BAYESIAN METHODS FOR CORRELATED PREDICTORS AND CONFOUNDING VARIABLES IN EPIDEMIOLOGY

Angel Davalos

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Public Health.

Chapel Hill
2019

Approved by:

Amy H. Herring

Jianwen Cai

Annie Green Howard

Andrew Olshan

Haibo Zhou

ABSTRACT

Angel Davalos: Bayesian Methods for Correlated Predictors and Confounding Variables in Epidemiology
(Under the direction of Amy H. Herring)

In epidemiology, it is common to have a set of outcomes, exposures, and confounding variables on different scales (i.e. continuous, count, categorical: nominal/ordinal). Confounding variables are expect to be correlated with exposures and at times exposures may be highly correlated among themselves which present model estimation complications. This is especially prevalent in environmental epidemiology, where studying the joint or simultaneous effect of chemical mixture or air pollution exposures on health for example is of interest. Dimension reduction techniques and shrinkage effect estimation are important tools to overcome these difficulties.

Studying the multivariate dependence among mixed scale variables can aid investigators in developing analysis plans but mixed-scale distribution modeling is not a simple task. Specifically, it may be of interest to assess and quantify the degree of correlation among variables and or characterize different exposure-confounding variable profiles. Certain dimension reduction methods, such as, mixture models can do both, as well as, jointly model variables of mixed-scales. Shrinkage methods, on the other hand, do not transform a set of correlated variables but implement a bias-variance trade-off to address effect estimation.

The overall goal of this research is to develop a suite of Bayesian methods for clustering via mixed-scale distributional modeling and variable selection. First, motivated by sophisticated Bayesian mixed-scale distribution modeling, we develop a joint model using modularized tensor factorization (MOTEF) as a simplification for ease of implementation and computation. The performance of MOTEF is assessed via a simulation study and applied to data from the National Birth Defects and Prevention Study (NBDPS) for mixed-scale multivariate profiling. Second, we develop a Bayesian semi-parametric model with variable selection for hierarchical interactions (BHIS). Its performance is assessed via simulation studies and applied to the Mount Sinai Children's Environmental Health Study. Lastly, building on Bayesian mixed-scale distribution modeling, we develop a joint mixture model for compositional data with essential zeros. The model is applied to accelerometry-assessed sedentary behavior and physical activity data from the Hispanic Community Health Study / Study of Latinos for describing activity profiles and health risk.

To my wife, Saraí, and kids, Alina Elienaí, Gianna Maria, Judit Saraí, and Mario Angel.

ACKNOWLEDGEMENTS

I would like to thank everyone who in ways big or small helped me through the completion of this work.

To the UNC Biostatistics Department, thank you for extending me the opportunity to become a part of the department and providing me all the necessary support. I would like to specifically thank my advisor, Dr. Herring, for her outstanding mentoring and guidance. She has been an inspiration and example to follow as researcher, teacher, and person for her dedication to and empathy for students. I thank my committee members, Dr. Cai, Dr. Howard, Dr. Olshan, and Dr. Zhou, for their time, patience, and comments. Thank you to Dr. Sotres-Alvarez whose help and encouragement were an invaluable part of this work. A thank you to Dr. Hudgens and Katie Mollan for the opportunity to join the CFAR. I would also like to thank my BIOS peers and the administrative staff.

To my family, thank you for blessing me with all the love and care through the difficult moments. Thank you to my mom who encouraged me to try and for incessantly praying for me. I also thank my wife and kids for being my motivation through this seemingly endless journey, and for giving me something to look forward to when the work was crushing my spirits. We made it through together. I also express my great appreciation to my in-laws for their sacrifices to help our family. Thank you to my siblings for emphasizing academics growing up and encouraging me throughout the years. Lastly and most importantly, I am forever grateful to God for all the blessings that made the completion of my doctoral studies possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MOTEF	Modularized tensor factorizations
MPK	Mixture of product kernels
BBVS	Bayesian blocked variable selection
BHIS	Bayesian hierarchical interaction selection
LASSO	Least absolute shrinkage and selection operator
GLINTERNET	Grouped lasso interaction network
BTMPK	Bayesian tensor mixture of product kernels on the simplex
BMPK	Bayesian mixture of product kernels on the simplex
BMMG	Bayesian mixture of multivariate Gaussians on the simplex

CHAPTER 1: INTRODUCTION

In environmental epidemiology, exposures and subject characteristics are measured on numerous scales. For instance, exposure to phthalates, a potential endocrine disruptor, is often quantified on a continuous scale as urine metabolite concentration, while subject characteristics such as gender and accelerometer-assessed physical activity are categorical and compositional, respectively. Atop exposures and characteristics being measured on different scales, some variables are highly correlated which introduce difficulties in estimation.

Recently, as the field of environmental epidemiology has moved toward studying simultaneous effects of highly correlated exposures and characteristics (e.g. environment gene interaction), from previous emphasis on effect of single exposures, novel statistical methods have been proposed to address this issue. One such technique is dimension reduction, often two step procedures, where predictors are transformed into a set of ideally lesser correlated variables, which are then used to model a health outcome of interest. Other methods attempt to stabilize the estimation of effect parameters by incorporating a degree of bias into the estimation procedure. These methods are generally referred to as shrinkage methods such as penalized regression or Bayesian methods. In this work, we focus our attention on particular dimension reduction and shrinkage methods.

The mixture model is often used as a dimension reduction technique because it offers distributional (or model-based) clustering. Distributional clustering, as opposed to distance-based clustering, models the joint distribution of a set of variables assuming a heterogeneous population. The clustering of a set of variables allows investigators the capability to profile the characteristics of subpopulations in the data. Some mixture models, such as the mixture of product kernels, additionally have the capability to jointly model mixed-scale variables which is highly useful for environmental epidemiology. However, mixed-scale distribution modeling requires non-standard ways of constructing a distribution. In the subsequent chapter, we review the literature on mixed-scale distribution modeling. At the end of the chapter we introduce a section on compositional data as a case study of a non-standard scale.

When it is of interest to quantify the effect of each predictor in assessing joint effects, certain shrinkage methods can both stabilize estimation and perform variable selection. To complicate matters more in the presence of highly correlated predictors, there may be non-homogeneous effects where it is necessary to assess the degree of interaction among the predictors. Certain shrinkage methods, such as penalized and Bayesian

techniques, have the capability to screen interactions while preserving hierarchy. Bayesian techniques offer greater modeling control in the type of bias introduced over penalized regression via the prior distributions assumed on model parameters. In the next chapter, we briefly review hierarchical interactions, and certain penalized and Bayesian methods.

Motivated by statistical methods for studying simultaneous health effects of correlated exposures and subject characteristics, the overall objective of this dissertation is to develop a suite of Bayesian methods for addressing modeling challenges in the presence of correlated predictors via mixed-scale distribution modeling and shrinkage methods. The following chapter reviews the literature of basic elements needed for understanding our proposed methods, and the subsequent chapters present our proposed methods which comprise our three paper dissertation work.

