and Little (2003, 2005) propose a nonparametric method which flexibly models the outcome given inclusion probability using penalized spline and Chen et al. (2010) extend the model for binary variables. Si et al. (2014) propose a nonparametric Bayesian model which jointly models outcome and survey weights and the interaction is flexibly induced by a Gaussian process regression. However, these approaches are developed mainly for estimation of the population mean of a univariate static outcome. Hence, it is not easy to extend them for mixed-scale longitudinal data.

In this article, we propose a flexible nonparametric Bayesian joint model for the analysis of Add Health sexual behaviour data in which mixed-scale variables are expressed by transforming latent continuous variables. The Dirichlet process mixture of Gaussian factor models are developed for the latent continuous variables. For unordered categorical variables, we employ the concept of utilities in multinomial probit models where the nominal variable is a manifestation of underlying continuous utility variables. The subject-specific dynamic variability can be captured by time-varying latent factors via Gaussian processes which can easily incorporate irregular time intervals for each subject. Also, the bias from sampling designs can be collected by adjusting the mixture weights in the Dirichlet process mixtures relying on the survey weights. Since we build a joint model of the response variables and covariates, we can easily impute missing values assuming missing at random. In addition, we can obtain a density for population-level inferences by integrating out covariates from the joint model. Also, from the joint density, we can construct densities conditional on covariates for sub-populations of interest. For posterior computation, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm in which we modify the mixture weights of the Dirichlet process mixture model taking into account uncertainty in the adjustment process.

Section 5.2 describes our data set from Add Health. Section 5.3 proposes a novel approach for mixed-scale longitudinal surveys. Section 5.4 develops an efficient MCMC algorithm. Section 5.5 applies the proposed method to the adolescent sexual behaviour analysis.

5.2 National Longitudinal Study of Adolescent to Adult Health

Add Health is a nation-wide longitudinal study of adolescents in grades 7-12 in the United States during the 1994-95 school year (http://www.cpc.unc.edu/projects /addhealth). A stratified sampling design is utilized for collection of data. Also, the study drew supplemental samples, oversampling groups of particular interest based on ethnicity, genetic relatedness to siblings, adoption status, disability, and black adolescents with highly educated parents. To collect for bias from the survey design, survey weights are constructed for cross section studies and longitudinal studies respectively. The full study design is described by Harris et al. (2009).

Our primary goal is studying associations among sexual behaviours in adolescence and their trends in the transition to adulthood. We use public-released data from three waves of surveys which are conducted when participants are in grades 7-12 (Wave 1: 1994-1995), young adult aged around 18-26 (Wave 3: 2001-2002) and adult aged around 24-32 (Wave 4: 2007-2008). We select four variables as responses; attraction to opposite sex (binary, 1: Yes, 0: No), attraction to same sex (binary, 1: Yes, 0: No), cumulative numbers of partners with all types of sexual relationships (count) and sexual self definition (unordered categorical, 1. 100% heterosexual, 2. mostly heterosexual, 3. bisexual, 4. mostly homosexual, 5. 100% homosexual, 6. no sexual attraction). The sexual self definition is designed-based missing in Wave 1. Figure 5.1 shows histogram of the response variables and age in each wave for longitudinal studies with $n = 4,208$ subjects. The attraction to opposite/same sex and sexual self definition roughly look similar over time. On the other hand, the

number of sex partners clearly change over these waves of surveys. We observe a big spike at zero in 1994-95, then more probabilities are assigned to positive counts in 2001-02. 2007-08 shows a heavier tail with relatively high spikes at multiples of five.

Also, we have interests in comparing trends of the associations in population to those in sub-populations with different biological backgrounds. As predictors, we consider gender (binary, 1: female, 0: male) and race (unordered categorical, 1: Caucasian, 2: African American, 3: Others). Figure 5.2 shows the mosaic plot of the covariates. Because of the sampling design, the ratios of these boxes do not correspond to those in the population. For example, the percentage of black people 24% is much higher than around 12% in the population level. For longitudinal studies with these three waves, the survey weights are constructed for $n = 4,208$ subjects given in Figure 5.3. Since a survey weight implies the number of people the corresponding observed subject represents, there can be large bias in the estimation result without incorporating the weight into statistical models.

5.3 Proposed modeling of mixed-scale longitudinal surveys

5.3.1 Modeling of mixed-scale data

Let $y = (y_1, \ldots, y_p)'$ be a response variable. Each univariate y_k is one of binary, count and nominal variables for $k = 1, \ldots, p$. Let $y^* = (y_1^*, \ldots, y_{p^*}^*)' \in \mathbb{R}^{p^*}$ be a latent multivariate continuous variable and we express the response y by transforming this underlying variable y^*. Define $[k] \subset \{1, \ldots, p^*\}$ as a set of indices such that y_k is induced by $y_{[k]}^* = \{y_r^*, r \in [k]\}$. A binary variable $y_k \in \{0, 1\}$ can be induced by

$$\text{binary}: \quad y_k = 1(y_{[k]}^* > 0), \quad y_{[k]}^* \in \mathbb{R},$$

where $1(\cdot)$ denotes an indicator function. A count variable $y_k \in \{0, \cdots, \infty\}$ can be expressed as

$$\text{count}: \quad y_k = \sum_{r=0}^{\infty} r \times 1(a_r < y_{[k]}^* \leq a_{r+1}), \quad y_{[k]}^* \in \mathbb{R},$$

where $-\infty = a_0 < a_1 < a_2 < \cdots < \infty$. For a nominal variable with d categories $y_k \in \{1, \cdots, d\}$, we introduce a d-dimensional vector $y^*_{[k]} = (y^*_{k_1}, \ldots, y^*_{k_d})' \in \mathbb{R}^d$ and set

$$\text{unordered categorical :} \quad y_k = \sum_{r=1}^{d} r \times 1(y^*_{k_r} = \max_{1 \le l \le d} y^*_{k_l}). \tag{5.1}$$

In the expression (5.1), $y^*_{k_l}$ can be interpreted as the utility for the lth category as in multinomial probit models (McCulloch and Rossi (1994); Imai and van Dyk (2005); Burgette and Nordheim (2012); Johndrow et al. (2013)). Since the order of $y^*_{j_1}, \ldots, y^*_{j_d}$ is unchanged with respect to adding any constant and multiplying any positive scale to all of them, it is common to assume one of the utilities is zero.

5.3.2 Proposed nonparametric Bayesian joint modeling

We apply the above framework to longitudinal data with mixed-scale margins. Let $y_{ij} = (y_{ij1}, \ldots, y_{ijp})'$ be a response variable for subject i at age t_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ with $y_{ijk} \in \mathcal{Y}_k$ allowed to be any type of univariate random variable for $k = 1, \ldots, p$. As in the previous subsection, we introduce $y^*_{ij} = (y^*_{ij1}, \ldots, y^*_{ijp^*})' \in \mathbb{R}^{p^*}$ as a latent continuous variable which induces y_{ij} through a function g such that $y_{ij} = g(y^*_{ij})$. Also, $x_i = (x_{i1}, \ldots, x_{iL})'$ denotes static covariates for ith subject.

Then, we develop a flexible joint model of the response and covariates, in which both the mean and covariance of the response can flexibly change over age and covariates. The joint modeling allows us to easily construct conditional densities for sub-groups based on the covariates and impute missing values assuming missing at random. Let $\mu_t = (\mu_{t1}, \ldots, \mu_{tQ_\mu})'$ be a time-effect vector for age t and $\eta_{ij} = (\eta_{ij1}, \ldots, \eta_{ijQ})'$ denotes a time-varying factor for subject i at age t_{ij} where $Q_\mu < p^*$ and $Q < p^*$. We first consider modeling of the response conditional on the covariates,

$$y^*_{ij} = Bx_i + \Omega\mu_{t_{ij}} + \Lambda\eta_{ij} + \epsilon_{ij}, \ \epsilon_{ij} \sim N(0, \Sigma), \ \ \Sigma = \text{diag}(\sigma^2_1, \ldots, \sigma^2_{p^*}), \tag{5.2}$$

where B is a $p^* \times L$ matrix of regression coefficients, Ω is a $p^* \times Q_\mu$ coefficient matrix for time effects, Λ is a $p^* \times Q$ factor loading matrix. In this model, we can capture age-specific effects for all individuals by μ_t and additional subject-specific variability is induced by the time-varying factor η_{ij}. Relying on covariance regressions by Hoff and Niu (2012) and Fox and Dunson (2011), we model the dynamic random effects,

$$\eta_{ij} = V x_i \eta_i^* + \xi_{t_{ij}} \tilde{\eta}_i, \quad \eta_i^* \sim N(0,1), \quad \tilde{\eta}_i \sim N(0, I_{Q_\eta}), \tag{5.3}$$

where η_i^* and $\tilde{\eta}_i$ are static subject-specific effects, V is a $Q \times L$ coefficient matrix for covariates and $\xi_t = \{\xi_{tql}\}$ is a $Q \times Q_\eta$ matrix for $t = 1, \ldots, T$. After integrating out the static random effects, we obtain a distribution with a covariance dependent on time and covariates,

$$y_{ij}^* \sim N(Bx_i + \Omega \mu_{t_{ij}}, \Lambda V x_i x_i' V' \Lambda' + \Lambda \xi_{t_{ij}} \xi_{t_{ij}}' \Lambda' + \Sigma). \tag{5.4}$$

For the time effects μ_t and ξ_t, we apply Gaussian processes,

$$\mu_{\cdot q} \sim \mathrm{GP}(0, c_\mu), \quad c_\mu(\mu_{tq}, \mu_{t'q}) = \exp(-\kappa_\mu |t - t'|^2), \tag{5.5}$$

$$\xi_{\cdot ql} \sim \mathrm{GP}(0, c_\xi), \quad c_\xi(\xi_{tql}, \xi_{t'ql}) = \exp(-\kappa_\xi |t - t'|^2), \tag{5.6}$$

where $\kappa_\mu > 0$, $\kappa_\xi > 0$ and $\mu_{\cdot q}$ and $\xi_{\cdot ql}$ are mutually independent with respect to q and l.

Since social science data often contain complex structures, we incorporate additional flexibility into the model (5.2) and (5.3) relying on Dirichlet process mixtures (Lo (1984); West et al. (1994); Escobar and West (1995); Müller et al. (1996)). Let $y_i = \{y_{ij},\ j \in \{1, \ldots, n_i\}\}$ and $\theta = \{B, \Omega, \Lambda, \Sigma, V, \phi\}$ where ϕ is a parameter set for x_i. Using the stick-breaking representation of the Dirichlet process mixtures (Sethuraman (1994); Muliere and Tardella (1998)), the joint density of y_i and x_i can be

expressed as

$$f(y_i, x_i) = \int_{R(y_i)} f(y_i^*, x_i \mid \eta_i^*, \tilde{\eta}_i) f(\eta_i^*) f(\tilde{\eta}_i) d\eta_i^* d\tilde{\eta}_i dy_i^*, \tag{5.7}$$

$$f(y_i^*, x_i \mid \eta_i^*, \tilde{\eta}_i) = \sum_{h=1}^{\infty} \pi_h \prod_{j=1}^{n_i} f(y_{ij}^* \,|\, x_i, \eta_i^*, \tilde{\eta}_i, \theta_h) \prod_{l=1}^{L} f_l(x_{il} \,|\, \phi_h), \tag{5.8}$$

$$\pi_h = v_h \prod_{b<h} (1 - v_b), \quad v_h \sim \text{Beta}(1, \alpha), \tag{5.9}$$

where $R(y_i) = \{y_i^* \in \mathbb{R}^{p^* \times n_i} : y_{ij} = g(y_{ij}^*),\ j = 1, \ldots, n_i\}$ and $f(y_{ij}^* \,|\, x_i, \eta_i^*, \tilde{\eta}_i, \theta_h)$ denotes a density from (5.2) and (5.3). With respect to the density $f_l(x_{il} \,|\, \phi)$ in (5.8), we assume a function depending on the type of the covariate such as multinomial for a categorical variable. The model (5.8)-(5.9) corresponds to the Dirichlet process mixture of Gaussian factor models.

5.4 Posterior computation

Relying on the blocked Gibbs sampler by Ishwaran and James (2001) with H mixture components, we develop an efficient posterior computation for the proposed model. We apply shrinkage priors to $U_h \equiv [B_h\ \Omega_h\ \Lambda_h]$ with high density at zero to exclude redundant covariates, time effects and factors from the model while the tails are heavy enough to capture important signals. Let U_{hkl} be the (k, l) component in U_h for $k = 1, \ldots, p^*$ and $l = 1, \ldots, L + Q_\mu + Q$. Then, we assume $U_{hkl} \sim N(0, \delta_{kl}^2)$, $\delta_{kl}^2 \sim \text{IG}(0.5, 0.5)$ where IG denotes an inverse-gamma distribution. After integrating out δ_{kl}^2, the prior corresponds to a Cauchy prior which is a commonly used shrinkage prior with heavy tails. We also apply the same type of shrinkage prior to V_h. Then, we propose the following MCMC algorithm.

1. Update v_h for $h = 1, \ldots, H - 1$ from

$$\text{Beta}\left(1 + \sum_{i=1}^{n} 1(s_i = h), \alpha + \sum_{l>h} \sum_{i=1}^{n} 1(s_i = l)\right),$$

where s_i is a latent mixture indicator for subject i.

2. Using the prior Gamma(a_α, b_α), update α from

$$\text{Gamma}\left(a_\alpha + H - 1, b_\alpha - \sum_{h=1}^{H-1} \log(1 - v_h)\right).$$

3. Update s_i for $i = 1, \ldots, n$ from

$$P(s_i = h \mid \cdots) = \frac{\pi_h \prod_{j=1}^{n_i} f(y_{ij}^* \mid x_i, \eta_i^*, \tilde{\eta}_i, \theta_h) \prod_{l=1}^{L} f(x_{il} \mid \phi_h)}{\sum_{m=1}^{H} \pi_m \prod_{j=1}^{n_i} f(y_{ij}^* \mid x_i, \eta_i^*, \tilde{\eta}_i, \theta_m) \prod_{l=1}^{L} f(x_{il} \mid \phi_m)}.$$

4. Using the prior IG$(\tilde{a}_k, \tilde{b}_k)$, update σ_{hk}^2 for $k = 1, \ldots, p^*$ and $h = 1, \ldots, H$ from

$$\text{IG}\left(\frac{\sum_{i:s_i=h} n_i + \tilde{a}_k}{2}, \frac{\sum_{i:s_i=h} \sum_{j=1}^{n_i} \tilde{y}_{ijk}^2 + \tilde{b}_k}{2}\right),$$

where $\tilde{y}_{ijk}$ is the kth component of $\tilde{y}_{ij} \equiv y_{ij}^* - B_h x_i - \Omega_h \mu_{t_{ij}} - \Lambda_h \eta_{ij}$.

5. Update the kth column $u_{hk} = (u_{hk1}, \ldots, u_{hkL^*})'$ in U_h with $L^* = L + Q_\mu + Q$ for $k = 1, \ldots, p^*$ and $h = 1, \ldots, H$ from $N(\mu_u^*, \Sigma_u^*)$ where

$$\mu_u^* = \Sigma_u^* \left(\sigma_{hk}^{-2} \sum_{i:s_i=h} \sum_{j=1}^{n_i} z_{ij} y_{ijk}^*\right), \quad \Sigma_u^* = \left(\sigma_{hk}^{-2} \sum_{i:s_i=h} \sum_{j=1}^{n_i} z_{ij} z_{ij}' + \Sigma_0^{-1}\right)^{-1},$$

where $z_{ij} = (x_i', \mu_{t_{ij}}', \eta_{ij}')'$ and $\Sigma_0 = \text{diag}(\delta_{k1}^2, \ldots, \delta_{kL^*}^2)$.

6. Update $\tilde{\eta}_i$ from $N(\mu_{\tilde{\eta}}, \sigma_{\tilde{\eta}}^2)$ for $i = 1, \ldots, n$ where

$$\mu_{\tilde{\eta}} = \sigma_{\tilde{\eta}}^2 x_i' V_{s_i}' \Lambda_{s_i}' \Sigma_{s_i}^{-1} \sum_{j=1}^{n_i} \tilde{y}_{ij}, \quad \sigma_{\tilde{\eta}}^2 = \left(n_i x_i' V_{s_i}' \Lambda_{s_i}' \Sigma_{s_i}^{-1} \Lambda_{s_i} V_{s_i} x_i + 1\right)^{-1},$$

where we set $\tilde{y}_{ij} = y_{ij}^* - B_{s_i} x_i - \Omega_{s_i} \mu_{t_{ij}} - \Lambda_{s_i} \xi_{t_{ij}}$.

7. Update η_i^* from $N(\mu_{\eta^*}^*, \Sigma_{\eta^*}^*)$ for $i = 1, \ldots, n$ where

$$\mu_{\eta^*}^* = \Sigma_{\eta^*}^* \left(\sum_{j=1}^{n_i} \xi_{t_{ij}}' \Lambda_{s_i}' \Sigma_{s_i}^{-1} \tilde{y}_{ij}\right), \quad \Sigma_{\eta^*}^* = \left(\sum_{j=1}^{n_i} \xi_{t_{ij}}' \Lambda_{s_i}' \Sigma_{s_i}^{-1} \Lambda_{s_i} \xi_{t_{ij}} + I\right)^{-1},$$

where we set $\tilde{y}_{ij} = y_{ij}^* - B_{s_i} x_i - \Omega_{s_i} \mu_{t_{ij}} - \Lambda_{s_i} V_{s_i} x_i \tilde{\eta}_i$.

8. Update κ_μ using the Griddy-Gibbs sampler by Ritter and Tanner (1992) from

$$P(\kappa_\mu = g_k \,|\, \cdots) = \frac{\prod_{q=1}^{Q_\mu} f_N(\mu_q \,|\, 0, C(g_k))}{\sum_{l=1}^{G} \prod_{q=1}^{Q_\mu} f_N(\mu_q \,|\, 0, C(g_l))},$$

where $g_1, \ldots, g_G$ are G grid points, $\mu_q = (\mu_{1q}, \ldots, \mu_{Tq})'$ and $C(g_k)$ is a covariance matrix in GP with the length-scale g_k.

9. Update κ_ξ using the Griddy-Gibbs sampler from

$$P(\kappa_\xi = g_k \,|\, \cdots) = \frac{\prod_{q=1}^{Q} \prod_{l=1}^{Q^*} f_N(\xi_{ql} \,|\, 0, C(g_k))}{\sum_{j=1}^{G} \prod_{q=1}^{Q} \prod_{l=1}^{Q^*} f_N(\xi_{ql} \,|\, 0, C(g_j))},$$

where $\xi_{ql} = (\xi_{1ql}, \ldots, \xi_{Tql})'$.

10. Update V_h for $h = 1, \ldots, H$ by generating lth column of V_h for $l = 1, \ldots, L$ from $N(\mu_v, \Sigma_v)$ with

$$\mu_v = \Sigma_v \left(\sum_{i:s_i=h} \sum_{j=1}^{n_i} x_i \Lambda_h' \Sigma_h^{-1} \tilde{y}_{ij} x_{il} \tilde{\eta}_i \right), \quad \Sigma_v = \left(\sum_{i:s_i=h} n_i \Lambda_h' \Sigma_h^{-1} \Lambda_h x_{il}^2 \tilde{\eta}_i^2 + \Sigma_{0v}^{-1} \right)^{-1},$$

where we set $\tilde{y}_{ij} = y_{ij}^* - B_h x_i - \Omega_h \mu_{t_{ij}} - \Lambda_h \xi_{t_{ij}} \eta_i^* - \Lambda_h V_{h,-l} x_{i,-l} \tilde{\eta}_i$ where $V_{h,-l}$ is a submatrice of V_h with lth column removed and $\Sigma_{0v} = \mathrm{diag}(\zeta_{1l}^2, \ldots \zeta_{Ql}^2)$

11. Update $\mu_{\cdot q} = (\mu_{1q}, \ldots, \mu_{Tq})'$ for $q = 1, \ldots, Q_\mu$ from $N(\mu_\mu, \Sigma_\mu)$ with

$$\mu_\mu = \Sigma_\mu A z_{iq}^*, \quad \Sigma_\mu = \left(A + K_\mu^{-1} \right)^{-1},$$

where $z_{iq}^* = (b_{1q}, \ldots, b_{Tq})'$ with $b_{tq} = \sum_{(i,j):t_{ij}=t} \Omega_{s_i,q}' \Sigma_{s_i}^{-1} \tilde{y}_{ij}$, $A = \mathrm{diag}(a_{1q}, \ldots, a_{Tq})$ with $a_{tq} = \sum_{(i,j):t_{ij}=t} \Omega_{s_i,q}' \Sigma_{s_i}^{-1} \Omega_{s_i,q}$, $\tilde{y}_{ij} = y_{ij}^* - B_h x_i - \Lambda_h \eta_{ij} - \Omega_{h,-q} \mu_{t,-q}$ for subject i with $s_i = h$ and $h = 1, \ldots, H$, $\Omega_{h,q}$ is qth column of Ω_h, $\Omega_{h,-q}$ is a submatrix of Ω_h with qth column removed.

12. Update $\xi_{\cdot ql} = (\xi_{1ql}, \ldots, \xi_{Tql})'$ for $q = 1, \ldots, Q_\mu$ and $l = 1, \ldots, L$ from $N(\mu_{\xi ql}, \Sigma_{\xi ql})$ with

$$\mu_{\xi ql} = \Sigma_{\xi ql} \left(\sum_{(i,j):t_{ij}=t} \Lambda_{s_i,q}' \Sigma_{s_i}^{-1} \tilde{y}_{ij} \eta_{il}^* \right), \quad \Sigma_{\xi ql} = \left(R_\xi + K_\xi^{-1} \right)^{-1},$$

where $R_\xi = \text{diag}(r_1, \ldots, r_p)$ with $r_k = \sum_{(i,j):t_{ij}=t} \Lambda_{s_i,q} \Sigma_{s_i}^{-1} \Lambda_{s_i,q} \eta_{il}^{*2}$, $\tilde{y}_{ij} = y_{ij}^* - B_{s_i} x_i - \Omega_{s_i} \mu_{t_{ij}} - \Lambda_{s_i} V_{s_i} x_i \tilde{\eta}_i - c_{ql}$ with $c_{ql} = (c_{1ql}, \ldots, c_{pql})'$ and $c_{kql} = \sum_{(q',l') \neq (q,l)} \Lambda'_{s_i k q'} \xi_{t q' l'} \eta_{il'}^*$.

13. Update ϕ_{hl} for $h = 1, \ldots, H$ and $l = 1, \ldots, L$ from

$$f(\phi_{hl} \mid \cdots) \propto \prod_{i:s_i=h} f(x_{il} \mid \phi_{hl}) \pi(\phi_{hl}),$$

where $\pi(\phi_{hl})$ is the prior density. The form of the posterior depends on the type of x_{itl}. For example, for a categorical variable, the posterior is a Dirichlet distribution with a conjugate prior.

14. Impute missing values x_{il} from

$$f(x_{il} \mid \cdots) \propto \prod_{j=1}^{n_i} f(y_{ij}^* \mid x_i, \eta_i^*, \tilde{\eta}_i, \theta_{s_i}) f(x_{il} \mid \phi_{s_i}).$$

15. Impute missing values y_{ijk}. Let $B_{h[k]}$, $\Omega_{h[k]}$, $\Lambda_{h[k]}$ and $\Sigma_{h[k]}$ be submatrices of B_h, Λ_h and Σ_h related to the kth response. Then,

    1. Generate $y_{ij[k]}^* \sim N(B_{s_i[k]} x_i + \Omega_{s_i[k]} \mu_{t_{ij}} + \Lambda_{s_i[k]} \eta_{ij}, \Sigma_{s_i[k]})$.
    2. Set $y_{ijk} = g_j(y_{ij[k]}^*)$.

16. Update y_{ijk}^* for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, p^*$ from the truncated normal distribution in which the constraint comes through y_{ijm} with $k \in [m]$,

$$f(y_{ijk}^* \mid \cdots) \propto f(y_{ijm} \mid y_{ijk}^*) f_N(y_{ijk}^* \mid b_{s_i k} x_i + \lambda_{s_i k} \eta_{ij}, \sigma_{s_i k}^2).$$

17. Update δ_{kl}^2 for $k = 1, \ldots, p^*$ and $l = 1, \ldots, L + Q_\mu + Q$ from

$$\text{IG}\left(\frac{H+1}{2}, \frac{\sum_{h=1}^H U_{hkl}^2 + 1}{2} \right).$$

18. Update ζ_{ql}^2 for $q = 1, \ldots, Q$ and $l = 1, \ldots, L$ from

$$\text{IG}\left(\frac{H+1}{2}, \frac{\sum_{h=1}^{H} V_{hql}^2 + 1}{2}\right).$$

19. Using the proposed adjustment method in the previous chapter with the prior Dirichlet$(a_1, \ldots, a_H)$, generate adjusted weights $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_H)'$ from

$$\tilde{\pi} \sim \text{Dirichlet}\left(a_1 + \frac{1}{c}\sum_{i:s_i=1} w_i, \ldots, a_H + \frac{1}{c}\sum_{i:s_i=H} w_i\right),$$

where $c = \sum_{i=1}^{n} w_i/N$.

5.5 Analysis of adolecent sexual behaviour data

This section applies the proposed model to the adolescent sexual behaviour data from Add Health. Our primary interest is in estimating interactions among the response variables and their trajectories in population and sub-populations based on biological backgrounds. As a scale-free measure of dependence between two variables, we employ Goodman and Kruskal's gamma (Goodman and Kruskal (1954); Goodman and Kruskal (1959); Goodman and Kruskal (1963); Goodman and Kruskal (1972)),

$$\gamma = \frac{N_c - N_d}{N_c + N_d}, \tag{5.10}$$

where N_c is the number of concordant pairs and N_d is the number of discordant pairs. A concordant pair can be defined as a pair of variables (X_1,Y_1) and (X_2,Y_2) such that $\text{sgn}(X_2 - X_1) = \text{sgn}(Y_2 - Y_1)$. On the other hand, a discordant pair means $\text{sgn}(X_2 - X_1) = -\text{sgn}(Y_2 - Y_1)$. γ takes values ranging from -1 (100% negative association) to 1 (100% positive association) and zero indicates absence of association.

As a function $f_l(x_{il} \,|\, \phi)$ for the covariates in (5.8), we assume multinomial distributions using Dirichlet priors with all concentration parameters 1. For other priors, we use $\alpha \sim \text{Gamma}(0.25, 0.25)$, $\sigma_{hk}^2 \sim \text{IG}(2, \tilde{b}_k)$ where $\tilde{b}_k$ is the sample variance of

$\log(y_{ik} + 0.5)$ for the count variables and 1 for other variables. We set $H = 50$, $Q = Q_\mu = Q_\eta = 4$ and put 30 grids on the $(0, 1]$ interval for κ_μ and κ_ξ. Also, we set the hyperparameter $a = a_h$ in the Dirichlet prior for adjusting the survey bias such that the prior sample size Ha is equal to 1% of the population size. We generate 20,000 samples after the initial 10,000 samples are discarded as a burn-in period and every 20th sample is saved. At each MCMC iteration, we generated $n = 4,208$ sample from the posterior predictive distribution to compute γ for each age. We observe γ can unstably return values -1 and 1 if the denominator in (5.10) is small. Therefore, we excluded such outliers from the analysis.

Figures 5.4 and 5.5 show the boxplots of γ in the population. Let Ao, As, #, He, MHe, Bi, MHo, Ho and No represent attraction to opposite sex, attraction to same sex, cumulative number of sex partners, heterosexual, mostly heterosexual, bisexual, mostly homosexual, homosexual and no sexual attraction. We observe various patterns of trajectories of the associations. For example, the correlation between Ao and As is around zero at the beginning but drops steeply to strongly negative values after 15 years old. Ao and # roughly show positive correlations over time but decrease to around zero after 25 years old. As and # are positively correlated over the period of time but the interaction shows a dip from 19 to 24 years old. With respect to the sexual self definition, we focus on the associations after 18 years old because it is observed only in wave 3 and 4. The correlation of {Ao and Bi} and {# and No} look stable taking negative values age after 18. {Ao and MHo}, {Ao and Ho}, {Ao and No} are {As and He} strongly negatively correlated. On the other hand, {Ao and He}, {As and MHe}, {As and Bi}, {As and MHo} and {As and Ho} show highly positive interactions after 18. The associations of {# and MHe}, {# and Bi}, {# and MHo} and {# and Ho} slightly increase after 25 years old. On the other hand, {# and He} drop after 25.

Figures 5.6 and 5.7 report the comparison of the posterior means in the population

with those in the sub-populations defined by gender or race. We observe similar trends among these groups for various pairs such as {Ao, As}. On the other hand, we can see differences, especially between male and female, in other associations. In {Ao and #}, male shows relatively high positive correlations for teenagers while the interaction for female is small. In {Ao, MHe}, {As, #}, {#, MHe}, {#, Bi}, {#, MHo} and {#, Ho}, female indicates higher associations but the interactions for male are low with relatively large difference. On the flip side, the associations for male are larger than female in {#, He}. For the sub-populations based on the interaction of gender and race, the comparison of the posterior means of γ is given in Figures 5.8 and 5.9. As in Figures 5.6 and 5.7, trends for many pairs look similar such as for {Ao, As}. On the other hand, we can see different patterns of trends between non-white male and white female for many pairs such as in {Ao, MHe}, {As, #}, {#, MHe}, {#, Bi}, {#, MHo}, {#, Ho} and {# and He}.

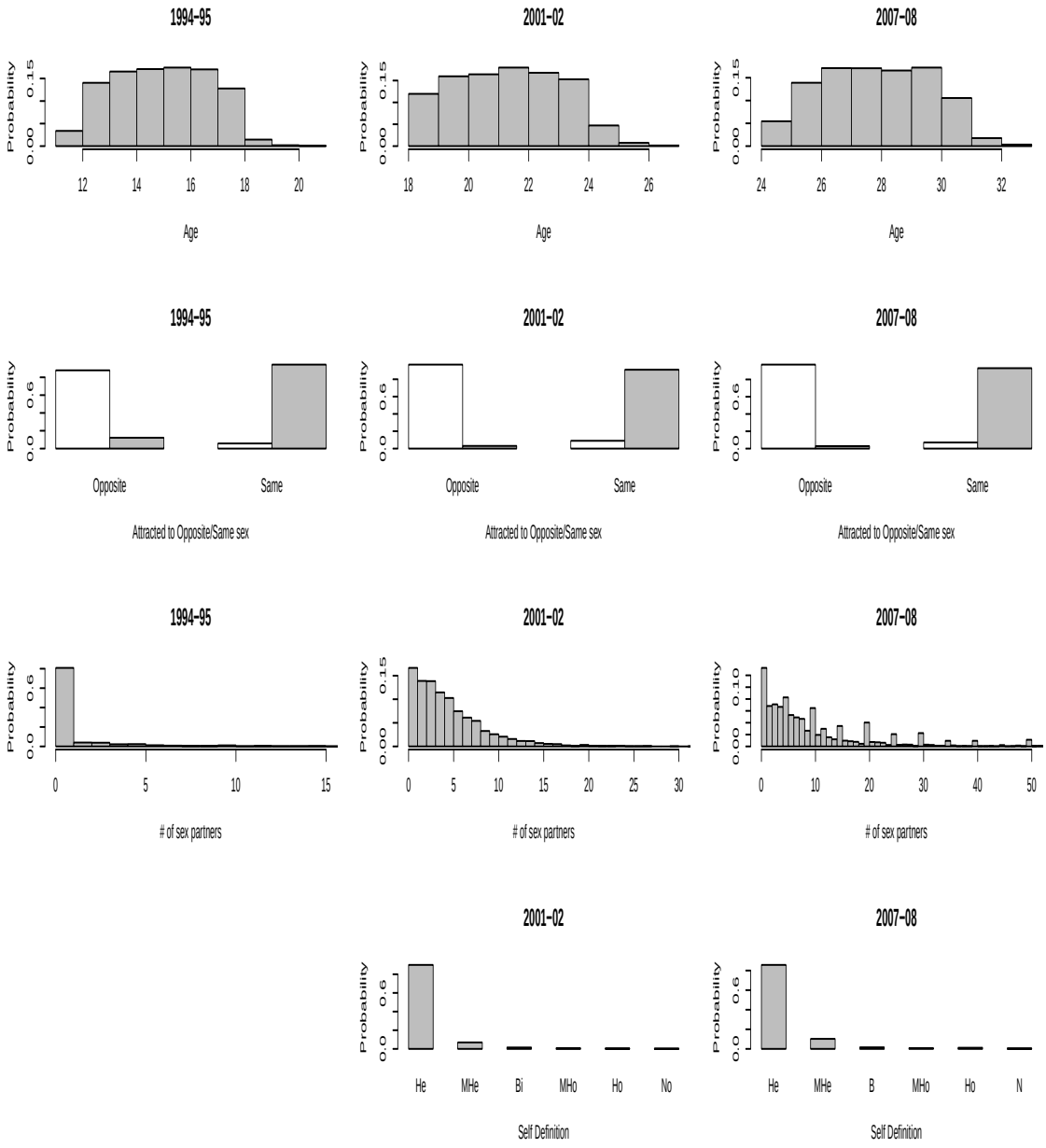


FIGURE 5.1: Histograms of sexual behaviour data. The first, second, third and forth columns show age, attraction to opposite/same sex (white: yes, gray: no), cumulative number of sex partners and sexual self definition. For the self definition, He, MHe, Bi, MHo, Ho and No mean 100% heterosexual, mostly heterosexual, bisexual, mostly homosexual, 100% homosexual and no sexual attraction respectively. The first, second and third rows correspond to wave 1, 3 and 4 respectively.

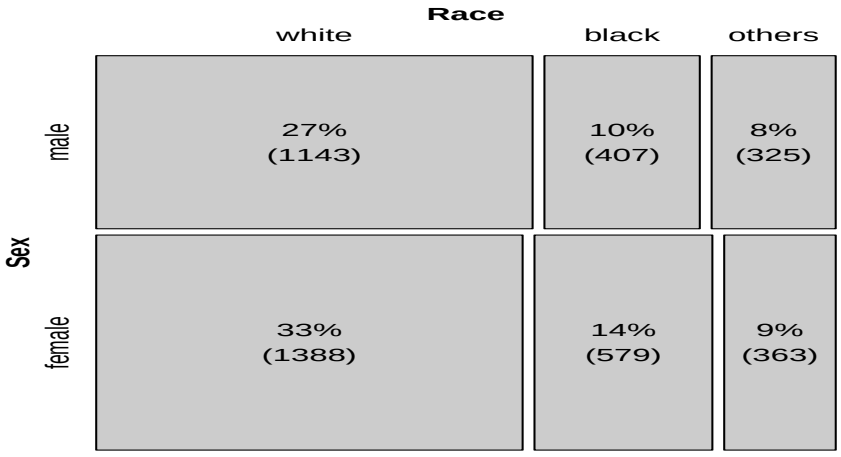


FIGURE 5.2: Mosaic plot of covariates. y-axis and x-axis correspond to gender and race respectively.