CHAPTER 2: REVIEW OF LITERATURE

2.1 Mixed-scale distribution modeling

The topic of modeling the distribution of mixed scale variables has roots in psychology where correlation models were used to assess the association between categorical and continuous variables (Tate, 1954; Olkin and Tate, 1961). Although these may not have yielded a unified modeling of the joint distribution between variables of different scales, it led to the eventual modeling of such for complete and missing data (Krzanowski, 1980; Little and Schluchter, 1985). During the 1980s and 1990s, mixed scale distribution modeling broadened its focus to include conditional mixed scale distribution modeling where the a mixed scale multivariate vector is modeled as an outcome conditional on covariates of interest, with a focus on repeated measures for applications in clustered and longitudinal studies. The latter part of this review focuses on developments within Bayesian non-parametric statistics as this is an area where most recent developments have occurred.

Many approaches have been proposed for modeling multivariate mixed-scale data but a standard distribution does not exist. The approaches can be broadly grouped into two categories: (1) product conditional-marginal modeling and (2) joint modeling. Consider the a multivariate mixed scale vector, $\boldsymbol{y} = (y_1, \ldots, y_p)'$, which can consist of multiple scales where s_j denotes the scale of the variable. For simplicity, let $s_j \in \{1, 2, 3\}$ which correspond to categorical with bounded support, count or ordinal categorical with unbounded support, and continuous variables with support on $\mathcal{R}$, respectively. The product conditional-marginal modeling approach defines a joint distribution though the product of a distribution of a set of variables of uniform scale conditional on another set of uniform scale variables times the marginal distribution of the latter set of variables. The joint modeling approach simply builds a joint model by incorporating various techniques that include: estimating equations, kernel density estimation, copulas, and latent variable approaches. The latent variable approach is also used within the first approach, however, it is commonly used to define a conditional or marginal distribution.

2.1.1 Product Conditional-Marginal Distributions

Among product conditional-marginal modeling, the order in which the different sets of variables are defined to take the conditional or marginal distribution define the model. We first begin by considering a case

where the categorical variables of a mixed scale vector are modeled conditional on the set set of continuous variables and the continuous variables are modeled marginally. For illustrative purposes let us assume a simple case where in a mixed scale vector $\boldsymbol{y}$ there is only one binary variable $s_1 = 1$ and the rest are continuous $s_j = 3$ for $j = 2, \ldots, p$. One possible model definition could be to assume a logistic regression on the conditional model, $\text{logit}(\gamma) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{y}_{(-1)}$ with $\text{E}(y_1) = \gamma$ and $\boldsymbol{y}_{(-1)}$ denotes all variables except the first, and $\boldsymbol{y}_{(-1)} \sim \text{N}_{p-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This is a special case of a general modeling framework first proposed by Cox and Wermuth (1992) called conditional Gaussian regression for mixed binary and continuous variables (Cox and Wermuth, 1992). This type of modeling can be extended to include categorical variables with bounded and unbounded support as well possibly via a three tier hierarchical structure and using say multivariate polytomous regression in tandem with Poisson regression and the marginal multivariate normal for continuous variables. However, this type of modeling involves a degree of subjectivity in defining the conditional distributions where a natural choice can be standard mean regression arguments.

The converse of the conditional Gaussian regression approach has also seen more study and development. One approach, termed conditional Gaussian distribution, was introduced by **?** was introduced for modeling categorical with bounded support and continuous variables. Their approach entails defining a variable u that is a linear index of all configurations of the categorical variables $\boldsymbol{u}_1 = \{y_j : j \in s_j = 1\}$ which has $d = \prod_{j:s_j=1} d_j$ levels where d_j is the number of levels for $\{j : s_j = 1\}$. Letting $\boldsymbol{u}_2 = \{y_j : s_j = 3\}$ they jointly model $(u, \boldsymbol{u}_2)$ by assuming $\boldsymbol{u}_2|u = h \sim \text{N}_{p_3}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ and $\Pr(u = h) = \pi_h$ where $p_3 = \sum_{j=1}^{p} I(s_j = 3)$. While this approach allows the modeling of nominal and bounded ordinal variables, the definition does not facilitate modeling count variables. Also, computational issues arise in the presence of sparse data or a large number of categorical variables since it will require the estimation of many covariance matrices. An approach that involves the use of latent variables for modeling bounded categorical variables has also been studied and at times referred to as the conditional grouped continuous model (Poon and Lee, 1987). This method involves assuming latent continuous variables $\boldsymbol{y}^*$ for the mixed scale vector $\boldsymbol{y}$ such that $\boldsymbol{y}^* \sim \text{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $F(y_j) = \Phi(\frac{y_j - \mu_j}{\sigma_{jj}})$ for $j \in \{j : s_j = 3\}$, and $\Pr(y_j = h) = \Phi(\frac{q_{hj} - \mu_j}{\sigma_{jj}}) - \Phi(\frac{q_{h-1,j} - \mu_j}{\sigma_{jj}})$ for $j \in \{j : s_j = 1\}$ and $h = 1, \ldots, d_j$ where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable. This idea of assuming an underlying continuous variable for categorical variables is at time referred to as thresholding or rounding in literature. An appealing feature of this model is that it allows multivariate analysis even in the presence of many categorical variables that may include count variables. This methods ability to model many categorical variables comes at the cost of not adequately being able to model nominal variables since the latent continuous variables assume some ordering of the categorical variable levels.

A model aimed at relieving the shortcomings of these two approaches has also been studied that is a hybrid between the conditional Gaussian distribution and the conditional grouped continuous model. The

joint distribution of $\boldsymbol{y}$ in this case is constructed by defining the triplet $(u, \boldsymbol{u}_1', \boldsymbol{u}_2')'$ such that u is a linear index of all configurations of the nominal variables, $\boldsymbol{u}_1$ corresponds to the set of ordinal categorical variables with corresponding continuous latent variables $\boldsymbol{u}_1^*$, and $\boldsymbol{u}_2$ corresponds to the set of continuous variables with corresponding continuous latent variables $\boldsymbol{u}_2^*$. The joint distribution of $\boldsymbol{y}$ is induced by assuming $(\boldsymbol{u}_1^*, \boldsymbol{u}_2^*)'|u = h \sim \mathrm{N}_{p^*}(\boldsymbol{\mu}_h, \Sigma)$ and $\Pr(u = h) = \pi_h$, such that $\boldsymbol{\mu}_h = (\boldsymbol{\mu}_{h1}', \boldsymbol{\mu}_{h2}')'$, $F(u_{j1}|u = h) = \Phi(\frac{u_{j1} - \mu_{hj1}}{\sigma_{jj1}})$, and $\Pr(u_{j2} = l|u = h) = \Phi(\frac{q_{lj} - \mu_{hj2}}{\sigma_{jj}}) - \Phi(\frac{q_{l-1,j} - \mu_{hj2}}{\sigma_{jj}})$. The idea behind this model is to assume a conditional grouped continuous model indexed by configurations of the nominal variables as in the conditional Gaussian distribution. The difference however is that homogeneous covariances are assumed across all configurations. This is a model in difference to previous models that can account for all kinds of scales. However, the computational limitations of the conditional Gaussian distribution are inherited because of the linear indexing of configurations. A potential difficulty that all these models face is whether the multivariate normality may hold although transformations may be applied to the continuous variables. This methodology called the general mixed data model has been very developed and includes likelihood ratio tests, a Malhalanobis distance, and methodology for classification, for further details, see de Leon and Carrièrgre (2007) (de Leon and Carrièrgre, 2007; de Leon and Carriégre, 2005; de Leon, 2007; De Leon et al., 2011).