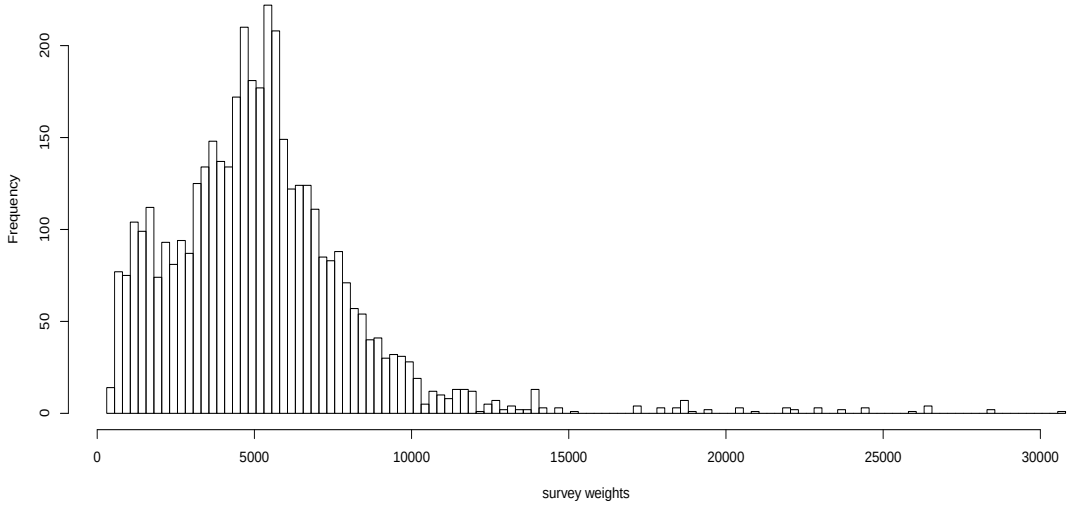


FIGURE 5.3: Histogram of survey weights for longitudinal study with wave 1, 3 and 4.

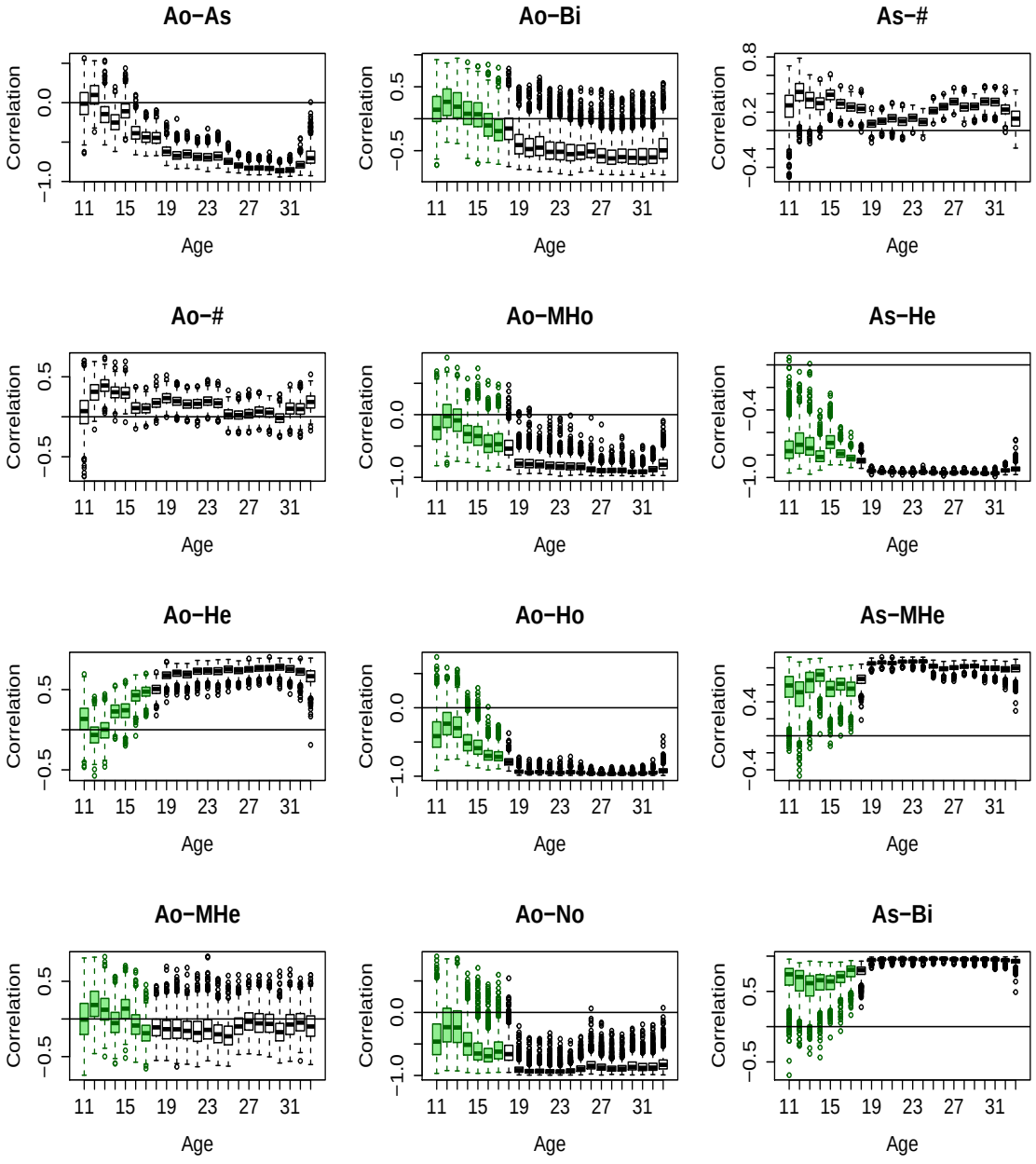


FIGURE 5.4: Boxplots of γ for the population [1]. Ao, As, #, He, MHe, Bi, MHo, Ho and No represent attraction to opposite sex, attraction to same sex, cumulative number of sex partners, hetero, mostly hetero, bisexual, mostly homo, homo and no sexual attraction. Green color means the sexual definition is design-based missing for the corresponding age.

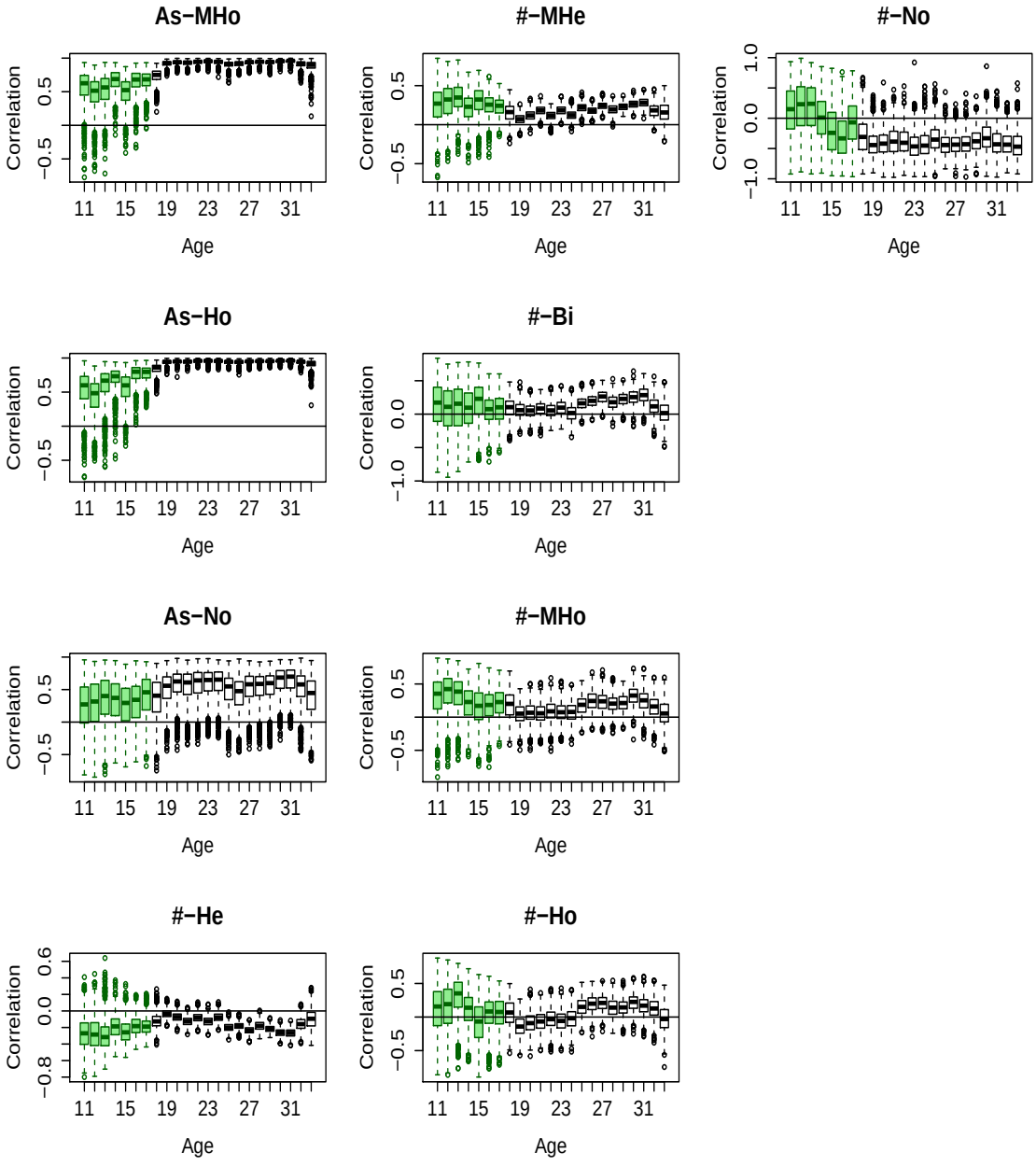


FIGURE 5.5: Boxplots of γ for the population [2]. Ao, As, #, He, MHe, Bi, MHo, Ho and No represent attraction to opposite sex, attraction to same sex, cumulative number of sex partners, hetero, mostly hetero, bisexual, mostly homo, homo and no sexual attraction. Green color means the sexual definition is design-based missing for the corresponding age.

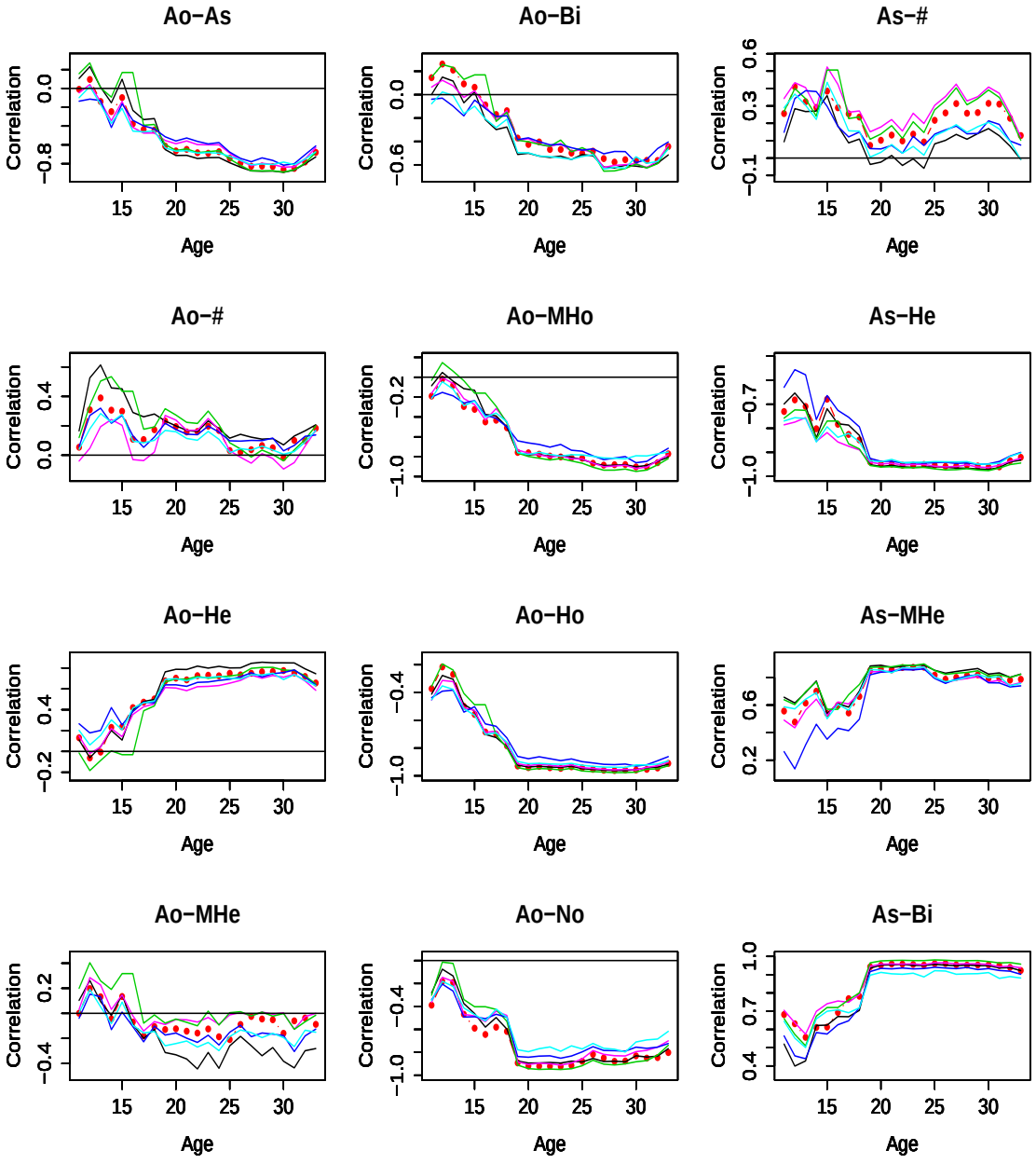


FIGURE 5.6: Comparison of the posterior means of γ for the population (red), male (black), female (magenta), white (green), black (blue) and other races (aqua) [1].

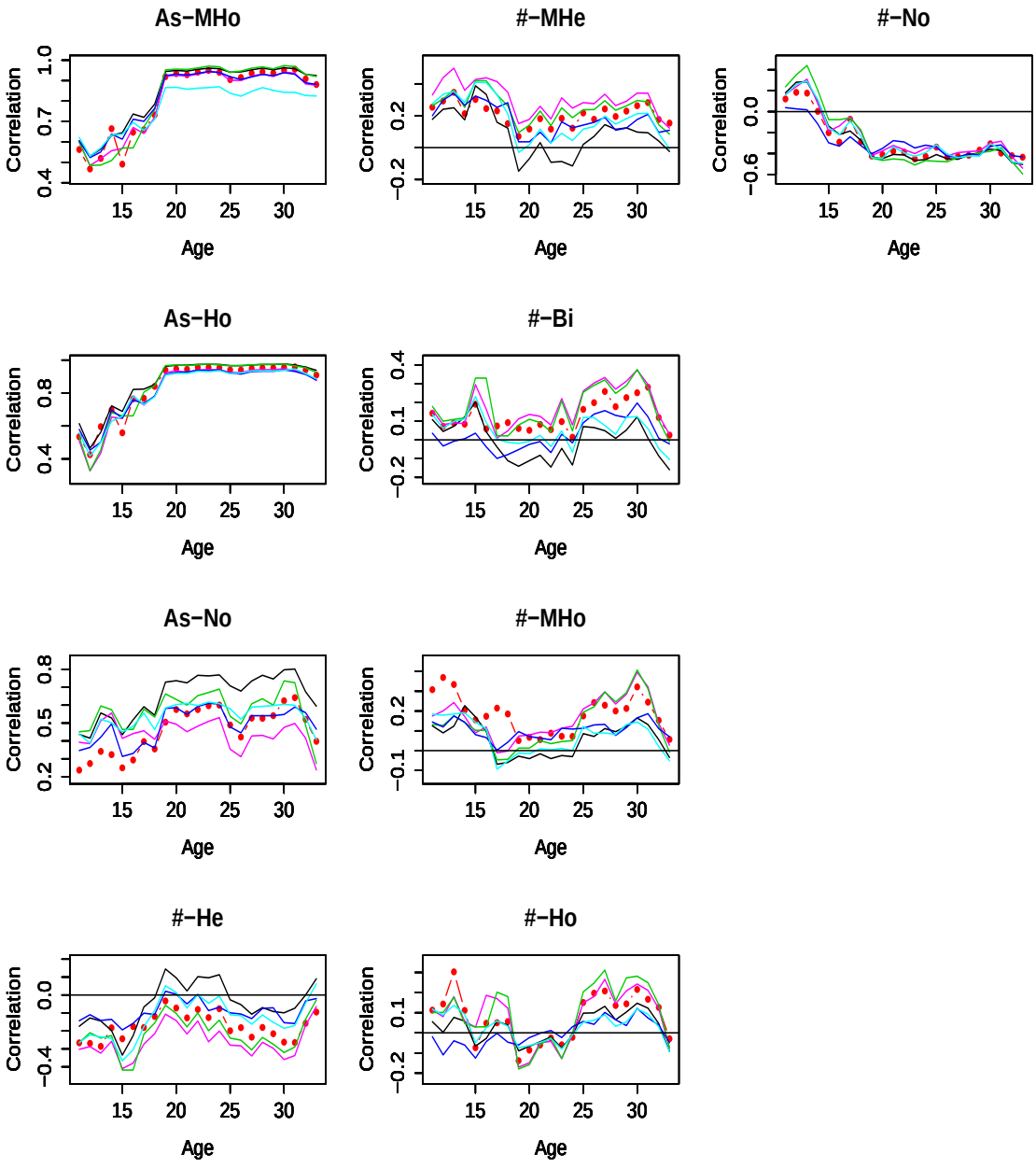


FIGURE 5.7: Comparison of the posterior means of γ for the population (red), male (black), female (magenta), white (green), black (blue) and other races (aqua) [2].

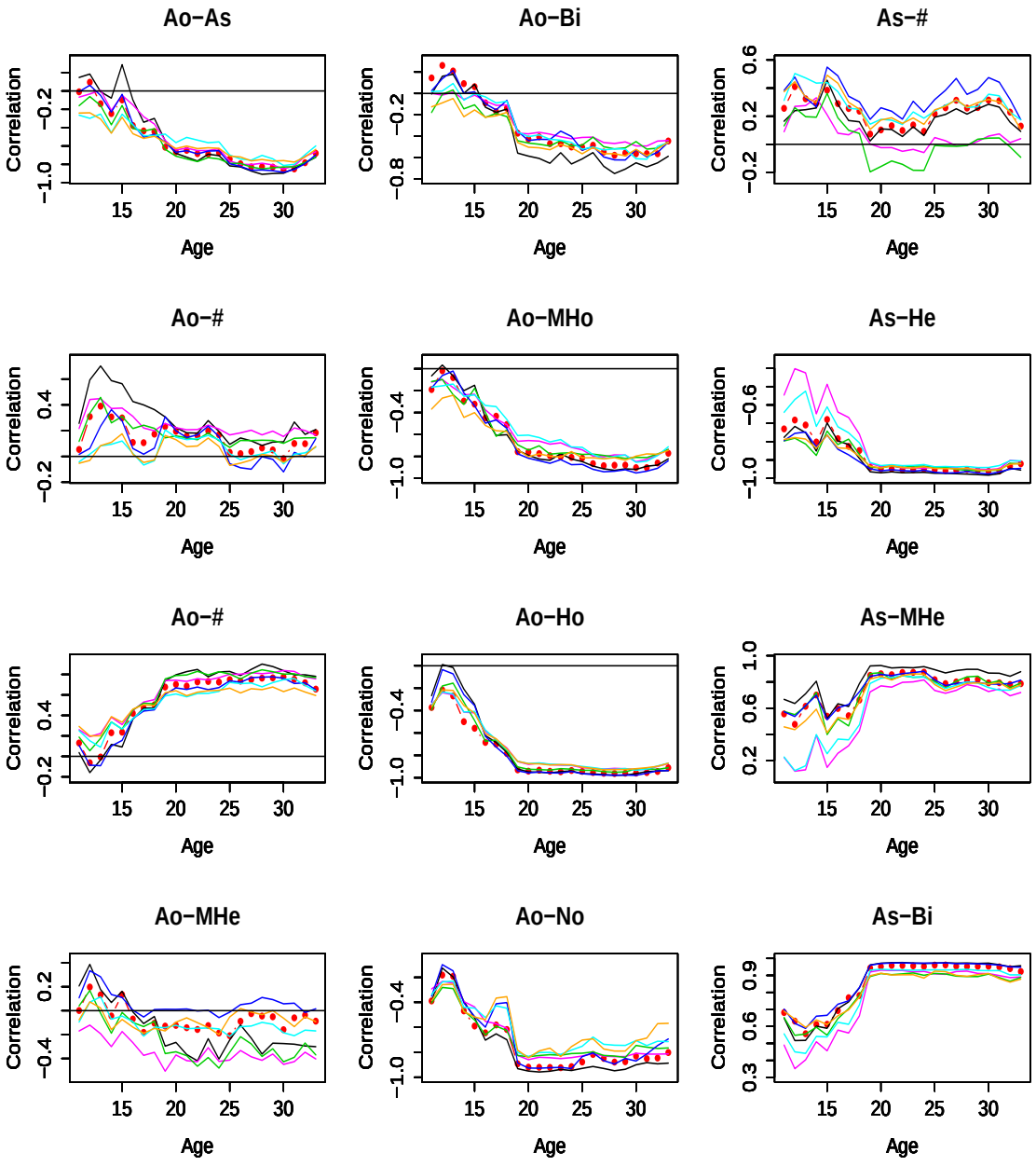


FIGURE 5.8: Comparison of the posterior means of γ for the population (red), white male (black), black male (magenta), other male (green), white female (blue), black female (aqua) and other female (orange) [1].

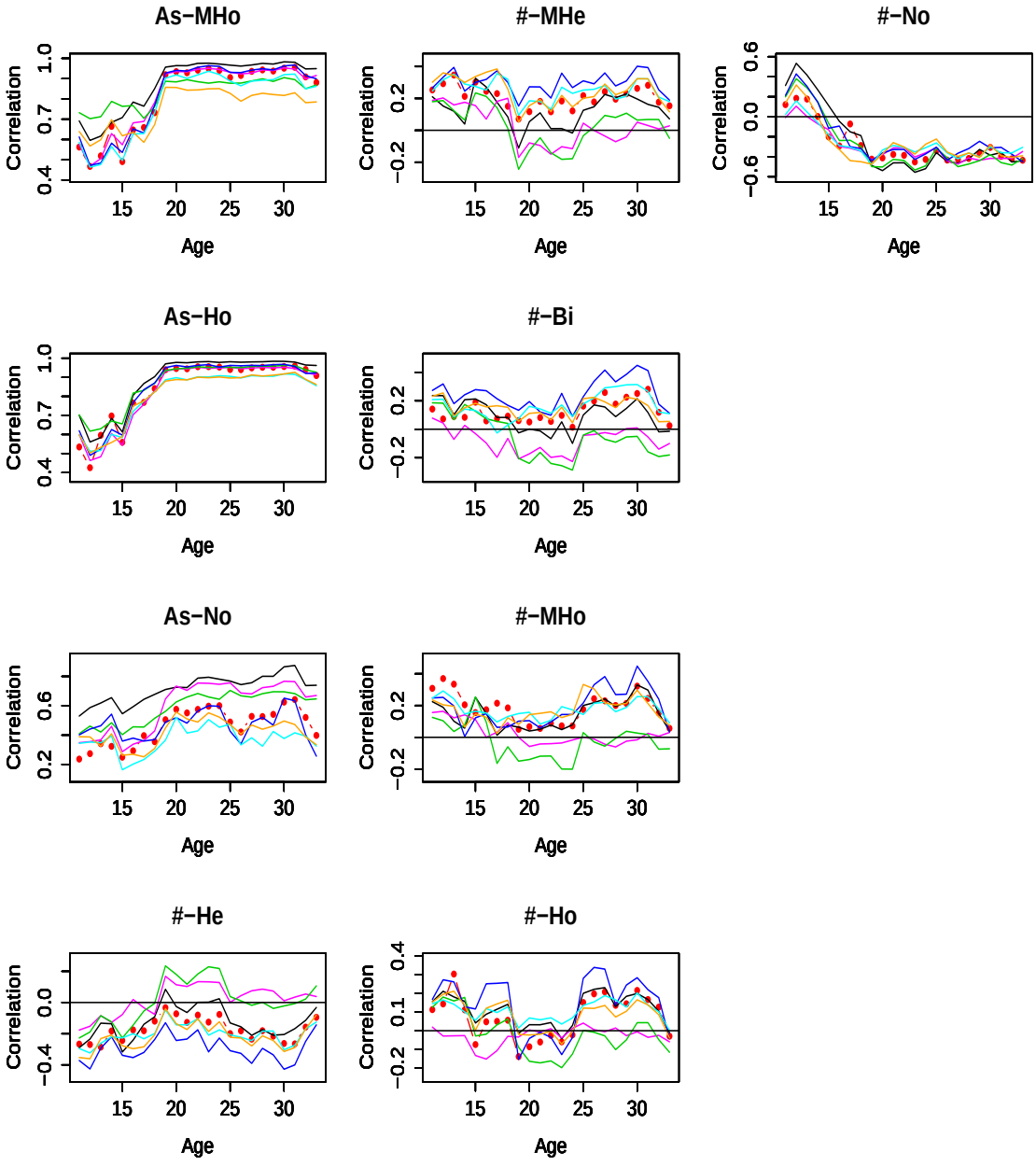


FIGURE 5.9: Comparison of the posterior means of γ for the population (red), white male (black), black male (magenta), other male (green), white female (blue), black female (aqua) and other female (orange) [2].

Appendix A

Supplementary materials for Chapter 2

A.1 Proof of Lemma 1

The expectation of cell probability is

$$E\{\pi_{tc_1\cdots c_p}\} = E\left\{\sum_{h=1}^{\infty}\nu_{th}\prod_{j=1}^{p}\psi_{hc_j}^{(j)}\right\} = \sum_{h=1}^{\infty}\left[E\{\nu_{th}\}\prod_{j=1}^{p}E\left\{\psi_{hc_j}^{(j)}\right\}\right],$$

$$= \prod_{j=1}^{p}E\left\{\psi_{hc_j}^{(j)}\right\}\sum_{h=1}^{\infty}E\{\nu_{th}\} = \prod_{j=1}^{p}E\left\{\psi_{hc_j}^{(j)}\right\} = \prod_{j=1}^{p}\frac{a_{jc_j}}{\hat{a}_j}.$$

The marginal distribution of W_{th} can be expressed as $N(\mu/(1-\phi), \sigma_\eta^2/(1-\phi^2)+\sigma_\varepsilon^2)$, independent of t and h. Hence, we set $\beta_1 = E\{g(W_{th})\}$ and $\beta_2 = E\{g^2(W_{th})\}$. The

second moment of cell probability is

$$
\begin{aligned}
E\{\pi^2_{tc_1\cdots c_p}\} &= E\left[\left\{\sum_{h=1}^{\infty}\nu_{th}\prod_{j=1}^{p}\psi^{(j)}_{hc_j}\right\}\left\{\sum_{l=1}^{\infty}\nu_{tl}\prod_{j=1}^{p}\psi^{(j)}_{lc_j}\right\}\right], \\
&= \sum_{h=1}^{\infty}\sum_{l=1}^{\infty}E\{\nu_{th}\nu_{tl}\}E\left\{\prod_{j=1}^{p}\psi^{(j)}_{hc_j}\psi^{(j)}_{lc_j}\right\}, \\
&= \left[\prod_{j=1}^{p}E\left\{\left(\psi^{(j)}_{hc_j}\right)^2\right\}-\prod_{j=1}^{p}E^2\left\{\psi^{(j)}_{hc_j}\right\}\right]\sum_{h=1}^{\infty}E\{\nu^2_{th}\} \\
&+\prod_{j=1}^{p}E^2\left\{\psi^{(j)}_{hc_j}\right\}\sum_{h=1}^{\infty}\sum_{l=1}^{\infty}E\{\nu_{th}\nu_{tl}\}, \\
&= \left(\prod_{j=1}^{p}\frac{a_{jc_j}(a_{jc_j}+1)}{\hat{a}_j(\hat{a}_j+1)}-\prod_{j=1}^{p}\frac{a^2_{jc_j}}{\hat{a}^2_j}\right)\sum_{h=1}^{\infty}E\{\nu^2_{th}\}+\prod_{j=1}^{p}\frac{a^2_{jc_j}}{\hat{a}^2_j},
\end{aligned}
$$

where

$$
\begin{aligned}
\sum_{h=1}^{\infty}E\{\nu^2_{th}\} &= \sum_{h=1}^{\infty}E\left[g^2(W_{th})\prod_{l<h}\{1-g(W_{tl})\}^2\right], \\
&= \sum_{h=1}^{\infty}\beta_2\{1-2\beta_1+\beta_2\}^{h-1}, \\
&= \frac{\beta_2}{2\beta_1-\beta_2}.
\end{aligned}
$$

Hence,

$$
V\{\pi_{tc_1\cdots c_p}\} = \left(\prod_{j=1}^{p}\frac{a_{jc_j}(a_{jc_j}+1)}{\hat{a}_j(\hat{a}_j+1)}-\prod_{j=1}^{p}\frac{a^2_{jc_j}}{\hat{a}^2_j}\right)\left(\frac{\beta_2}{2\beta_1-\beta_2}\right). \tag{A.1}
$$

Similarly,

$$E\{\pi_{tc_1\cdots c_p}\pi_{t+kc'_1\cdots c'_p}\} = E\left[\left\{\sum_{h=1}^{\infty}\nu_{th}\prod_{j=1}^{p}\psi_{hc_j}^{(j)}\right\}\left\{\sum_{l=1}^{\infty}\nu_{t+kl}\prod_{i=1}^{p}\psi_{lc'_i}^{(i)}\right\}\right],$$

$$= \left[\prod_{j=1}^{p}E\left\{\psi_{hc_j}^{(j)}\psi_{hc'_j}^{(j)}\right\} - \prod_{j=1}^{p}E\left\{\psi_{hc_j}^{(j)}\right\}E\left\{\psi_{lc'_j}^{(j)}\right\}\right]\sum_{h=1}^{\infty}E\{\nu_{th}\nu_{t+kh}\} + \prod_{j=1}^{p}E\left\{\psi_{hc_j}^{(j)}\right\}E\left\{\psi_{lc'_j}^{(j)}\right\},$$

$$= \left(\prod_{j=1}^{p}\frac{a_{jc_j}\{a_{jc'_j}+1(c_j=c'_j)\}}{\hat{a}_j(\hat{a}_j+1)} - \prod_{j=1}^{p}\frac{a_{jc_j}a_{jc'_j}}{\hat{a}_j^2}\right)\sum_{h=1}^{\infty}E\{\nu_{th}\nu_{t+kh}\} + \prod_{j=1}^{p}\frac{a_{jc_j}a_{jc'_j}}{\hat{a}_j^2},$$

where

$$E\{\nu_{th}\nu_{t+kh}\} = E\left\{\left[g(W_{th})\prod_{l<h}\{1-g(W_{tl})\}\right]\left[g(W_{t+kh})\prod_{l<h}\{1-g(W_{t+kl})\}\right]\right\},$$

$$= E\left\{g(W_{th})g(W_{t+kh})\right\}\prod_{l<h}E\left[\{1-g(W_{tl})\}\{1-g(W_{t+kl})\}\right],$$

$$= E\left\{g(W_{th})g(W_{t+kh})\right\}\prod_{l<h}\left[1-2\beta_1+E\{g(W_{tl})g(W_{t+kl})\}\right].$$

From (2.7) and (2.8), $E\left\{g(W_{th})g(W_{t+kh})\right\}$ can be expressed as

$$E\left\{g(W_{th})g(W_{t+kh})\right\} = E\left\{g(\alpha_{th}+\varepsilon_{th})g(\alpha_{t+kh}+\varepsilon_{t+kh})\right\},$$

$$= E\left\{g\left(\alpha_{th}+\varepsilon_{th}\right)g\left(\frac{1-\phi^k}{1-\phi}\mu+\phi^k\alpha_{th}+\sum_{i=0}^{k-1}\phi^i w_{t+k-ih}+\varepsilon_{t+kh}\right)\right\}.$$

Since α_{th}, ε_{th}, w_{t+k-ih} $(i=0,\ldots,k-1)$ and ε_{t+kh} are independent of one another and their distributions do not depend on t or h, hence $\gamma_k \equiv E\left\{g(W_{th})g(W_{t+kh})\right\}$ is dependent on time difference k but independent of time t.

In addition,

$$\sum_{h=1}^{\infty}E\{\nu_{th}\nu_{t+kh}\} = \sum_{h=1}^{\infty}\gamma_k\prod_{l<h}\left\{1-2\beta_1+\gamma_k\right\},$$

$$= \frac{\gamma_k}{2\beta_1-\gamma_k}.$$

Hence,

$$Cov\{\pi_{tc_1\cdots c_p}, \pi_{t+kc'_1\cdots c'_p}\} = \left(\prod_{j=1}^{p} \frac{a_{jc_j}\{a_{jc'_j} + 1(c_j = c'_j)\}}{\hat{a}_j(\hat{a}_j + 1)} - \prod_{j=1}^{p} \frac{a_{jc_j}a_{jc'_j}}{\hat{a}_j^2}\right)\left(\frac{\gamma_k}{2\beta_1 - \gamma_k}\right).$$

Since $\beta_2/(2\beta_1 - \beta_2) > 0$, $\gamma_k/(2\beta_1 - \gamma_k) > 0$ and (A.1), cell probabilities with $c_j = c'_j$ for all j have positive covariance and, on the other hand, those with $c_j \neq c'_j$ for all j have negative covariance.

In a case where $a_{j1} = \cdots = a_{jc_j} = a$, the variance and covariance are expressed as

$$V\{\pi_{tc_1\cdots c_p}\} = \left(\prod_{j=1}^{p} \frac{1 + 1/a}{d_j^2 + d_j/a} - \prod_{j=1}^{p} \frac{1}{d_j^2}\right)\left(\frac{\beta_2}{2\beta_1 - \beta_2}\right),$$

$$Cov\{\pi_{tc_1\cdots c_p}, \pi_{t+kc'_1\cdots c'_p}\} = \left(\prod_{j=1}^{p} \frac{1 + 1(c_j = c'_j)/a}{d_j^2 + d_j/a} - \prod_{j=1}^{p} \frac{1}{d_j^2}\right)\left(\frac{\gamma_k}{2\beta_1 - \gamma_k}\right).$$

Hence, $V\{\pi_{tc_1\cdots c_p}\} \to 0$ and $Cov\{\pi_{tc_1\cdots c_p}, \pi_{t+kc'_1\cdots c'_p}\} \to 0$ as $a \to \infty$.