de Leon and Carrièrgre (2007) have also suggested alternative approaches to circumvent the computational issues that arise with estimating the general mixed data model. They propose using a pairwise likelihood approach to estimating their model where a pseudo-likelihood is specified as a simplification to the conditional component that corresponds to the latent variables of the ordinal variables using only bivariate normal distribution functions. They suggest this results in significant reduction of computing time and easily implemented using S-plus. Generalized estimating equations as a generalization to the pseudo-likelihood approach, naturally, has also been used for analyzing mixed scale data albeit in the context of mixed scale outcomes in correlated data or longitudinal studies. For a thorough and historical account of the development of mixed scale data methods within the context of both joint and outcome analysis which also include copula regression methods, see De Leon and Chough (2013). The GEE introduces an alternative idea not considered up to this point where modeling mixed scale data is done without the use of product conditional-marginal statements but rather by using a joint approach.

2.1.2 Joint Modeling Approaches

The use of conditioning arguments to derive a mixed-scale distribution has been a very useful tool because it facilitated the use of standard distributions with numerous choices in how to define. Each choice involves a degree of subjectivity in defining with varying degrees of computational difficulty, and defines the assumed

underlying data generating mechanism. Alternatively, methods for analyzing mixed scale data without the use of conditioning arguments have been studied. We have briefly mentioned one approach that altogether avoids the specification of a likelihood in generalized estimating equations. The complication with not using conditional arguments is that no standard mixed scale joint distribution exists and as such have had to use to creative tools to achieve analyzing mixed scale distributions. These methods have mainly applied non-parametric and Bayesian approaches.

2.1.2.1 Kernel Density Estimation

In non-parametric statistics, an approach to density estimation is to smooth the unknown true density of a vector. Non-parametric kernel density estimation has been very well developed for multivariate continuous and discrete cases and methods for mixed scale distributions are developing (Nagler, 2017). Briefly, the aim of kernel density estimation is to smooth the unknown density f of the mixed scale vector, $\boldsymbol{y} \sim f$. The smooth form of the unknown density is defined:

$$\tilde{f}(\boldsymbol{y}) = \sum_{l=1}^{L} \alpha_l \mathcal{K}(\boldsymbol{y}_l, \boldsymbol{y}; \boldsymbol{\theta}), \tag{2.1}$$

where $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L\} \in \mathcal{Y}$, $\mathcal{Y} = \otimes_{j=1}^{p} \mathcal{Y}_j$, $\mathcal{Y}_j$ denotes the support of the j^{th} variable in the vector $\boldsymbol{y}$, and for some constants α_l for $l = 1, \ldots, L$ and $\boldsymbol{\theta}$. The kernel $\mathcal{K}$ is defined as a measure of similarity with certain regularity assumptions and often depends on some smoothing parameters. The form of (2.1) in certain cases can be in an alternative basis formulation

$$\tilde{f}(\boldsymbol{y}) = \sum_{h=1}^{H} \omega_h \psi_h(\boldsymbol{y}), \tag{2.2}$$

where $\{\psi_h\}_{h=1}^{H}$ is a set of orthogonal basis functions, each basis function may have a constant parameter, and $\boldsymbol{\omega} = \{\omega_h\}_{h=1}^{H}$ is a set of constant coefficients.

From the few efforts at establishing kernel estimating procedures for the mixed scale setting, Li and Racine (2003) were the first to establish asymptotic normality their proposed smoothing parameters for the class of product kernels. In continuous kernel estimation, a kernel $\mathcal{K}$ can be assigned to measure a multivariate distance $\|\boldsymbol{y}_l - \boldsymbol{y}\|$, however, Li and Racine (2003) use the product kernel formulation. Let $A = \{j : s_j = 3\}$ denote the set of indexes corresponding to continuous variables in $\boldsymbol{y}$ and assume $s_j \in \{1, 3\}$ for all j. The continuous and discrete components of $\boldsymbol{y}$ are separated into $\boldsymbol{y}_A$ and $\boldsymbol{y}_{(-A)}$ with their product kernel density

estimate defined as:

$$\tilde{f}(\boldsymbol{y}) = \frac{1}{L}\sum_{l=1}^{L}\left[\left\{g(\boldsymbol{\theta}^*)\prod_{j\in A} w\big(\tfrac{y_{lj}-y_j}{\theta_j^*}\big)\right\} \times \left\{\prod_{j'\in S\setminus A}\tfrac{1}{d_{j'}-1}(1-\lambda)^{1-I(y_{lj'}-y_{j'})}\lambda^{I(y_{lj'}-y_{j'})}\right\}\right], \quad (2.3)$$

where $S = \{1, \ldots, p\}$, $I(\cdot)$ denotes the indicator function, w is a univariate kernel, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{*'}, \lambda)'$, and d_j denotes the number of levels in $y_j \in \{1, \ldots, d_j\}$ for $j \in S \setminus A$. The product kernel assigns a univariate kernel with respect to each component of the multivariate mixed scale vector and takes the mutual product across all components. They estimate $\boldsymbol{\theta}$ using a cross-validation approach which was shown to be consistent and achieves asymptotic normality. One limitation of this approach is the assumption that each of the continuous components of the mixed vector must come from a density that is four times differentiable.

Efromovich (2011) extended the Li and Racine idea through the use of a tensor product kernel. The tensor product kernel can be loosely defined as:

$$\tilde{f}(\boldsymbol{y}) = \sum_{h_1=1}^{H_1}\cdots\sum_{h_p=0}^{H_p}\theta_{h_1\cdots h_p}\prod_{j=1}^{p}\psi_{h_j,j}(y_j), \quad (2.4)$$

where $\boldsymbol{\Psi} = \{\psi_{0j}, \ldots, \psi_{H_j j}\}_{j=1}^{p}$ is a set of basis functions for each variable j chosen with respect to the measurement scale s_j and $\boldsymbol{\theta} = \{\theta_{h_1\cdots h_p} : h_j = 0, 1, \ldots, H_j\}_{j=1}^{p}$ is a multidimensional fixed coefficient. Through the use of mixed tensor-product basis defined by assigning cosine and discrete cosine basis with respect to the scale of the variable they appealed to Parseval's theorem where any multivariate mixed scale density with a finite norm can be written as a Fourier series, he proposed an oracle estimator (Efromovich, 2011). Appealing features of this cosine tensor product kernel include simultaneous variable selection and dimension reduction and the ability to allow different degrees of smoothness in each continuous variable thereby relaxing the four times differentiable assumption of Li and Racine (2003). Even though this method is an improvement on the relaxing of assumptions, it is limited by an assumption of its own where the continuous variables were assumed bounded on the unit interval. This limitation may require a transformation of the continuous variables perhaps via the inverse logit. Another possible limitation is the potential for computational difficulties since it relies on truncation of the infinite Fourier series. Lastly, the estimation aspects of this method are far from straightforward which may limit routine implementation.