A.2 Proof of Lemma 2

To prove $\sum_{h=1}^{\infty} \nu_{th} = 1$ a.s., it is enough to show $\sum_{h=1}^{\infty} E\{\log(1 - g(W_{th})\} = -\infty$ (Ishwaran and James (2001)). g is a non-negative monotone increasing link function: $\Re \to (0, 1)$, therefore $0 < \beta_1 = E\{g(W_{th})\} < 1$. Then, using Jensen's inequality,

$$E[\log\{1 - g(W_{th})\}] \leq \log[1 - E\{g(W_{th})\}] = \log(1 - \beta_1) < 0.$$

Therefore, $\sum_{h=1}^{\infty} E\{\log(1 - g(W_{th})\} = -\infty$ at each time point.

A.3 Proof of Theorem 3

The proposed prior probability assigned to $\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$ can be expressed as

$$\mathcal{Q}\left\{\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)\right\} = \int 1(\|\boldsymbol{\pi} - \boldsymbol{\pi}^0\| < \epsilon) d\mathcal{Q}(\boldsymbol{\nu}_t, \boldsymbol{\psi}_h^{(j)}, t \in \{1, \ldots, T\}, h = 1, \ldots, \infty, j = 1, \ldots, p).$$

where $\boldsymbol{\nu}_t$ is a probability vector induced by the proposed stick breaking process and we use the L_1 distance

$$\|\boldsymbol{\pi} - \boldsymbol{\pi}^0\| = \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} |\pi_{tc_1\cdots c_p} - \pi^0_{tc_1\cdots c_p}|,$$

where p_t is a probability mass function for time $t \in \{1, \ldots, T\}$.

For any $\boldsymbol{\pi}^0 \in \Pi$, each component in $\boldsymbol{\pi}^0$ can be expressed as

$$\boldsymbol{\pi}^0_t = \sum_{h=1}^{k_t} \nu^0_{th} \Psi_{th}, \qquad \Psi_{th} = \boldsymbol{\psi}^{(1)}_{th} \otimes \cdots \otimes \boldsymbol{\psi}^{(p)}_{th},$$

where $k_t \in \mathbb{N}$, $\boldsymbol{\nu}^0_t = (\nu^0_{t1}, \ldots, \nu^0_{tk_t})'$ is a probability vector, $\Psi_{th} \in \Pi_{d_1\cdots d_p}$ and $\boldsymbol{\psi}^{(j)}_{th} = (\psi^{(j)}_{th1}, \ldots, \psi^{(j)}_{thd_j})'$ is a $d_j \times 1$ probability vector. We define $k^+_0 = 0$ and $k^+_t = \sum_{i=1}^{t} k_i$ for $t = 1, \ldots, T$. Then, we construct $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t, t \in \{1, \ldots, T\}\} \in \Pi$ induced by the proposed prior such that the component with the index h in $\boldsymbol{\pi}^0_t$ is approximated by the component with the index $k^+_{t-1} + h$ in $\boldsymbol{\pi}_t$. For any ϵ, we define a set $D(\boldsymbol{\pi}^0, \epsilon) \subset \Pi$ such that for any $\boldsymbol{\pi} \in D(\boldsymbol{\pi}^0, \epsilon)$, each $\boldsymbol{\pi}_t$ can be expressed as (4) satisfying $\boldsymbol{\nu} \in \mathcal{N}_{\epsilon'}(\tilde{\boldsymbol{\nu}})$ where $\boldsymbol{\nu} = \{\boldsymbol{\nu}_t, t \in \{1, \ldots, T\}\}$, $\tilde{\boldsymbol{\nu}} = \{\tilde{\boldsymbol{\nu}}_t, t \in \{1, \ldots, T\}\}$ and $\tilde{\boldsymbol{\nu}}_t = (\tilde{\nu}_{t1}, \tilde{\nu}_{t2}, \ldots)'$ is a probability vector where

$$\tilde{\nu}_{tm} = \begin{cases} \nu^0_{tm-k^+_{t-1}}, & (k^+_{t-1} < m \le k^+_t), \\ 0, & (\text{otherwise}), \end{cases}$$

therefore $\tilde{\nu}_{tf(t,h)} = \nu^0_{th}$ where $f(t,h) = k^+_{t-1} + h$ for $1 \le h \le k_t$. Also, $\epsilon' = \epsilon/2 \prod_{j=1}^{p} d_j$, and $\boldsymbol{\psi}^{(j)}_{k^+_{t-1}+h} \in \mathcal{N}_{\epsilon''}\left(\boldsymbol{\psi}^{(j)}_{th}\right)$ for $h = 1, \ldots, k_t$ and $t = 1, \ldots, T$ where $\epsilon'' = \epsilon/2 \sum_t p_t k_t p \prod_j d_j$.

We consider the intervals (a_{th}, b_{th}) in the real line for W_{th} in the proposed prior for $h = 1, \ldots, k^+_t$ and $t = 1, \ldots, T$ where

$$a_{th} = \begin{cases} g^{-1}\{\max(\tilde{\nu}_{th} - \tilde{\epsilon}, 0)\}, & (h = 1), \\ g^{-1}\left\{\frac{\max(\tilde{\nu}_{th} - \tilde{\epsilon}, 0)}{\prod_{l<h}\{1 - g(W_{tl})\}}\right\}, & (h = 2, \ldots, k^+_t), \end{cases}$$

$$b_{th} = \begin{cases} g^{-1}\{\tilde{\nu}_{th} + \tilde{\epsilon}\}, & (h = 1), \\ g^{-1}\left\{\frac{\tilde{\nu}_{th}+\tilde{\epsilon}}{\prod_{l<h}\{1-g(W_{tl})\}}\right\}, & (h = 2, \ldots, k_t^+), \end{cases}$$

where $\tilde{\epsilon} = \epsilon'/2\sum_t p_t k_t^+$. In this case, it is straightforward to check $|\nu_{th} - \tilde{\nu}_{th}| < \tilde{\epsilon}$ for $h = 1, \ldots, k_t^+$ and the proposed prior assigns positive probability to these intervals. Then, the distance between $\boldsymbol{\nu}$ and $\tilde{\boldsymbol{\nu}}$ is

$$\begin{aligned} \|\boldsymbol{\nu} - \tilde{\boldsymbol{\nu}}\| &= \sum_{t=1}^{T} p_t \sum_{h=1}^{\infty} |\nu_{th} - \tilde{\nu}_{th}|, \\ &= \sum_{t=1}^{T} p_t \sum_{h=1}^{k_t^+} |\nu_{th} - \tilde{\nu}_{th}| + \sum_{t=1}^{T} p_t \sum_{h>k_t^+} \nu_{th}, \qquad \text{(A.2)} \\ &< 2\tilde{\epsilon} \sum_{t=1}^{T} p_t k_t^+ = \epsilon'. \end{aligned}$$

For the second component in (A.2), $\sum_{h>k_t^+} \nu_{th} < k_t^+ \tilde{\epsilon}$ because $\nu_{th} > \tilde{\nu}_{th} - \tilde{\epsilon}$ for $h = 1, \ldots, k_t^+$ and $\sum_{h=1}^{k_t^+} \nu_{th} > 1 - k_t^+ \tilde{\epsilon}$. In addition, it is straightforward to show that the proposed prior assigns positive probability to $\mathcal{N}_{\epsilon''}\left(\boldsymbol{\psi}_{th}^{(j)}\right)$. Therefore, since $D(\boldsymbol{\pi}^0, \epsilon)$ contains such case, $\mathcal{Q}\{D(\boldsymbol{\pi}^0, \epsilon)\} > 0$.

For any $\boldsymbol{\pi} \in D(\boldsymbol{\pi}^0, \epsilon)$, $\|\boldsymbol{\pi} - \boldsymbol{\pi}^0\|$ is equal to

$$\sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} |\pi_{tc_1\cdots c_p} - \pi^0_{tc_1\cdots c_p}|,$$

$$= \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left| \sum_{h=1}^{\infty} \nu_{th} \prod_{j=1}^{p} \psi^{(j)}_{hc_j} - \sum_{l=1}^{k_t} \nu^0_{tl} \prod_{j=1}^{p} \psi^{(j)}_{tlc_j} \right|,$$

$$= \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left| \sum_{h=1}^{k_t} \left(\nu_{tk^+_{t-1}+h} \prod_{j=1}^{p} \psi^{(j)}_{k^+_{t-1}+hc_j} - \nu^0_{th} \prod_{j=1}^{p} \psi^{(j)}_{thc_j} \right) + \sum_{l \le k^+_{t-1}, k^+_t < l} \nu_{tl} \prod_{j=1}^{p} \psi^{(j)}_{lc_j} \right|,$$

$$\le \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left(\sum_{h=1}^{k_t} \left| \nu_{tk^+_{t-1}+h} \prod_{j=1}^{p} \psi^{(j)}_{k^+_{t-1}+hc_j} - \nu^0_{th} \prod_{j=1}^{p} \psi^{(j)}_{thc_j} \right| + \sum_{l \le k^+_{t-1}, k^+_t < l} \nu_{tl} \right),$$

$$\le \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left(\sum_{h=1}^{k_t} \left| \nu_{tk^+_{t-1}+h} - \nu^0_{th} \right| + \sum_{l=1}^{k_t} \sum_{j=1}^{p} \left| \psi^{(j)}_{k^+_{t-1}+lc_j} - \psi^{(j)}_{tlc_j} \right| + \sum_{l \le k^+_{t-1}, k^+_t < l} \nu_{tl} \right),$$

$$= \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \sum_{t=1}^{T} p_t \sum_{h=1}^{\infty} |\nu_{th} - \tilde{\nu}_{th}| + \sum_{t=1}^{T} p_t \sum_{l=1}^{k_t} \sum_{j=1}^{p} \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left| \psi^{(j)}_{k^+_{t-1}+lc_j} - \psi^{(j)}_{tlc_j} \right|,$$

$$< \prod_{j=1}^{p} d_j \epsilon' + \sum_{t=1}^{T} p_t k_t p \prod_{j=1}^{p} d_j \epsilon'',$$

$$= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Therefore $\boldsymbol{\pi} \in \mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$ and $D(\boldsymbol{\pi}^0, \epsilon) \subset \mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$. Hence, $\mathcal{Q}\{\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)\} > 0$.

A.4 Table of categorical variables

Table A.1: List of categorical variables.

No.	Categorical variable (Name in GSS)
1	Age group* (AGE)
2	Sex (SEX)
3	Race (RACE)
4	Religious preference** (RELIG)
5	Region (REGION)
6	Attitude toward abortion (ABANY)
7	Should Govetnment help pay for medical care? (HELPSICK)
8	Highest degree (DEGREE)
9	Political party affiliation (PARTYID)
10	Current marital status (MARITAL)
11	Astrological sign (ZODIAC)
12	Confidence in banks and financial institutions (CONFINAN)
13	Confidence in U.S. Supreme Court (CONJUDGE)
14	Think of self as liberal or conservative (POLVIEWS)
15	Belief in life after death (POSTLIFE)
16	Attitude toward homosexual sex relations (HOMOSEX)
17	Have gun in home (OWNGUN)
18	Subjective class identification (CLASS)
19	Should Marijuana be made legal (GRASS)
20	Total family income (INCOME)
21	Favor or oppose death penalty for murder (CAPPUN)
22	Attitude toward spending money on space exploration program (NATSPAC)
23	Attitude toward spending money on improving and protecting environment (NATENVIR)
24	Attitude toward spending money on improving and protecting the nations's health (NATHEAL)
25	Attitude toward spending money on halting the rising crime rate (NATCRIME)
26	Attitude toward spending money on dealing with drug addiction (NATDRUG)
27	Attitude toward spending money on improving the nation's education system (NATEDUC)
28	Attitude toward spending money on the military, armaments and defense (NATARMS)
29	Attitude toward spending money on foreigh aid (NATAID)

*The category of Age group is different from the original one: 1. 18 or 19 years old, 2. 20s, 3. 30s, 4. 40s, 5. 50s, 6. 60s, 7. 70s, 8. more than 80 years old.
**The category of Religious preference is different from the original one: 1. Protestant, 2. Catholic, 3. Jewish, 4. None, 5. Others.

Appendix B

Supplementary materials for Chapter 3

B.1 Proof of Theorem 4

For $\epsilon > 0$, we define $E = [f : KL\{f_0(y, x, z), f(y, x, z)\} < \epsilon]$. Then, there exists N such that for $n > N$ and $f \in E$,

$$|\zeta(f, P_n) - \zeta_0| = \left| \int \log \frac{f(y,x,z)f(z)}{f(y,z)f(x,z)} dP_n - \int \log \frac{f_0(y,x,z)f_0(z)}{f_0(y,z)f_0(x,z)} dP_0 \right|,$$

$$\leq \sup_{f \in \mathcal{F}} \left| \int \log \frac{f_0(y,x,z)}{f(y,x,z)} dP_n - \int \log \frac{f_0(y,x,z)}{f(y,x,z)} dP_0 \right| \tag{B.1}$$

$$+ \sup_{f \in \mathcal{F}} \left| \int \log \frac{f_0(y,z)}{f(y,z)} dP_n - \int \log \frac{f_0(y,z)}{f(y,z)} dP_0 \right| \tag{B.2}$$

$$+ \sup_{f \in \mathcal{F}} \left| \int \log \frac{f_0(x,z)}{f(x,z)} dP_n - \int \log \frac{f_0(x,z)}{f(x,z)} dP_0 \right| \tag{B.3}$$

$$+ \sup_{f \in \mathcal{F}} \left| \int \log \frac{f_0(z)}{f(z)} dP_n - \int \log \frac{f_0(z)}{f(z)} dP_0 \right| \tag{B.4}$$

$$+ \left| \int \log \frac{f_0(y,x,z)f_0(z)}{f_0(y,z)f_0(x,z)} dP_n - \int \log \frac{f_0(y,x,z)f_0(z)}{f_0(y,z)f_0(x,z)} dP_0 \right| \tag{B.5}$$

$$+ \int \log \frac{f_0(y,x,z)}{f(y,x,z)} dP_0 + \int \log \frac{f_0(y,z)}{f(y,z)} dP_0 \tag{B.6}$$

$$+ \int \log \frac{f_0(x,z)}{f(x,z)} dP_0 + \int \log \frac{f_0(z)}{f(z)} dP_0 \leq 9\epsilon, \quad \text{almost surely.} \tag{B.7}$$

Each term in (B.1)-(B.4) can be bounded by ϵ almost surely from the definition of P_0-Glivenko-Cantelli classes. (B.11) goes to zero by the strong law of large numbers. The terms in (B.6) and (B.7) are bounded by 2ϵ almost surely respectively. This comes from the non-negativity of the Kullback-Leibler divergence, for example,

$$\begin{aligned} \int \log \frac{f_0(y,z)}{f(y,z)} dP_0 &\leq \int \log \frac{f_0(y,z)}{f(y,z)} dP_0 + \int \log \frac{f_0(x \mid y,z)}{f(x \mid y,z)} dP_0 \\ &= \int \log \frac{f_0(y,x,z)}{f(y,x,z)} dP_0 < \epsilon. \end{aligned}$$

Hence, by setting $\epsilon' = 9\epsilon$, $E \subset \{f : |\zeta(f, P_n) - \zeta_0| < \epsilon'\}$. The argument by a 2012 unpublished technical paper of A. Norets shows if $[\log\{f_0(y,x,z)/f(y,x,z)\}, f \in \mathcal{F}]$ is P_0-Glivenko-Cantelli and the Kullback-Leibler support condition (3.3) is satisfied, then the posterior converges to the true data-generating function in the Kullback-Leibler distance. Therefore, $\Pi\{|\zeta(f, P_n) - \zeta_0| < \epsilon' \mid D_n\} \geq \Pi(E \mid D_n) \to 1$ almost surely P_0^∞.

B.2 Proof of Theorem 5

For $\epsilon > 0$, we define $E = [f : KL\{f_0(y,x), f(y,x)\} < \epsilon]$. Then, there exists N such that for $n > N$ and $f \in E$,

$$\begin{aligned}
\max_{1\le j\le p} |\zeta_j(f, P_n) - \zeta_0| &= \max_{1\le j\le p} \left| \int \log \frac{f(y,x)f(x_{-j})}{f(y,x_{-j})f(x)} dP_n - \int \log \frac{f_0(y,x)f_0(x_{-j})}{f_0(y,x_{-j})f_0(x)} dP_0 \right|, \\
&\le \sup_{f\in\mathcal{F}} \left| \int \log \frac{f_0(y,x)}{f(y,x)} dP_n - \int \log \frac{f_0(y,x)}{f(y,x)} dP_0 \right| && \text{(B.8)} \\
&+ \max_{1\le j\le p} \sup_{f\in\mathcal{F}} \left| \int \log \frac{f_0(y,x_{-j})}{f(y,x_{-j})} dP_n - \int \log \frac{f_0(y,x_{-j})}{f(y,x_{-j})} dP_0 \right| && \text{(B.9)} \\
&+ \sup_{f\in\mathcal{F}} \left| \int \log \frac{f_0(x)}{f(x)} dP_n - \int \log \frac{f_0(x)}{f(x)} dP_0 \right| && \text{(B.10)} \\
&+ \max_{1\le j\le p} \sup_{f\in\mathcal{F}} \left| \int \log \frac{f_0(x_{-j})}{f(x_{-j})} dP_n - \int \log \frac{f_0(x_{-j})}{f(x_{-j})} dP_0 \right| && \text{(B.11)} \\
&+ \max_{1\le j\le p} \left| \int \log \frac{f_0(y,x)f_0(x_{-j})}{f_0(y,x_{-j})f_0(x)} dP_n - \int \log \frac{f_0(y,x)f_0(x_{-j})}{f_0(y,x_{-j})f_0(x)} dP_0 \right| && \text{(B.12)} \\
&+ \int \log \frac{f_0(y,x)}{f(y,x)} dP_0 + \max_{1\le j\le p} \int \log \frac{f_0(y,x_{-j})}{f(y,x_{-j})} dP_0 && \text{(B.13)} \\
&+ \int \log \frac{f_0(x)}{f(x)} dP_0 + \max_{1\le j\le p} \int \log \frac{f_0(x_{-j})}{f(x_{-j})} dP_0, && \text{(B.14)} \\
&\le 9\epsilon, \quad \text{almost surely.}
\end{aligned}$$

(B.8)-(B.11) are less than ϵ almost surely from the definition of P_0-Glivenko-Cantelli classes. (B.12) converges to zero by the strong law of large numbers. Each term in (B.13) and (B.14) are bounded by $KL\{f_0(y,x), f(y,x)\}$, which is less than ϵ almost surely. Therefore, $E \subset \{f : \max_{1\le j\le p} |\zeta_j(f, P_n) - \zeta_0| < \epsilon'\}$ where $\epsilon' = 9\epsilon$ and $\Pi\{\max_{1\le j\le p} |\zeta_j(f, P_n) - \zeta_0| < \epsilon' \mid D_n\} \ge \Pi(E \mid D_n) \to 1$ almost surely P_0^∞ from the posterior consistency of the joint densities in Kullback-Leibler divergence from the

argument by A. Norets.

B.3 Proof of Lemma 6

Without loss of generality, we assume $p = 2$ and $\beta_0 = 0$. We first show that the Kullback-Leibler support condition holds for the encompassing model. Since Q_0 and G have compact support, we suppose $Q_0(A) = 1$ and $Q(B) = 1$ for Q in the support of Π^Q where $A = \{(\beta, \mu) : -k \leq \beta_1, \beta_2, \mu_1, \mu_2 \leq k\}$ and $B = \{(\beta, \mu) : -k' \leq \beta_1, \beta_2, \mu_1, \mu_2 \leq k'\}$. We can check f_0 has moments of all orders. Hence, for any $\eta > 0$, there exists a such that $\int_{|y|>a} g(y,x) f_0(y,x) dydx < \eta$, $\int_{|x_1|>a} g(y,x) f_0(y,x) dydx < \eta$ and $\int_{|x_2|>a} g(y,x) f_0(y,x) dydx < \eta$ where $g(y,x) = 1 + |x_1| + |x_2| + x_1^2 + x_2^2 + |y||x_1| + |y||x_2| + |x_1||x_2|$. The Kullback-Leibler divergence between f_0 and f can be expressed as

$$\int f_0 \log \frac{f_0}{f} = \int f_0(y,x) \log \frac{\int \phi_{\sigma_0}(y - x'\beta)\phi_{\tau_{0,1}}(x_1 - \mu_1)\phi_{\tau_{0,2}}(x_2 - \mu_2) dQ_0(\beta, \mu)}{\int \phi_{\sigma}(y - x'\beta)\phi_{\tau_1}(x_1 - \mu_1)\phi_{\tau_2}(x_2 - \mu_2) dQ_0(\beta, \mu)} dydx \tag{B.15}$$

$$+ \int f_0(y,x) \log \frac{\int \phi_{\sigma}(y - x'\beta)\phi_{\tau_1}(x_1 - \mu_1)\phi_{\tau_2}(x_2 - \mu_2) dQ_0(\beta, \mu)}{\int \phi_{\sigma}(y - x'\beta)\phi_{\tau_1}(x_1 - \mu_1)\phi_{\tau_2}(x_2 - \mu_2) dQ(\beta, \mu)} dydx. \tag{B.16}$$

With respect to the integral (B.16), we divide the support $\mathcal{R}^3$ into $C = \{(y,x) \in \mathcal{R}^3 : -a \leq y, x_1, x_2 \leq a\}$ and its complement C^C. For the complement, we consider

the subspace $\{(y,x) \in \mathcal{R}^3 : y < -a, -a \le x_1, x_2 \le a\}$ for example.

$$\int_{-\infty}^{-a}\int_{-a}^{a}\int_{-a}^{a} f_0(y,x)\log\frac{\int \phi_\sigma(y-x'\beta)\phi_{\tau_1}(x_1-\mu_1)\phi_{\tau_2}(x_2-\mu_2)dQ_0(\beta,\mu)}{\int \phi_\sigma(y-x'\beta)\phi_{\tau_1}(x_1-\mu_1)\phi_{\tau_2}(x_2-\mu_2)dQ(\beta,\mu)}dydx,$$

$$\int_{-\infty}^{-a}\int_{-a}^{a}\int_{-a}^{a} f_0(y,x)\log\frac{\sup_{(\beta,\mu)\in A}\phi_\sigma(y-x'\beta)\phi_{\tau_1}(x_1-\mu_1)\phi_{\tau_2}(x_2-\mu_2)}{\inf_{(\beta,\mu)\in B}\phi_\sigma(y-x'\beta)\phi_{\tau_1}(x_1-\mu_1)\phi_{\tau_2}(x_2-\mu_2)}dydx,$$

$$\le \int_{-\infty}^{-a}\int\int \frac{1}{2\sigma^2}\{(k^2+k'^2)(x_1^2+x_2^2)+2(k+k')(|x_1|+|x_2|)|y|+2(k^2+k'^2)|x_1||x_2|\}$$

$$\times f_0(y,x)dydx$$

$$+\int_{-\infty}^{-a}\int\int\left(\frac{k+k'}{\tau_1^2}|x_1|+\frac{k^2+k'^2}{2\tau_1^2}+\frac{k+k'}{\tau_2^2}|x_2|+\frac{k^2+k'^2}{2\tau_2^2}\right)f_0(y,x)dydx$$

$$<\left(\frac{k+k'}{\sigma^2}+\frac{3(k^2+k'^2)}{2\sigma^2}+\frac{k+k'}{\tau_1^2}+\frac{k^2+k'^2}{2\tau_1^2}+\frac{k+k'}{\tau_2^2}+\frac{k^2+k'^2}{2\tau_2^2}\right)\eta. \qquad \text{(B.17)}$$

For other regions in C^C where one of y, x_1 and x_2 is larger than a or smaller than $-a$, the corresponding integral can be bounded by (B.17). Following the proof of Theorem 3 in Ghosal et al. (1999), there exists a set E with $\Pi^Q(E) > 0$ and for $Q \in E$, the integral over C is less than $3\tilde{\eta}/(1-3\tilde{\eta})$ where $0 < \tilde{\eta} < 1/3$. Therefore, for $Q \in E$, the integral (B.16) is less than

$$6\left(\frac{k+k'}{\sigma^2}+\frac{3(k^2+k'^2)}{2\sigma^2}+\frac{k+k'}{\tau_1^2}+\frac{k^2+k'^2}{2\tau_1^2}+\frac{k+k'}{\tau_2^2}+\frac{k^2+k'^2}{2\tau_2^2}\right)\eta+\frac{3\tilde{\eta}}{1-3\tilde{\eta}}.$$

Also, we can show the right term in (B.15) converges to 0 as $\sigma \to \sigma_0$, $\tau_j \to \tau_{0,j}$ with $j = 1, 2$ by the dominated convergence theorem with the inequality

$$\frac{\int \phi_{\sigma_0}(y-x'\beta)\phi_{\tau_{0,1}}(x_1-\mu_1)\phi_{\tau_{0,2}}(x_2-\mu_2)dQ_0(\beta,\mu)}{\int \phi_\sigma(y-x'\beta)\phi_{\tau_1}(x_1-\mu_1)\phi_{\tau_2}(x_2-\mu_2)dQ_0(\beta,\mu)},$$

$$\le \sup_{(\beta,\mu)\in A}\frac{\phi_{\sigma_0}(y-x'\beta)\phi_{\tau_{0,1}}(x_1-\mu_1)\phi_{\tau_{0,2}}(x_2-\mu_2)}{\phi_\sigma(y-x'\beta)\phi_{\tau_1}(x_1-\mu_1)\phi_{\tau_2}(x_2-\mu_2)}.$$

For any $\epsilon > 0$, we can choose η, $\tilde{\eta}$ and a small neighborhood of σ_0 and τ_0 such that both the integrals in (B.15) and (B.16) are less than $\epsilon/2$ respectively. Then, the Kullback-Leibler support condition is satisfied.

Next, we check the Glivenko-Cantelli conditions. For simplicity, we show only $[\log\{f_0(x_1)/f(x_1)\}, f \in \mathcal{F}]$ is P_0-Glivenko-Cantelli but we can similarly prove that other classes of functions also satisfy the condition. According to Theorem 3 in van der Vaart and Wellner (2000), if two classes of functions $\mathcal{F}_0$ and $\mathcal{F}_1$ are P_0-Glivenko-Cantelli, then $g(\mathcal{F}_0, \mathcal{F}_1)$ is also P_0-Glivenko-Cantelli with g a continuous function provided that it has an integrable envelope function. We set $\mathcal{F}_0 = \{f_0(x_1)\}$, $\mathcal{F}_1 = \{f(x_1), f \in \mathcal{F}\}$ and g is a log ratio function. It is clear $\mathcal{F}_0$ is P_0-Glivenko-Cantelli. Then, we show $\mathcal{F}_1$ is P_0-Glivenko-Cantelli by proving $\mathcal{F}_1$ satisfies the sufficient condition, $N_{[]}\{\epsilon, \mathcal{F}_1, L_1(P_0)\} < \infty$ for any $\epsilon > 0$ where $N_{[]}\{\epsilon, \mathcal{F}_1, L_1(P_0)\}$ is the minimum number of ϵ-brackets with which $\mathcal{F}_1$ can be covered in $L_1(P_0)$ distance.

We first construct bracket functions. Let $[\underline{\tau}, \overline{\tau}]$ be the support of τ_1. Because the support of (μ_1, τ_1) is compact, for any $\epsilon > 0$ we can take $h > 0$ such that $f(x_1) = \int \phi_{\tau_1}(x_1 - \mu_1) dQ(\mu_1) < \epsilon$ for $|x_1| > h$ and any $\tau_1 \in [\underline{\tau}, \overline{\tau}]$. Also, we can show that $|f'(x_1)| < K$ for $x_1 \in [-h, h]$ with some constant K. Then, we take $0 < \epsilon' < \epsilon/(K+1)$ and divide the interval $[-h, h]$ into sub-intervals $\{I_i, i = 1, \ldots, G\}$ of equal length less than ϵ' with $[-h, h] = \cup_i I_i$ and $I_i \cap I_j = \emptyset$ for $i \neq j$. On each interval I_i, we define $u_{ij} = (j\epsilon' + \epsilon)1_{I_i}$ and $l_{ij} = (j\epsilon')1_{I_i}$ for $j = 0, \ldots, J$ such that $J\epsilon' > \max_{x_1 \in [-h,h]} \max_{\tau_1 \in [\underline{\tau}, \overline{\tau}]} f(x_1)$ where 1_I is an indicator function on the interval I. Letting $m_i \in \{1, \ldots, J\}$ and $m = (m_1, \ldots, m_G)$, we define $u_m = \sum_{i=1}^G u_{im_i} + \epsilon 1_{[-h,h]^C}$ and $l_m = \sum_{i=1}^G l_{im_i}$. Then, it is straightforward to check $l_m < u_m$ and $||u_m - l_m||_{L_1(P_0)} \leq ||u_m - l_m||_\infty < \epsilon$. Because $|f'(x_1)| < K$ and $\epsilon/\epsilon' > K + 1$, for any $f \in \mathcal{F}_1$ there exists m_i such that $l_{im_i} \leq f \leq u_{im_i}$ on the interval I_i and further we can find some m such that $l_m \leq f \leq u_m$ on $\mathcal{R}$. Since $m \in \{1, \ldots, J\}^G$, the set $\{(l_m, u_m)\}$ consists of a finite number of functions. Therefore, $N_{[]}\{\epsilon, \mathcal{F}_1, L_1(P_0)\} < \infty$.

With respect to the envelop function,

$$\begin{aligned}\left|\log\frac{f_0(x_1)}{f(x_1)}\right| &\leq \log\max\left(\overline{\tau}\tau_{0,1}^{-1}, \tau_{0,1}\underline{\tau}^{-1}\right) + (\tau_{0,1}^{-2} + \underline{\tau})x_1^2 + 2(\tau_{0,1}^{-2}k + \underline{\tau}^{-2}k')|x_1| \\ &\quad + \tau_{0,1}^{-2}k^2 + \underline{\tau}^{-2}k'^2, \\ &\equiv B(x_1).\end{aligned}$$

It is easy to check $\int B(x_1)dP_0 < \infty$. As a result, $[\log\{f_0(x_1)/f(x_1)\}, f \in \mathcal{F}]$ is P_0-Glivenko-Cantelli.

B.4 Supplemental materials for application to criminology data

B.4.1 Data in the criminology application

The whole data set can be downloaded from the University of California Irvine machine learning repository website. Further information is given in [1] United States Department of Commerce, Bureau of the Census, census of population and housing 1990 United States: summary tape file 1a and 3a, [2] United States Department of Commerce, Bureau of the Census Producer, Washington, DC and Inter-university consortium for political and social research, Ann Arbor, Michigan in 1992, [3] United States Department of Justice, Bureau of Justice Statistics, law enforcement management and administrative statistics, [4] United States Department of Justice, Federal Bureau of Investigation, crime in the United States in 1995.

As for the predictors, Table B.1 and Table B.2 give the whole list.

B.4.2 Markov chain Monte Carlo Algorithm

Relying on the blocked Gibbs sampler by Ishwaran and James (2001), we develop an efficient posterior computation method for the proposed Dirichlet process mixture model. Let $s = (s_1, \ldots, s_n)'$ be the latent cluster index variables. Then, we propose the following Markov chain Monte Carlo algorithm:

1. Update V_h for $h = 1, \ldots, H-1$ from

$$\text{Be}\left(1 + n_h, \alpha_0 + \sum_{l>h} n_l\right),$$

where $n_h = \sum_{i=1}^{n} 1(s_i = h)$.

2. Using the prior Gamma(a_α, b_α), update α_0 from

$$\text{Gamma}\left\{a_\alpha + H - 1, b_\alpha - \sum_{h=1}^{H-1} \log(1 - V_h)\right\}.$$

3. Update s_i for $i = 1, \ldots, n$ from

$$\text{pr}(s_i = h \mid \cdots) = \frac{\pi_h f(y_i \mid x_i, \theta_h) \prod_{j=1}^{p} f(x_{i,j} \mid \theta_h)}{\sum_{l=1}^{H} \pi_l f(y_i \mid x_i, \theta_l) \prod_{j=1}^{p} f(x_{i,j} \mid \theta_l)}.$$

4. Update $\mu_{j,h}$ for $j = 1, \ldots, p$ and $h = 1, \ldots, H$ from $N(\tilde{\mu}_{j,h}, \tilde{\tau}^2_{j,h})$ where

$$\tilde{\mu}_{j,h} = \tilde{\tau}^2_{j,h}\left(\frac{\sum_{i:s_i=h} x_{i,j}}{\tau^2_{j,h}} + \frac{\bar{\mu}_j}{s_j^2}\right), \quad \tilde{\tau}^2_{j,h} = \left(\frac{n_h}{\tau^2_{j,h}} + \frac{1}{s_j^2}\right)^{-1}, \quad n_h = \sum_{i=1}^{n} 1(s_i = h).$$

5. Update $\tau^2_{j,h}$ for $j = 1, \ldots, p$ and $h = 1, \ldots, H$ from

$$\text{IG}\left\{\frac{n_h + 3}{2}, \frac{\sum_{i:s_i=h}(x_{i,j} - \mu_{j,h})^2 + s_j^2}{2}\right\}.$$

6. Update σ^2_h for $h = 1, \ldots, H$ from

$$\text{IG}\left\{\frac{n_h + 3}{2}, \frac{\sum_{i:s_i=h}(y_i - \tilde{x}_i'\beta_h)^2 + s_y^2}{2}\right\}.$$

7. Update $\beta_{j,h}$ for $j = 0, \ldots, p$ and $h = 1, \ldots, H$ from

$$\pi(\beta_{j,h} \mid \cdots) = \hat{p}_{j,h}\delta_0(\beta_{j,h}) + (1 - \hat{p}_{j,h})N(\beta_{j,h} \mid \mu_{\beta_{j,h}}, \sigma^2_{\beta_{j,h}}),$$

where

$$\mu_{\beta_{j,h}} = \sigma^2_{\beta_{j,h}} \left\{ \sum_{i:s_i=h} \frac{x_{i,j}\left(y_i - \tilde{x}'_{i,-j}\beta_{-j,h}\right)}{\sigma^2_h} \right\}, \quad \sigma^2_{\beta_{j,h}} = \left(\sum_{i:s_i=h} \frac{x^2_{i,j}}{\sigma^2_h} + \frac{1}{\lambda^2_{j,h}} \right)^{-1},$$

$$\hat{p}_{j,h} = \left\{ 1 + \frac{1-p_{0j}}{p_{0j}} \frac{N(0 \mid 0, \lambda^2_{j,h})}{N(0 \mid \mu_{\beta_{j,h}}, \sigma^2_{\beta_{j,h}})} \right\}^{-1}.$$

8. Update $\lambda^2_{j,h}$ for $j = 1, \ldots, p$ and $h = 1, \ldots, H$ from

$$\text{IG}\left\{ \frac{1(\beta_{j,h} \neq 0) + 1}{2}, \frac{\beta^2_{j,h} + 1}{2} \right\}.$$

9. Update p_0 from

$$\text{Be}\left\{ 4.75 + \sum_{j,h} 1(\beta_{j,h} = 0), 0.25 + \sum_{j,h} 1(\beta_{j,h} \neq 0) \right\}.$$

10. Impute missing values y_i^{mis} in the response.

    (a) Generate $y_i^* \sim N(\tilde{x}'_i\beta_{s_i}, \sigma^2_{s_i})$.