The non-parametric kernel density methods offer nice data-driven alternatives to the parametric counterparts which use product conditional-marginal arguments by completely avoiding the use of conditional arguments. However, the non-parametric methods may have more in common with the product conditional-marginal

since both methods assume their target joint distribution f is of the form, $f(\boldsymbol{y}) = f(\boldsymbol{y}_A|\boldsymbol{y}_{(-A)})p(\boldsymbol{y}_{(-A)})$. It is possible that the true underlying distribution may not be in the class of assumed target distributions. Given that the true underlying data generating mechanism in practice will never be known, this is not to say that this methodology may not be useful, where in fact it may be very useful even in the presence of misspecification.

Recently, a non-parametric analogue to the use of latent continuous variables in analyzing mixed scale data was studied. Following the often used technique of introducing noise to categorical data in non-parametric statistics, the idea was formalized and studied by Nagler (2017). The continuous convolution product kernel density estimator was defined as:

$$\tilde{f}(\boldsymbol{y}) = \frac{g_1(\boldsymbol{\theta}_1)g_2(\boldsymbol{\theta}_2)}{L}\sum_{l=1}^{L}\left[\left\{\prod_{j\in A} w_1\big(\tfrac{y_{lj}-y_j}{\theta_{j1}}\big)\right\}\times\left\{\prod_{j'\in S\setminus A} w_2\big(\tfrac{y_{lj'}-y_{j'}}{\theta_{j'2}}\big)\right\}\right], \tag{2.5}$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ are smoothing parameters, $\boldsymbol{y}_l = \boldsymbol{y}_l^* + \boldsymbol{e}_l$, $\boldsymbol{e}_l = (e_{l1}, \ldots, e_{lp})'$, $e_{lj} \sim \eta$ if $j \in S \setminus A$ and zero otherwise, and $\boldsymbol{y}_l^* \in \mathcal{Y}$. Note this estimator is essentially the same as (2.3) with the exception that the categorical components have noise added to them. The introduction of the noise variable implies the product kernel with w_2 components is now random. It was once thought that the introduction of noise would bias results, however, it was shown that under certain assumptions on the distribution η adding noise to the discrete variables does not negatively impact estimation (Nagler, 2017). Even though this is reassuring and appealing for use in practice, it introduces a complication for nominal variables. Also, it is possible that the continuous convolution technique modifies the target distribution.

2.1.2.2 Bayesian Approaches

We turn our focus to Bayesian methods that have in some cases facilitated mixed scale data analysis. The methods we review use most of the techniques already presented albeit from slightly different viewpoints. Bayesian methods assume parameters from a model are random variables themselves therefore inference is based on the posterior distribution, $f(\theta|y) \propto f(y|\theta)p(\theta)$ as opposed to maximizing a likelihood or estimating equation for a single point estimate. When inference is based on the posterior distribution, one may consider different summaries of it (e.g. median, mean, mode) that may be more appropriate for its shape. The use of Markov Chain Monte Carlo methods allow approximate sampling from the posterior distribution (joint if multivariate parameters or latent variables) which at times make it possible to address problems where maximization may be intractable or computationally difficult. For example, consider the conditional grouped continuous model where latent continuous forms of the discrete variables are integrated out and the resulting likelihood is maximized over the thresholds and covariance parameters. In a Bayesian framework, sampling

from latent variables with a standard distribution is significantly easier to do when coupled with carefully chosen priors for the parameters so as to facilitate a Gibbs sampler.

2.1.2.3 Copulas

The extended rank likelihood approach proposed by Hoff (2007) was developed to jointly analyze mixed scale data. Motivated by the degree of subjectivity involved with degree of subjectivity it takes to define a mixed scale joint distribution, this method was developed with the use of a copula and estimated through a Bayesian semi-parametric approach (Hoff, 2007). A copula is a multivariate distribution where the marginal distribution of each variable is uniform on the unit interval. By Sklar's theorem, any continuous multivariate distribution can be elucidated by a unique copula, $\mathbb{C}$:

$$F(\boldsymbol{y}) = \mathbb{C}(F_1(y_1), \ldots, F_p(y_p)), \tag{2.6}$$

where F_j is the marginal distribution of the j^{th} variable and the marginal distributions of the variables. This is a greatly convenient theorem where one can jointly model a set of continuous variables by assuming a copula structure and marginal distributions. The extended rank likelihood is semi-parametric in the sense that it takes a parametric copula model, the Gaussian copula, and combines it with a non-parametric assumption of unknown marginals. The extended rank likelihood is defined as $\Pr(\boldsymbol{Z} \in R(\boldsymbol{Y})|\boldsymbol{C})$ which is a component of the likelihood induced by the Gaussian copula:

$$\begin{aligned} \Pr(\boldsymbol{Y}|\boldsymbol{C}, F_1, \ldots, F_p) &= \Pr(\boldsymbol{Y}, \boldsymbol{Z} \in R(\boldsymbol{Y})|\boldsymbol{C}, F_1, \ldots, F_p) && (2.7) \\ &= \Pr(\boldsymbol{Z} \in R(\boldsymbol{Y})|\boldsymbol{C}) \Pr(\boldsymbol{Y}|\boldsymbol{Z} \in R(\boldsymbol{Y}), \boldsymbol{C}, F_1, \ldots, F_p), && (2.8) \end{aligned}$$

where $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^n$, $\boldsymbol{z}_i \sim N_p(\boldsymbol{0}, \boldsymbol{C})$, $R(\boldsymbol{Y}) = \{\boldsymbol{Z} : z_{ij} < z_{i'j} \text{ if } y_{ij} < y_{i'j}\}$, $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$, and $\boldsymbol{y}_i$ mixed scale data vector of the i^{th} observation. The set $R(\boldsymbol{Y})$ introduces a partial ordering preserved from the ranks of the observations. Posterior inference is carried out by assigning an inverse-Wishart prior on the correlation matrix $\boldsymbol{C}$ and carried out via Gibbs sampling (Hoff, 2007). The extended rank likelihood has been shown to achieve posterior consistency and was extended to include parsimonious decomposition of the correlation matrix via the Gaussian Copula Factor model (Murray et al., 2013). One of the appealing features of this model is that pairwise and higher order dependence can be carried out from the posterior correlation matrix sample. The derivation of this idea is very different from from the Gaussian threshold idea from the conditional grouped continuous model but ended up with similar modeling elements, as such it unfortunately is not useful for nominal data.