    (b) Set $y_i^{\text{mis}} = l$ if $a_l < y_i^* \leq a_{l+1}$.

11. Update latent variables y_i^* and $x_{i,j}^*$ for count and percentage variables.

    (a) For the response variable, $y_i^* \sim TN(\tilde{x}'_i\beta_{s_i}, \sigma^2_{s_i}, a_{y_i}, a_{y_{i+1}})$,

    (b) For the count predictor, $x_{i,j}^* \sim TN(\mu_{j,s_i}, \tau^2_{j,s_i}, a_{x_i}, a_{x_{i+1}})$,

    (c) (c) For the percentage predictor, $x_{i,j}^* \sim TN(\mu_{j,s_i}, \tau^2_{j,s_i}, -\infty, 0)$ if $x_{i,j} = 0$ and $x_{i,j}^* \sim TN(\mu_{j,s_i}, \tau^2_{j,s_i}, 100, \infty)$ if $x_{i,j} = 100$,

    where $TN(a, b, c, d)$ denotes a truncated normal with the location a, scale b, lower bound c and upper bound d.

12. Compute and save $\zeta_j(f, P_n)$ for $j = 1, \ldots, p$.

B.4.3 Additional estimation results

Tables B.3-B.12 show lists of the selected predictors by the proposed method for murders, rapes, robberies, assaults, burglaries, larcenies, auto thefts, arsons, violent crimes and non-violent crimes, respectively. The predictors are listed in descending order of the posterior mean of the conditional mutual information.

Table B.1: List of 1st to 34th predictors

No.	Predictor
1	population for community
2	mean people per household
3	% of population that is african american
4	% of population that is caucasian
5	% of population that is of asian heritage
6	% of population that is of hispanic heritage
7	% of population that is 16-24 in age
8	% of population that is 65 and over in age
9	# of people living in areas classified as urban
10	median household income
11	% of households with wage or salary income in 1989
12	% of households with farm or self employment income in 1989
13	% of households with investment / rent income in 1989
14	% of households with social security income in 1989
15	% of households with public assistance income in 1989
16	% of households with retirement income in 1989
17	median family income
18	per capita income
19	# of people under the poverty level
20	% of people 25 and over with less than a 9th grade education
21	% of people 25 and over that are not high school graduates
22	% of people 25 and over with a bachelors degree or higher education
23	% of people 16 and over, in the labor force, and unemployed
24	% of people 16 and over who are employed
25	% of people 16 and over who are employed in manufacturing
26	% of people 16 and over who are employed in professional services
27	% of males who are divorced
28	% of males who have never married
29	% of females who are divorced
30	% of population who are divorced
31	mean number of people per family
32	% of families (with kids) that are headed by two parents
33	% of kids in family housing with two parents
34	% of kids 4 and under in two parent households

Table B.2: List of 35th to 68th predictors

No.	Predictor
35	% of kids age 12-17 in two parent households
36	% of moms of kids 6 and under in labor force
37	% of moms of kids under 18 in labor force
38	# of kids born to never married
39	total number of people known to be foreign born
40	% of immigrants who immigated within last 5 years
41	% of population who have immigrated within the last 5 years
42	% of people who speak only English
43	% of people who do not speak English well
44	% of family households that are large (6 or more)
45	% of all occupied households that are large (6 or more people)
46	% of people in owner occupied households
47	% of persons in dense housing (more than 1 person per room)
48	% of housing units with less than 3 bedrooms
49	# of vacant households
50	% of housing occupied
51	% of households owner occupied
52	% of vacant housing that is boarded up
53	% of vacant housing that has been vacant more than 6 months
54	owner occupied housing: lower quartile value
55	owner occupied housing: median value
56	owner occupied housing: upper quartile value
57	rental housing: lower quartile rent
58	rental housing: median rent
59	rental housing: upper quartile rent
60	median gross rent
61	median gross rent as % of household income
62	# of people in homeless shelters
63	# of homeless people counted in the street
64	% of people born in the same state as currently living
65	% of people living in the same city as in 1985 (5 years before)
66	land area in square miles
67	population density in persons per square mile
68	% of people using public transit for commuting

Table B.3: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with murders as the response

j	Mean	90%CI	Predictor
66	0.2587	[0.2157, 0.2936]	land area in square miles
67	0.1188	[0.0905, 0.1454]	population density in persons per square mile
4	0.0507	[0.0302, 0.0678]	% of population that is caucasian
9	0.0250	[0.0043, 0.0636]	# of people living in areas classified as urban
1	0.0250	[0.0015, 0.0469]	population for community
3	0.0192	[0.0058, 0.0374]	% of population that is african american
57	0.0177	[0.00007, 0.0463]	rental housing: lower quartile rent
13	0.0075	[0.0004, 0.0149]	% of households with investment / rent income in 1989
6	0.0067	[0.0021, 0.0125]	% of population that is of hispanic heritage
64	0.0039	[0.0005, 0.0067]	% of people born in the same state as currently living
49	0.0030	[0.0003, 0.0089]	# of vacant households
42	0.0027	[0.0001, 0.0092]	% of people who speak only English
27	0.0019	[0.0001, 0.0055]	% of males who are divorced
52	0.0018	[0.0002, 0.0051]	% of vacant housing that is boarded up

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.4: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with rapes as the response

j	Mean	90%CI	Predictor
66	0.4168	[0.3929, 0.4428]	land area in square miles
67	0.1964	[0.1727, 0.2217]	population density in persons per square mile
1	0.0680	[0.0523, 0.0865]	population for community
9	0.0359	[0.0086, 0.0608]	# of people living in areas classified as urban
30	0.0189	[0.0013, 0.0379]	% of population who are divorced
32	0.0178	[0.0009, 0.0398]	% of families (with kids) that are headed by two parents
33	0.0174	[0.0006, 0.0389]	% of kids in family housing with two parents
29	0.0156	[0.0005, 0.0330]	% of females who are divorced
27	0.0123	[0.0009, 0.0265]	% of males who are divorced
39	0.0051	[0.0004, 0.0118]	total number of people known to be foreign born
5	0.0046	[0.0002, 0.0092]	% of population that is of asian heritage
35	0.0031	[0.0001, 0.0125]	% of kids age 12-17 in two parent households
7	0.0027	[0.0008, 0.0056]	% of population that is 16-24 in age
50	0.0023	[0.0002, 0.0053]	% of housing occupied
12	0.0022	[0.0001, 0.0059]	% of households with farm or self employment income in 1989
19	0.0021	[0.0003, 0.0062]	# of people under the poverty level
38	0.0017	[0.0001, 0.0065]	# of kids born to never married
18	0.0015	[0.00008, 0.0050]	per capita income
28	0.0014	[0.0002, 0.0036]	% of males who have never married
63	0.0011	[0.0002, 0.0020]	# of homeless people counted in the street

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.5: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with robberies as the response

j	Mean	90%CI	Predictor
66	0.6074	[0.5551, 0.6554]	land area in square miles
67	0.5080	[0.4548, 0.5605]	population density in persons per square mile
33	0.0859	[0.0545, 0.1203]	% of kids in family housing with two parents
4	0.0652	[0.0353, 0.0953]	% of population that is caucasian
3	0.0530	[0.0211, 0.0865]	% of population that is african american
9	0.0469	[0.0078, 0.0926]	# of people living in areas classified as urban
1	0.0388	[0.0268, 0.0623]	population for community
47	0.0277	[0.0084, 0.0493]	% of persons in dense housing
30	0.0159	[0.0007, 0.0348]	% of population who are divorced
18	0.0139	[0.0009, 0.0326]	per capita income
32	0.0122	[0.0006, 0.0340]	% of families (with kids) that are headed by two parents
29	0.0107	[0.0002, 0.0258]	% of females who are divorced
6	0.0106	[0.0006, 0.0237]	% of population that is of hispanic heritage
64	0.0094	[0.0045, 0.0146]	% of people born in the same state as currently living
42	0.0090	[0.0002, 0.0217]	% of people who speak only English
22	0.0079	[0.0001, 0.0198]	% of people 25 and over with a bachelors degree or higher education
46	0.0071	[0.0006, 0.0183]	% of people in owner occupied households
56	0.0064	[0.0001, 0.0182]	owner occupied housing: upper quartile value
25	0.0062	[0.0002, 0.0125]	% of people 16 and over who are employed in manufacturing
68	0.0055	[0.0015, 0.0099]	% of people using public transit for commuting
34	0.0054	[0.0004, 0.0183]	% of kids 4 and under in two parent households
51	0.0050	[0.0006, 0.0142]	% of households owner occupied
19	0.0030	[0.0003, 0.0072]	# of people under the poverty level
38	0.0029	[0.0005, 0.0077]	# of kids born to never married
49	0.0021	[0.0001, 0.0056]	# of vacant households

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.6: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with assaults as the response

j	Mean	90%CI	Predictor
66	0.3380	[0.2897, 0.3914]	land area in square miles
67	0.1760	[0.1318, 0.2267]	population density in persons per square mile
9	0.0760	[0.0451, 0.0996]	# of people living in areas classified as urban
1	0.0413	[0.0186, 0.0641]	population for community
33	0.0350	[0.0114, 0.0571]	% of kids in family housing with two parents
13	0.0348	[0.0234, 0.0478]	% of households with investment / rent income in 1989
32	0.0176	[0.0010, 0.0403]	% of families (with kids) that are headed by two parents
47	0.0171	[0.0057, 0.0283]	% of persons in dense housing
4	0.0168	[0.0046, 0.0284]	% of population that is caucasian
3	0.0070	[0.0004, 0.0174]	% of population that is african american
43	0.0050	[0.0013, 0.0102]	% of people who do not speak English well
45	0.0027	[0.0003, 0.0074]	% of all occupied households that are large
50	0.0025	[0.0007, 0.0046]	% of housing occupied
34	0.0024	[0.0001, 0.0075]	% of kids 4 and under in two parent households
44	0.0023	[0.0003, 0.0064]	% of family households that are large
23	0.0014	[0.0001, 0.0041]	% of people 16 and over, in the labor force, and unemployed

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.7: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with burglaries as the response

j	Mean	90%CI	Predictor
66	0.9177	[0.8717, 0.9492]	land area in square miles
67	0.7075	[0.6639, 0.7464]	population density in persons per square mile
33	0.0508	[0.0241, 0.0796]	% of kids in family housing with two parents
47	0.0281	[0.0146, 0.0444]	% of persons in dense housing
29	0.0173	[0.0100, 0.0276]	% of females who are divorced
50	0.0152	[0.0071, 0.0236]	% of housing occupied
13	0.0135	[0.0008, 0.0303]	% of households with investment / rent income in 1989
6	0.0097	[0.00007, 0.0166]	% of population that is of hispanic heritage
30	0.0083	[0.0001, 0.0224]	% of population who are divorced
9	0.0078	[0.0004, 0.0258]	# of people living in areas classified as urban
4	0.0070	[0.0007, 0.0163]	% of population that is caucasian
68	0.0057	[0.0004, 0.0126]	% of people using public transit for commuting
65	0.0048	[0.0001, 0.0116]	% of people living in the same city as in 1985
49	0.0046	[0.0005, 0.0125]	# of vacant households
7	0.0031	[0.0002, 0.0066]	% of population that is 16-24 in age
19	0.0028	[0.0001, 0.0110]	# of people under the poverty level
61	0.0024	[0.00008, 0.0069]	median gross rent as % of household income
36	0.0008	[0.00001, 0.0025]	% of moms of kids 6 and under in labor force

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.8: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with larcenies as the response

j	Mean	90%CI	Predictor
66	0.9425	[0.9149, 0.9682]	land area in square miles
67	0.8035	[0.7707, 0.8359]	population density in persons per square mile
32	0.0305	[0.0003, 0.0505]	% of families (with kids) that are headed by two parents
2	0.0233	[0.0126, 0.0397]	mean people per household
22	0.0219	[0.00001, 0.0436]	% of people 25 and over with a bachelors degree or higher education
35	0.0217	[0.0085, 0.0383]	% of kids age 12-17 in two parent households
65	0.0165	[0.0062, 0.0256]	% of people living in the same city as in 1985
8	0.0163	[0.0008, 0.0321]	% of population that is 65 and over in age
45	0.0135	[0.00002, 0.0520]	% of all occupied households that are large
33	0.0133	[0.00002, 0.0422]	% of kids in family housing with two parents
7	0.0106	[0.0002, 0.0180]	% of population that is 16-24 in age
68	0.0105	[0.0062, 0.0154]	% of people using public transit for commuting
25	0.0084	[0.0056, 0.0111]	% of people 16 and over who are employed in manufacturing
47	0.0070	[0.0001, 0.0178]	% of persons in dense housing
23	0.0054	[0.0004, 0.0103]	% of people 16 and over, in the labor force, and unemployed
42	0.0054	[0.0003, 0.0163]	% of people who speak only English
64	0.0053	[0.0005, 0.0123]	% of people born in the same state as currently living
5	0.0038	[0.0001, 0.0077]	% of population that is of asian heritage
14	0.0022	[0.0003, 0.0056]	% of households with social security income in 1989

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.9: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with auto thefts as the response

j	Mean	90%CI	Predictor
66	0.7650	[0.7310, 0.8011]	land area in square miles
67	0.6471	[0.6098, 0.6847]	population density in persons per square mile
47	0.0298	[0.0164, 0.0437]	% of persons in dense housing
30	0.0245	[0.0008, 0.0541]	% of population who are divorced
18	0.0229	[0.0001, 0.0626]	per capita income
13	0.0211	[0.0054, 0.0405]	% of households with investment / rent income in 1989
46	0.0197	[0.00001, 0.0899]	% of people in owner occupied households
60	0.0138	[0.0054, 0.0342]	median gross rent
53	0.0119	[0.0050, 0.0178]	% of vacant housing that has been vacant more than 6 months
4	0.0095	[0.0004, 0.0214]	% of population that is caucasian
42	0.0087	[0.0014, 0.0190]	% of people who speak only English
12	0.0081	[0.0045, 0.0120]	% of households with farm or self employment income in 1989
2	0.0081	[0.0004, 0.0234]	mean people per household
68	0.0075	[0.0022, 0.0138]	% of people using public transit for commuting
40	0.0071	[0.0032, 0.0144]	% of immigrants who immigated within last 5 years
43	0.0041	[0.00005, 0.0123]	% of people who do not speak English well
58	0.0034	[0.0007, 0.0106]	rental housing: median rent
59	0.0030	[0.0005, 0.0138]	rental housing: upper quartile rent
57	0.0022	[0.0005, 0.0057]	rental housing: lower quartile rent
50	0.0021	[0.0002, 0.0047]	% of housing occupied

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.10: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with arsons as the response

j	Mean	90%CI	Predictor
66	0.3030	[0.2517, 0.3593]	land area in square miles
67	0.1619	[0.1226, 0.2084]	population density in persons per square mile
1	0.0394	[0.0131, 0.0689]	population for community
9	0.0152	[0.0010, 0.0471]	# of people living in areas classified as urban
19	0.0131	[0.0005, 0.0323]	# of people under the poverty level
27	0.0119	[0.0022, 0.0229]	% of males who are divorced
13	0.0085	[0.0004, 0.0168]	% of households with investment / rent income in 1989
29	0.0078	[0.0001, 0.0212]	% of females who are divorced
41	0.0039	[0.0013, 0.0071]	% of population who have immigrated within the last 5 years
15	0.0031	[0.0004, 0.0065]	% of households with public assistance income in 1989

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.11: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with violent crimes as the response

j	Mean	90%CI	Predictor
66	0.5254	[0.4868, 0.5763]	land area in square miles
67	0.3515	[0.3106, 0.4052]	population density in persons per square mile
9	0.1004	[0.0589, 0.1498]	# of people living in areas classified as urban
33	0.0751	[0.0412, 0.1058]	% of kids in family housing with two parents
47	0.0272	[0.0140, 0.0417]	% of persons in dense housing
32	0.0242	[0.0012, 0.0581]	% of families (with kids) that are headed by two parents
13	0.0242	[0.0094, 0.0451]	% of households with investment / rent income in 1989
4	0.0163	[0.0029, 0.0329]	% of population that is caucasian
1	0.0153	[0.0003, 0.0394]	population for community
3	0.0137	[0.0014, 0.0278]	% of population that is african american
15	0.0080	[0.00002, 0.0165]	% of households with public assistance income in 1989
6	0.0080	[0.0004, 0.0195]	% of population that is of hispanic heritage
43	0.0053	[0.0013, 0.0125]	% of people who do not speak English well
68	0.0036	[0.0004, 0.0072]	% of people using public transit for commuting
49	0.0031	[0.0001, 0.0101]	# of vacant households
50	0.0031	[0.0007, 0.0067]	% of housing occupied
62	0.0027	[0.0002, 0.0068]	# of people in homeless shelters
38	0.0025	[0.00007, 0.0103]	# of kids born to never married
45	0.0024	[0.0004, 0.0072]	% of all occupied households that are large
44	0.0023	[0.0005, 0.0068]	% of family households that are large
31	0.0020	[0.00009, 0.0077]	mean number of people per family
41	0.0018	[0.00009, 0.0051]	% of population who have immigrated within the last 5 years
5	0.0017	[0.00008, 0.0042]	% of population that is of asian heritage
23	0.0013	[0.00008, 0.0038]	% of people 16 and over, in the labor force, and unemployed

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Table B.12: List of the selected predictors by the proposed method in descending order of the posterior means of conditional mutual information with non-violent crimes as the response

j	Mean	90%CI	Predictor
66	0.9859	[0.9500, 1.0189]	land area in square miles
67	0.8282	[0.7870, 0.8700]	population density in persons per square mile
32	0.0300	[0.0082, 0.0518]	% of families (with kids) that are headed by two parents
28	0.0217	[0.0011, 0.0475]	% of males who have never married
33	0.0217	[0.0015, 0.0484]	% of kids in family housing with two parents
9	0.0200	[0.0017, 0.0518]	# of people living in areas classified as urban
30	0.0183	[0.0006, 0.0399]	% of population who are divorced
27	0.0182	[0.0001, 0.0443]	% of males who are divorced
47	0.0181	[0.0043, 0.0353]	% of persons in dense housing
1	0.0174	[0.0001, 0.0426]	population for community
29	0.0086	[0.0003, 0.0223]	% of females who are divorced
64	0.0072	[0.0007, 0.0155]	% of people born in the same state as currently living
50	0.0039	[0.0010, 0.0075]	% of housing occupied
52	0.0023	[0.0001, 0.0058]	% of vacant housing that is boarded up

j, j-th predictor; Mean, posterior mean; 90%CI refers to a 90% credible interval.

Bibliography

Agresti, A. (2002), *Categorical Data Analysis.*, Wiley, New York, second edn.

Albert, J. H. and Chib, S. (2001), "Sequential ordinal modeling with applications to survival data," *Biometrics*, 57, 829–836.

An, Q., Wang, C., Shterev, I., Wang, E., Carin, L., and Dunson, D. B. (2008), "Hierarchical kernel stick-breaking process for multi-task image analysis," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 17–24, Helsinki, Finland.

Baghfalakia, T., Ganjalia, M., and Berridge, D. (2014), "Joint modeling of multivariate longitudinal mixed measurements and time to event data using a Bayesian approach," *Journal of Applied Statistics*, 41, 1934–1955.

Bandyopadhyay, S., Ganguli, B., and Chatterjee, A. (2011), "A review of multivariate longitudinal data analysis," *Statistical Methods in Medical Research*, 20, 299330.

Bellhouse, D. R. and Stafford, J. E. (1999), "Density estimation from complex surveys," *Statistica Sinica*, 9, 407–424.

Bhattacharya, A. and Dunson, D. B. (2012), "Simplex factor models for multivariate unordered categorical data," *Journal of the American Statistical Association*, 107, 362–377.

Bouezmarni, T., Rombouts, J. V. K., and Taamouti, A. (2012), "Nonparametric Copula-Based Test for Conditional Independence with Applications to Granger Causality," *Journal of Business & Economic Statistics*, 30, 275–287.

Burgette, L. F. and Nordheim, E. V. (2012), "The Trace Restriction: An Alternative Identification Strategy for the Bayesian Multinomial Probit Model," *Journal of Business and Economic Statistics*, 30, 404–410.

Bush, C. and MacEachern, S. (1996), "A semiparametric Bayesian model for randomised block designs," *Biometrika*, 83, 275–285.

Buskirk, T. D. (1998), "Nonparametric density estimation using complex survey data," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 799–801, Washington, DC.

Buskirk, T. D. and Lohr, S. L. (2005), "Asymptotic properties of kernel density estimation with complex survey data," *Journal of Statistical Planning and Inference*, 128, 165–190.

Canale, A. and Dunson, D. B. (2011), "Bayesian Kernel Mixtures for Counts," *Journal of the American Statistical Association*, 106, 1528–1539.

Carter, C. K. and Kohn, R. (1994), "On Gibbs sampling for state space models," *Biometrika*, 81, 541–553.

Chen, Q., Elliott, M. R., and Little, R. J. A. (2010), "Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling," *Survey Methodology*, 36, 23–34.

Chung, Y. and Dunson, D. B. (2009), "Nonparametric Bayes conditional distribution modeling with variable selection," *Journal of the American Statistical Association*, 104, 1646–1660.

Chung, Y. and Dunson, D. B. (2011), "The local Dirichlet process," *Annals of the Institute for Statistical Mathematics*, 63, 59–80.

Cover, T. M. and Thomas, J. A. (2006), *Elements of Information Theory*, John Wiley & Sons, New York.

Dahinden, C., Parmigiani, G., Emerick, M., and Buehlmann, P. (2007), "Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries," *BMC Bioinformatics*, 8, 476.

Dahinden, C., Kalisch, M., and Buehlmann, P. (2010), "Decomposition and model selection for large contingency tables," *Biometrical Journal*, 52, 233–252.

de Jong, P. and Shephard, N. (1995), "The simulation smoother for time series models," *Biometrika*, 82, 339–350.

Diks, C. and DeGoede, J. (2001), "A general nonparametric bootstrap test for Granger causality," pp. 391–403, in: Broer, Krauskopf, Vegter (Eds.), Global Analysis of Dynamical Systems, Chapter 16.

Dobra, A. and Lenkoski, A. (2011), "Copula Gaussian graphical models and their application to modeling functional disability data," *Annals of Applied Statistics*, 5, 969–993.

Dunson, D. B. (2000), "Bayesian Latent Variable Models for Clustered Mixed Outcomes," *Journal of the Royal Statistical Society-B*, 62, 355–366.

Dunson, D. B. (2003), "Dynamic Latent Trait Models for Multidimensional Longitudinal Data," *Journal of the American Statistical Association*, 98, 555–563.

Dunson, D. B. (2006), "Bayesian dynamic modeling of latent trait distributions," *Biostatistics*, 7, 551–568.

Dunson, D. B. and Park, J. H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95, 307–323.

Dunson, D. B. and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051.

Durbin, J. and Koopman, S. J. (2002), "Simple and efficient simulation smoother for state space time series analysis," *Biometrika*, 89, 603–616.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Fienberg, S. and Rinaldo, A. (2007), "Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation," *Journal of Statistical Planning and Inference*, 137, 3430–3445.

Fieuws, S. and Verbeke, G. (2006), "Pairwise Fitting of Mixed Models for the Joint Modeling of Multivariate Longitudinal Profiles," *Biometrics*, 62, 424–431.

Fox, E. B. and Dunson, D. B. (2011), "Pairwise Fitting of Mixed Models for the Joint Modeling of Multivariate Longitudinal Profiles," Technical report.

Früwirth-Schnatter, S. (1994), "Data augmentation and dynamic linear models," *Journal of Time Series Analysis*, 15, 183–202.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008), "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems 21*, pp. 489–496.

Gelman, A. (2007), "Struggles with survey weighting and regression modeling," *Statistical Science*, 22, 153–164.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), "Posterior consistency of Dirichlet mixtures in density estimation," *Annals of Statistics*, 27, 143–158.

Ghosh, J. and Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, Springer Verlag.

Ghosh, P. and Hanson, T. (2010), "A Semiparametric Bayesian Approach to Multivariate Longitudinal Data," *Australian & New Zealand Journal of Statistics*, 52, 275–288.

Goodman, L. A. and Kruskal, W. H. (1954), "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, 49, 732764.

Goodman, L. A. and Kruskal, W. H. (1959), "Measures of Association for Cross Classifications. II: Further Discussion and References," *Journal of the American Statistical Association*, 54, 123163.

Goodman, L. A. and Kruskal, W. H. (1963), "Measures of Association for Cross Classifications III: Approximate Sampling Theory," *Journal of the American Statistical Association*, 58, 310–364.

Goodman, L. A. and Kruskal, W. H. (1972), "Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances," *Journal of the American Statistical Association*, 67, 415421.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179–194.

Gruhl, J., Erosheva, E. A., and Crane, P. K. (2013), "A Semiparametric Approach to Mixed Outcome Latent Variable Models: Estimating the Association between Cognition and Regional Brain Volumes," *Annals of Applied Statistics*, 7, 2361–2383.

Gueorguieva, R. V. and Sanacora, G. (2006), "Joint analysis of repeatedly observed continuous and ordinal measures of disease severity," *Statistics in Medicine*, 25, 1307–1322.

Györfi, L. and Walk, H. (2012), "Strongly consistent nonparametric tests of conditional independence," *Statistics & Probability Letters*, 82, 1145–1150.

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011), "Dirichlet Process Mixtures of Generalized Linear Models," *Journal of Machine Learning Research*, 12, 1923–1953.

Harris, K. M., Halpern, C. T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and Udry, J. R. (2009), "The National Longitudinal Study of Adolescent Health: Research Design [WWW document]," URL: http://www.cpc.unc.edu/projects/addhealth/design.

Harshman, R. A. (1970), "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, 16, 1–84.

Hoff, P. D. (2007), "Extending the rank likelihood for semiparametric copula estimation," *Annals of Applied Statistics*, 1, 265–283.

Hoff, P. D. and Niu, X. (2012), "A Covariance Regression Model," *Statistica Sinica*, 22, 729–753.

Horvitz, D. G. and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of American Statistical Association*, 47, 663–685.

Imai, K. and van Dyk, D. (2005), "A Bayesian analysis of the multinomial probit model using marginal data augmentation," *Journal of Econometrics*, 124, 311–334.

Ishwaran, H. and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Joe, H. (1989), "Relative Entropy Measures of Multivariate Dependence," *Journal of the American Statistical Association*, 84, 157–164.

Johndrow, J., Dunson, D. B., and Lum, K. (2013), "Diagonal Orthant Multinomial Probit Models," in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pp. 29–38, Scottsdale, AZ, USA.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 65, 93–105.

Kim, J. S. and Ratchford, B. T. (2013), "A Bayesian multivariate probit for ordinal data with semiparametric random-effects," *Computational Statistics & Data Analysis*, 64, 192–208.

Kolda, T. G. (2001), "Orthogonal tensor decompositions," *SIAM Journal on Matrix Analysis and Applications*, 23, 243–255.

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer.

Kottas, A., M. P. Q. F. (2005), "Nonparametric Bayesian Modeling for Multivariate Ordinal Data," *Journal of Computational and Graphical Statistics*, 14, 610–625.

Kunihama, T. and Dunson, D. B. (2013), "Bayesian modeling of temporal dependence in large sparse contingency tables," *Journal of the American Statistical Association*, 108, 1324–.

Kunihama, T. and Dunson, D. B. (2014), "Nonparametric Bayes inference on conditional independence," Technical report.

Kunihama, T., Herring, A. H., Halpern, C. T., and Dunson, D. B. (2014), "Nonparametric Bayes modeling with sample survey weights," Technical report.

Levy, P. S. and Lemeshow, S. (2008), *Sampling of Populations: Methods and Applications*, Wiley.

Little, R. J. A. (2004), "To model or not to model? Competing modes of inference for finite population sampling," *Journal of the American Statistical Association*, 99, 546–556.

Liu, X., Daniels, M. J., and Marcus, B. (2009), "Joint Models for the Association of Longitudinal Binary and Continuous Processes With Application to a Smoking Cessation Trial," *Journal of the American Statistical Association*, 104, 429–438.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *Annals of Statistics*, 12, 351–357.

Luo, S. and Wang, J. (2014), "Bayesian hierarchical model for multiple repeated measures and survival data: an application to Parkinson's disease," *Statistics in Medicine*, 33, 4279–4291.

Ma, L. (2013), "Adaptive testing of conditional association through recursive mixture modeling," *Journal of the American Statistical Association*, 108, 1493–1505.

MacEachern, S. N. (1999), "Dependent Nonparametric Processes," *In ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association*, pp. 50–55.

MacEachern, S. N. (2000), "Dependent Dirichlet processes," Technical report, Ohio State University, Department of Statistics.

MacKay, D. J. C. (2003), *Information Theory, Inference and Learning Algorithms*, Cambridge University Press.

Manrique-Vallier, D. and Reiter, J. (2012), "Bayesian Estimation of Discrete Truncated Latent Structure Models," Technical report.

McCulloch, R. and Rossi, P. (1994), "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64, 207–240.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W., and Collinson, M. A. (2014), "Clustering South African households based on their asset status using latent variable models," *Annals of Applied Statistics*, 8, 747776.

Moustaki, I. and Knott, M. (2000), "Generalized Latent Trait Models," *Psychometrika*, 65, 391–411.

Muliere, P. and Tardella, L. (1998), "Approximating Distributions of Random Functionals of Ferguson Dirichlet Priors," *Canadian Journal of Statistics*, 26, 283297.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Murray, J. S. and Reiter, J. P. (2014), "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence," Technical report.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013), "Bayesian Gaussian copula factor models for mixed data," *Journal of the American Statistical Association*, 108, 656–665.

Muthén, B. (1984), "A General Structural Equation Model With Dichotomous, Ordered Categorical and Continuous Latent Variable Indicators," *Psychometrika*, 49, 115–132.

Papaspiliopoulos, O. and Roberts, G. O. (2008), "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models," *Biometrika*, 95, 169–186.

Pérez-Cruz, F. (2008), "Estimation of Information Theoretic Measures for Continuous Random Variables," in *Advances in Neural Information Processing Systems 21*, pp. 1257–1264.

Rao, J. N. K. (2011), "Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal," *Statistical Science*, 26, 240–256.

Reich, B. J., Kalendra, E., Storlie, C. B., Bondell, H. D., and Fuentes, M. (2012), "Variable selection for high dimensional Bayesian density estimation: application to human exposure simulation," *Journal of the Royal Statistical Society Series C*, 61, 47–66.

Ren, L., Dunson, D., Lindroth, S., and Carin, L. (2010), "Dynamic Nonparametric Bayesian Models for Analysis of Music," *Journal of the American Statistical Association*, 105, 458–472.

Ritter, C. and Tanner, M. A. (1992), "Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy Gibbs sampler," *Journal of American Statistical Association*, 87, 861–868.

Rodriguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6, 145–177.

Rodriguez, A. and Horst, E. T. (2008), "Bayesian dynamic density estimation," *Bayesian Analysis*, 3, 339–366.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997), "Latent Variable Models for Mixed Discrete and Continuous Outcomes," *Journal of the Royal Statistical Society-B*, 59, 667–678.

Seth, S. and Principe, J. C. (2010), "A conditional distribution function based approach to design nonparametric tests of independence and conditional independence," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2066–2069.

Seth, S. and Príncipe, J. C. (2012a), "Assessing Granger non-causality using nonparametric measure of conditional independence," *IEEE Transactions on Neural Networks and Learning Systems*, 23, 47–59.

Seth, S. and Príncipe, J. C. (2012b), "Conditional association," *Neural Computation*, 24, 1882–1905.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Si, Y., Pillai, N., and Gelman, A. (2014), "Bayesian nonparametric weighted sampling inference," Technical report.

Song, K. (2009), "Testing conditional independence via Rosenblatt transforms," *Annals of Statistics*, 37, 4011–4045.

Su, L. and White, H. (2007), "A consistent characteristic function-based test for conditional independence," *Journal of Econometrics*, 141, 807–834.

Su, L. and White, H. (2008), "A nonparametric Hellinger metric test for conditional independence," *Econometric Theory*, 24, 829–864.

Tokdar, S. T. (2006), "Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression," *Sankhya*, 67, 90–110.

van der Vaart, A. and Wellner, J. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer.

van der Vaart, A. and Wellner, J. (2000), "Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes," 47, 115–133, in: High Dimensional Probability II, Progress in Probability.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014), "The analysis of multivariate longitudinal data: A review," *Statistical Methods in Medical Research*, 23, 42–59.

Walker, S. G. (2007), "Sampling the Dirichlet Mixture Model With Slices," *Communications in Statistics-Simulation and Computation*, 36, 45–54.

West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical priors and mixture models, with application in regression and density estimation," in *Aspects of uncertainty: A Tribute to DV Lindley*, eds. P. R. Freeman and A. F. M. Smith, pp. 363–386, Wiley.

Wu, Y. and Ghosal, S. (2008), "Kullback Leibler property of kernel mixture priors in Bayesian density estimation," *Electronic Journal of Statistics*, 2, 298–331.

Wyner, A. D. (1978), "A definition of conditional mutual information for arbitrary ensembles," *Information and Control*, 38, 51–59.

Zangeneh, S. Z. and Little, R. J. (2012), "Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample," *Proceedings of the Joint Statistical Meetings*.

Zheng, H. and Little, R. J. A. (2003), "Penalized spline model-based estimation of finite population total from probability-proportionalto-size samples," *Journal of Official Statistics*, 19, 99–107.

Zheng, H. and Little, R. J. A. (2005), "Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model," *Journal of Official Statistics*, 21, 1–20.

Biography

Tsuyoshi Kunihama was born in Tokushima, Japan. He received his bachelor degree and master degree in Economics from University of Tokyo in 2007 and 2009. He graduated from Duke with his Ph.D. in statistics in 2015, under the direction of David B. Dunson.