2.1.2.4 Mixture Models

The Bayesian mixture model has been heavily utilized in statistical applications because of its flexibility in modeling non-standard data. In univariate densities ($s_j = 3$) or probability mass functions for count variables ($s_j = 2$), the Bayesian mixture model can be defined as:

$$f_j(y_j) = \int_{\Theta_j} \mathcal{K}_j(y_j|\theta_j) dP_j(\theta_j), \tag{2.9}$$

where $\mathcal{K}_j$ is a pdf or pmf scale appropriate for s_j, P_j is a distribution of the kernel parameters θ_j over the space Θ_j, and Θ_j is countable. By letting d_j denote the size of the space Θ_j then an alternative formulation can be:

$$f_j(y_j) = \sum_{h=1}^{d_j} \omega_{hj} \mathcal{K}_j(y_j|\theta_{hj}), \tag{2.10}$$

where $P_j(\cdot) = \sum_{h=1}^{d_j} \omega_{hj} \delta_{\theta_{hj}}(\cdot)$. With this formulation, we can see that the mixture model can be finite or infinite by setting $d_j < \infty$ or $d_j = \infty$, respectively. For a full Bayesian specification, a prior can be assigned to the kernel weights, where the standard choice for the finite mixture model is the Dirichlet distribution $\boldsymbol{\omega}_j \sim \text{Dir}(\boldsymbol{a}_j)$ for $d_j < \infty$. For the infinite case, $d_j = \infty$, an example may be specifying geometric weights where $\Pr(\theta_{hj}) = \omega_{hj} = q_j(1-q_j)^{h-1}$ and the weight component can be assumed beta distributed $q_j \sim \text{beta}(a_j, b_j)$. Most of the time in practice, the values of the parameter index is unknown which gives rise to the use of random measures for P_j where both the kernel weights and parameter index are assumed random. The use of random measures as priors define Bayesian non-parametric methods, and in the case of the mixture model the parameter index parameters are often chosen to be iid to be conjugate to the kernel. The ubiquitous choice for P_j is the Dirichlet process where through its stick breaking construction, defines the Dirichlet Process Mixture model (DPM) as $P_j \sim DP(\alpha_j P_{0j})$:

$$\begin{aligned} f_j(y_j) &= \textstyle\sum_{h=1}^{\infty} \omega_{hj} \mathcal{K}_j(y_j|\theta^*_{hj}), \\ \boldsymbol{\omega}_j &\sim \text{stick}(\alpha_j), \quad \boldsymbol{\theta}^*_{hj} \overset{iid}{\sim} P_{0j}, \end{aligned} \tag{2.11}$$

where $\boldsymbol{\omega}_j \sim \text{stick}(\alpha_j)$ implies $\omega_{hj} = v_{hj} \prod_{h'<h}(1 - v_{h'j})$ and $\boldsymbol{v}_j \overset{iid}{\sim} \text{beta}(1, \alpha_j)$ (Ferguson, 1973; Sethuraman, 1994). The infinite nature of the DPM traditionally made it difficult for routine implementation but numerous samplers have been devised that have streamlined its use, see for example (Ishwaran and James, 2001; Walker, 2007; Papaspiliopoulos and Roberts, 2008; Kalli et al., 2011; Hastie et al., 2015).

The Bayesian mixture model can be thought of as a special class of the kernel density framework with a positive unit constraint on the basis coefficients and parametric distributions acting as basis functions.

Notice the similarity between (2.2) and (2.10) which in turn makes up the basis for Bayesian density estimation. The connection between the Bayesian mixture model and the constrained positive unit coefficient kernel density framework is a greatly appealing feature because of its intuitive interpretation as the average of the kernels. When a latent indicator variable is introduced, this framework reveals another appealing feature in effectively clustering observations, which is easily accommodated in the Bayesian setting. The posterior distribution of the latent allocation variables reveals that observations are clustered according to the indexed kernel with greatest probability. Thus, the Bayesian mixture model provides a flexible, intuitive to implement and interpret, and rich (i.e. clustering) perspective to kernel density estimation. The mixture model, however, suffers from identifiability issues, which imply that different configurations of the weights and parameter indexes lead to the same model and are manifested in two different forms. The first is widely known as the label switching problem, which implies that any mixture model can be written as $f_j(y_j) = \sum_{h=1}^{d_j} \omega_{hj} \mathcal{K}_j(y_j|\theta_{hj}) = \sum_{l=1}^{d_j} \omega_{\tau_l j} \mathcal{K}_j(y_j|\theta_{\tau_l j})$ where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{d_j})'$ is any permutation of the labels $\{1, 2, \ldots, d_j\}$. The second is a lesser emphasized problem where it is possible to have $f_j(y_j) = \sum_{h=1}^{d_j} \omega_{hj} \mathcal{K}_j(y_j|\theta_{hj}) = \sum_{l=1}^{d'_j} \pi_{lj} \mathcal{K}_j(y_j|\eta_{lj})$ where $\boldsymbol{\theta}_j, \boldsymbol{\eta}_j \in \Theta_j$ and $\boldsymbol{\omega}_j \neq \boldsymbol{\pi}_j$. Numerous investigators have attempted to address these identifiability problems, mostly in the label switching context, for some references see (Richardson and Green, 1997; Stephens, 2000; Jasra et al., 2005; Rodríguez and Walker, 2014; Mena and Walker, 2015). These identifiability issues do not concern density estimation since the focus is on the overall density f_j or functionals of it and not specific components of the weight and kernel index parameter posteriors.

2.1.2.5 Induced Densities

We elaborate on the Bayesian mixture models because it has greatly facilitated the development of mixed-scale distribution modeling. Most of these techniques have leveraged mixtures of multivariate Gaussians coupled with data augmentation in the sense of thresholding/rounding for categorical variables while few others have proposed alternative mixture modeling specifications that can avoid the use of rounding. Data augmentation for rounding to model categorical data can be traced back to Albert and Chib (1993) in the Bayesian setting and easily sampled via truncated variables. When this idea is coupled with the mixture model, the estimation of the threshold cut-offs can be avoided if the mixture as whole is chosen to provide sufficient flexibility as opposed to having to estimate the thresholds when a single Gaussian distribution is specified (Canale and Dunson, 2011; Carmona et al., 2016). Further, in non-parametric Bayes literature, posterior consistency has been demonstrated for multivariate kernels in multivariate density estimation (Wu and Ghosal, 2010). Poisson latent variables have also been leveraged for modeling mixed discrete outcomes which can

serve as a foundation for alternative specifications since they can have appealing properties such as closed form expressions for marginal summaries and improved interpretability (Dunson and Herring, 2005).

The rounding prior for mixed-scale density estimation can generally be described as:

$$f(\boldsymbol{y}) = \int_{C_{\boldsymbol{y}_{(-A)}}} f^*(h_1^{-1}(\boldsymbol{y}_A), \boldsymbol{y}^*_{(-A)})|J_{h_1^{-1}(\boldsymbol{y}_A)}|d\boldsymbol{y}^*_{(-A)}, \tag{2.12}$$

where $\boldsymbol{y} = (\boldsymbol{y}'_A, \boldsymbol{y}'_{(-A)})' = (h_1(\boldsymbol{y}^*_A)', h_2(\boldsymbol{y}^*_{(-A)})')' = h(\boldsymbol{y}^*)$, $\boldsymbol{y}^* \sim f^*$ with f^* a density with respect to Lesbegue measure over $\mathbb{R}^p$, $A = \{j : s_j \geq 3\}$ is the set of indexes corresponding to continuous variables, h_1 is a set of monotone differentiable functions that individually map the real line to the support of y_j for $s_j \in A$, h_2 are thresholding functions for each categorical variable ($s_j \notin A$) that replace elements of $\mathbb{R}$ with non-negative integers according to the support of y_j, $C_{\boldsymbol{y}_{(-A)}} = \{\boldsymbol{y}^*_{(-A)} : y^*_j \in C_{y_j,j}, j \notin A\}$, and $C_{\cdot j} = \{C_{1j}, \ldots, C_{d_j j}\}$ is a mutually exclusive partition of $\mathbb{R}$ defined for each categorical variable $s_j \notin A$ (Canale and Dunson, 2015). The behavior of any specified density of this type will largely be driven by the assumed underlying continuous distribution f^* and the continuous transformation components h_1. Norets and Pelenis (2012) were the first to study theoretical properties of this mixed scale distribution when f^* is chosen to be a finite mixture of multivariate Guassians with identity transformations for continuous variables (h_1), and showed posterior consistency for the density estimator (Norets and Pelenis, 2012). The appealing property of this specification is it can be thought of as building upon some of the previously described methods by adding flexibility with the mixture model. Canale and Dunson (2015) extended the theoretical findings to include a larger class of functions f^* and established sufficient conditions for posterior consistency. They note that the choice of h_1 the continuous transformation functions impact the smoothness of the continuous component of the density and subsequently the convergence rate. Thus, when the identity function is chosen as in (Norets and Pelenis, 2012) then the optimal rate is preserved for the continuous components. The DPM of multivariate Gaussians has been shown to meet the necessary conditions as f^* for mixed-scale density posterior consistency (Canale and Dunson, 2015). Given its intuitive construction and great flexibility, it has been proposed for use in potential spatial statistics applications and even extended to accommodate complex survey designs albeit with a slightly more flexible Poisson-Dirichlet mixture (Molitor et al., 2016; Carmona et al., 2016). It has also been extended to the mixed-scale multivariate regression setting with modeled threshold cut-offs (Papageorgiou and Richardson, 2016). These developments are useful to know when deciding which model would be appropriate for the situation at hand and whether one should use a finite or infinite mixture. As a compromise between the two, one may choose an over-fitted mixture model as it has been shown to adequately zero out extra components and is significantly simpler to implement (Rousseau and Mengersen, 2011). A drawback to this mixed-scale modeling framework is its inability to appropriately account for nominal data.

Recently, this was addressed by incorporating a set of binary indicator variables corresponding to each level for each nominal variable where then a latent continuous variable is assumed for each component, and the indicator variable corresponding to the observed nominal level is set to one when its continuous latent form is the maximum value among the set of continuous latent forms (Kunihama et al., 2016). This methodology was introduced as an extension for modeling mixed-scale variables in longitudinal studies. Related rounding Bayesian methods where the underlying latent continuous distribution is not a mixture model but simply a multivariate Gaussian that account for nominal variables as well, have also been proposed for mixed-scale modeling, (Zhang et al., 2015; Mirkamali and Ganjali, 2016). One of the advantages of rounding with a Gaussian kernel is that the covariance matrix facilitates assessing pairwise associations among variables across the different scales. In the case of sets of underlying variables used for nominal variables, it is unclear how association can be assessed between nominal and variables of other scales. An alternative to the mixture model that can be used for mixed-scale modeling are Pólya trees, however, because the methodology requires an iterative partitioning of the domain it produces non-smooth density estimates which is appropriate for discrete data but not ideal for continuous (Lavine, 1992). The mixture model as we have seen relies on a kernel specification that can produce both smooth or non-smooth density estimates for the continuous and non-continuous components, respectively.

2.1.2.6 Mixture of Product Kernels

Another approach to modeling mixed scale variables within the mixture model setting which avoids the use of latent continuous variables is defined by specifying the product kernel. The product kernel allows one to choose a variable specific kernels thereby providing a technique which can avoid rounding for categorical variables. It can be described as:

$$f(\boldsymbol{y}) = \int_{\Theta} \prod_{j=1}^{p} \mathcal{K}_j(y_j|\theta_j) dP(\boldsymbol{\theta}), \tag{2.13}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, $\Theta = \{\Theta_j\}_{j=1}^{p}$ is a multidimensional parameter space, and $\theta_j \in \Theta_j$ for all j (Dunson and Bhattacharya, 2011). As in the case of the univariate mixture model, an appropriate prior can be assumed for P which could yield the finite or infinite mixture model as:

$$f(\boldsymbol{y}) = \sum_{h=1}^{H} \lambda_h \prod_{j=1}^{p} \mathcal{K}_j(y_j|\theta_{hj}), \tag{2.14}$$

where $\Theta = \{\boldsymbol{\theta}_h\}_{h=1}^{H}$ and $\boldsymbol{\theta}_h = (\theta_{h1}, \ldots, \theta_{hp})'$. With this specification, it is clear to see that rounding can be bypassed since an appropriate kernel can be chosen for all types of scales individually. Further, this specification implies that multivariate dependence is induced as an average of independence models, and

allows the ability to identify sub-classes where when a latent indicator is introduced implies local independence within classes. Some have used this as a basis for the analysis of heterogeneous data recently albeit not within the Bayesian setting (Amiri et al., 2017). The product kernel formulation has been used for characterizing the population distribution of fibers between brain regions within and across individuals through the use of the Nested Dirichlet Process (Zhang et al., 2016). The product kernel mixture model framework provides notable simplifications to mixed-scale modeling over the rounding prior with its ease in incorporation of categorical variables, which when coupled with a random distribution on the kernel parameter indexes maintains great flexibility for modeling unconventional distributions. However, it is limited in assessing the strength of association among the variables as opposed to the ease in assessing dependence in previously described Bayesian methods albeit none facilitate cross nominal variable associations.

The product kernel infinite mixture model has been successfully used to model multivariate categorical data where strengths of associations among nominal and ordinal variables are easily assessed. From a generalization of singular value decomposition, Dunson and Xing (2009) showed that any probability tensor has a parallel factors (PARAFAC) decomposition as:

$$\boldsymbol{\pi} = \sum_{h=1}^{k} \lambda_h \bigotimes_{j=1}^{p} \boldsymbol{\psi}_{hj}, \tag{2.15}$$

where $\boldsymbol{\pi} = \{\pi_{h_1 \cdots h_p},\ h_j = 1, \ldots, d_j,\ j = 1, \ldots, p\}$, $\boldsymbol{\lambda}$ is a $k \times 1$ probability vector, $\boldsymbol{\psi}_{hj}$ is a $d_j \times 1$ probability vector for $h = 1, \ldots, k$, and $\otimes$ denotes the outer product. This provided a basis for using mixture models for analyzing multivariate unordered categorical data where they proposed the DPM of product multinomials for a Bayesian non-parametric implementation which effectively automates the selection of the rank k. Within this modeling framework, a model-based Cramer's V statistic was proposed to quantify pairwise associations and test pairwise independence which can be used for gauging multivariate dependence. It has been successfully used to facilitate associations among nominal variables and across ordinal and nominal scales. The immediate drawback is it is not directly useful for mixed scale data, however, it has provided a framework for extensions of the mixture model in mixed-scale applications.

Murray and Reiter (2016) married multiple mixed-scale modeling ideas to jointly model continuous and bounded categorical variables. They developed a mixture of mixtures called the hierarchically coupled mixture model with local dependence (HCMM-LD) by combining mixture and general location-scale modeling ideas. The HCMM-LD can be defined as:

$$f(\boldsymbol{y}) = \sum_{h=1}^{H} \lambda_h f_h(\boldsymbol{y}_A | \boldsymbol{y}_{(-A)}) \pi_{h, \boldsymbol{y}_{(-A)}}, \tag{2.16}$$

where the marginal of the categorical and conditional of the continuous components are modeled as indexed PARAFAC tensor decomposition (2.15) and DPM of multivariate Gaussians, respectively, as:

$$\begin{aligned} f_h(\boldsymbol{y}_A|\boldsymbol{y}_{(-A)}) &= \textstyle\sum_{r=1}^{R} \omega_{hr,1} \mathrm{N}_{p_c}(\boldsymbol{y}_A|D(\boldsymbol{y}_{(-A)})B_r, \Sigma_r), \\ \pi_{h,\boldsymbol{y}_{(-A)}} &= \textstyle\sum_{s=1}^{S} \omega_{hs,2} \prod_{j\notin A} \psi_{sy_j,j}, \end{aligned} \tag{2.17}$$

$p_c = \sum_{j=1}^{p} I(s_j \geq 3)$, and $D(\boldsymbol{y}_{(-A)})$ denotes a design matrix on the set of categorical variables that can include various forms of dependence (Murray and Reiter, 2016). The conditional distribution specification on the continuous components differs from the general location model in that it includes a set of regression variable specific regression coefficients as opposed to the linear indexing of all combinations of the categorical variable levels. It is a technique commonly implemented to overcome the sparsity induced by $\prod_{j\notin A} d_j$ dimensional split that the data would be required in the general location model. This formulation provides a sound basis that could facilitate the assessment of association among variables truly across all scales since it incorporates PARAFAC decomposition (2.15) components.

The mixture model provides infinite ways to formulate a mixed-scale distribution with great flexibility but the unidirectional indexing of the parameter space may require a greater number of components. A multi-directional exploration of the parameter space may be more efficient in capturing dependence with a smaller number of clusters. To illustrate how the tensor mixture model can provide a more efficient way of exploring the indexed kernel parameter space, consider the multivariate mixture of product kernels model defined in (2.13). The parameter space Θ and the corresponding probability P define the type of induced mixture model. Often for mathematical convenience, the parameter space is specified as a product space of each component $\Theta = \bigotimes_{j=1}^{p} \Theta_j$. When Θ is specified on a linear index such that $\boldsymbol{\theta}_h = (\theta_{h1}, \ldots, \theta_{hp})' \in \Theta$ for all h then P is univariate, however, when Θ is specified on its multivariate index such that $\boldsymbol{\theta}_{h_1\cdots h_p} = (\theta_{h_1 1}, \ldots, \theta_{h_p p})' \in \Theta$ for all h_j, j then P is a tensor. The mixture model subsequently, as opposed to the tensor mixture model, may require the selection of more clusters in order to model the same structure when elements of the parameters space are assumed random (i.e. Bayesian non-parametric modeling). Banerjee et al. (2013) extended the Dunson and Xing (2009) tensor factorization (2.15) to the infinite tensor factorization (ITF) where the arms $\boldsymbol{\psi}_{hj}$ are assumed to be infinite probability vectors and assumed to have stick-breaking priors $\boldsymbol{\psi}_{hj} \sim \text{stick}(\boldsymbol{a}_j)$, and simultaneously defined the infinite tensor mixture (ITM) model via product kernel:

$$f(\boldsymbol{y}) = \sum_{h_1=1}^{\infty} \cdots \sum_{h_p=1}^{\infty} \pi_{h_1\cdots h_p} \prod_{j=1}^{p} \mathcal{K}_j(y_j|\theta_{h_j j}), \tag{2.18}$$

which implies $P(\cdot) = \sum_{h_1=1}^{\infty} \cdots \sum_{h_p=1}^{p} \pi_{h_1 \cdots h_p} \delta_{\boldsymbol{\theta}_{h_1 \cdots h_p}}(\cdot)$ and $\boldsymbol{\pi}$ is assumed ITF. Atop of ease in model specification and enhanced flexibility in exploration of the index parameter space, the tensor mixture model also facilitates assessing the strength of association between disparate scales. Latent cluster indicator variables $C_1, \ldots, C_p$ are inherently present in ITM such that $\Pr(C_1 = h_1, \ldots, C_p = h_p) = \pi_{h_1 \cdots h_p}$ where the dependence between the variables is induced via the dependence in the cluster indicators, therefore, variable dependence can assessed without regard to the disparate scales. Further, this also alleviates complications with nominal variables. Banerjee et al. (2013) proposed estimation of the ITM via Bayesian non-parametric techniques and Kullback Leibler divergence quantities for assessing pairwise and higher dependencies. The main drawbacks of this methodology are sampling requires a sophisticated blocked MCMC with partially collapsed steps and incorporation of slice and retrospective sampling ideas in order to avoid approximations, and computational complications may arise under certain scenarios with undue influence of the joint structure hindering flexibility in the marginal distributions.

It is clear that investigators have had to develop ingenious techniques for addressing the complications that arise when analyzing mixed-scale data and it is unlikely for there to be one one-size-fits-all technique since the true data generating mechanism will seldom be known. We seek to study an exploratory mixed-scale data analysis method that facilitates the assessment of dependence among all variables. Motivated by the ITM, we seek to provide simplifications to this methodology where latent cluster indicators are also levied to model dependence and alleviate some of the computational complications via a separation of marginal and joint structures (i.e. modularization). This results in the study of a copula-like approach that assesses joint dependence from unknown marginal distributions in chapter 3 .

2.2 Variable Selection for Hierarchical Interactions

In environmental epidemiology, it is of interest to model the exposure-response surface associations between multiple exposures and characteristics and a health outcome of interest for quantifying and describing the relationship. Inference often entails simultaneous interest in estimating regression coefficients, variable selection, and grouping exposures based on magnitude of associations. Maximum likelihood estimation is not reliable for quantifying the relationship and by itself unable to simultaneously accomplish variable selection and grouping. Alternatively, certain shrinkage methods have the capability to accomplish all these. We focus on methods for linear models as opposed to non-linear models.

Within the context of linear models, it is also of interest to determine whether there is evidence of non-homogeneous associations via interactions. Interactions add another dimension of difficulty as some statisticians argue there is a need for incorporating hierarchy in the selection because violations may result in

non-sensible interpretation and reduced power (Bien et al., 2013; Cox, 1984; McCullagh and Nelder, 2019). Incorporating hierarchy in interaction selection is hereby referred to as hierarchical interactions selection. For illustration consider the pairwise interactions model,

$$g(\mu_i) = \mathbf{z}_i'\boldsymbol{\kappa} + \boldsymbol{x}_i'\boldsymbol{\beta}_1 + \sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p} \beta_{j_1 j_2} x_{ij_1} x_{ij_2}, \tag{2.19}$$

where $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)'$, $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \ldots, \kappa_{p_0})'$, y_i is a health outcome of interest, $E(y_i) = \mu_i$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ is a set of predictors, and $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{ip_0})'$ is a set of covariates. We make a distinction between predictors and covariates where predictors take the role of exposures being assessed and subject to variable selection. Covariates on the other hand are a set of variables, often confounding variables, for which investigators want to adjust for but not subject to variable selection. We hereafter use predictors and exposures, and covariates and confounding variables (i.e. confounders) interchangeably. Hierarchical interactions impose restrictions on the values higher order terms can take and have different strengths. Within the context of 2.19, weak hierarchy implies $\beta_{j_1 j_2} \neq 0$ only if $\beta_{j_1} \neq 0$ or $\beta_{j_2} \neq 0$. Strong hierarchy, on the other hand, implies $\beta_{j_1 j_2} \neq 0$ only if $\beta_{j_1} \neq 0$ and $\beta_{j_2} \neq 0$. Strong and weak hierarchy also hold within the n-way interaction model where the degree of the interaction model (i.e. n) is less than or equal to the number of predictors (p). Shrinkage methods are highly appealing in this setting because these not only make estimation possible but some methods offer simultaneous variable selection.

Constrained optimization (i.e. penalized regression) methods offer shrinkage but not all have the capability to simultaneously perform variables selection, and only recent methods have begun to incorporate ordered selection for hierarchical interactions. Ridge regression as one of the first penalized regression methods offered parameter stabilization but the L_2 constraint on the regression parameters is not able to shrink estimates to zero thereby incapable of performing simultaneous variable selection (Hoerl and Kennard, 1970). The Least Absolute Shrinkage and Selection Operator (LASSO) is a widely used shrinkage estimator that uses the L_1 norm constraint which allows parameter estimates to be shrunk to exactly zero (Tibshirani, 1996). LASSO is a favorite and widely popular shrinkage technique because of its theoretical properties, fast and scalable algorithms, and implementation in software (Lockhart et al., 2014). There have been numerous extension of LASSO that can incorporate hierarchical interactions such as the hiernet, group-lasso interaction network (GLINTERNET), and generalization framework for modeling interactions with a convex penalty (FAMILY) (Bien et al., 2013; Haris et al., 2016; Lim and Hastie, 2015). LASSO and GLINTERNET have been carefully studied in comparison to other variable selection methods within the context of correlated predictors for interaction selection, see Barrera-Gómez et al. (2017); Sun et al. (2013). For all of the great theoretical and

computational advances of LASSO and variants, there are still advances being made in quantifying uncertainty in selection and estimation (Lockhart et al., 2014).

Bayesian methods offer an alternative way of achieving shrinkage and in some instances simplify characterizing uncertainty with direct connections to some penalized regression methods. In Bayesian analyses, all model parameters are assumed to have an underlying distribution called a prior, where then inference is based on the posterior distribution. The optimization equation of the LASSO,

$$L(\lambda, \boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.20}$$

shows there is an equivalence to assuming an i.i.d. double exponential prior with zero mean on the regression coefficients. Assuming this prior defines the Bayesian LASSO in the linear regression setting, where then inference is based on the posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ with samples obtained via Markov Chain Monte Carlo algorithms (Li and Lin, 2010; Park and Casella, 2008). The variability can be characterized by the posterior samples and testing performed by computing posterior inclusion probabilities as the proportion of $\beta_j \notin (-\epsilon, \epsilon)$ for some small $\epsilon > 0$. The Bayesian analogue of the group LASSO has been developed, however, constructing a prior that preserves hierarchy is not trivial (Xu et al., 2015). Priors can be carefully constructed to place greater sparsity on higher order interaction terms, for example see Antonelli et al. (2017); Herring (2010), but few if any exist that uphold strong hierarchy.

An alternative to LASSO type priors for shrinkage and variable selection is to use spike-slab priors on regression coefficients. These priors offer simplification in testing for inclusion over continuous sparsity inducing priors (e.g. LASSO type priors) because these allow parameters to assume the value zero. Thus posterior inclusion probabilities can be easily constructed as the proportion not assuming zero, as opposed to using interval type testing procedures. Spike-slab priors are defined as a two-component mixture distribution between degenerate at zero (i.e. spike) and continuous distributions (e.g. Normal, double exponential, etc), and generally require including more parameters (George and McCulloch, 1993; Geweke, 1996). Shrinkage can be incorporated in two ways, through the prior: 1) probability of exclusion (i.e. probability of point mass at zero) and 2) hyper-parameters on the slab distribution. These features allow investigators greater control and facilitate the characterization of uncertainty in both estimation and variable selection. Spike-slab priors however do not offer the capability to formally group or cluster predictors based on magnitude of associations.

Few Bayesian methods have been proposed which can simultaneously group predictors. The multiple shrinkage and effect fusion priors are priors on regression coefficients that can be used to group predictors but have not been explicitly used for assessing non-homogeneous associations MacLehose and Dunson (2010); Malsiner-Walli et al. (2018). Herring (2010) proposed a semi-parametric pairwise interaction model with a

modified spike-slab prior in the form:

$$\begin{aligned}
\beta_j &\sim \pi_{01}\delta_0(\cdot) + (1-\pi_{01})G_1(\cdot), \text{ for } j = 1, \ldots, p \\
\beta_{j_1 j_2} &\sim \pi_{02}\delta_0(\cdot) + (1-\pi_{02})G_2(\cdot), \text{ for } 1 \leq j_1 < j_2 \leq p \\
\pi_{0l} &\sim \text{Beta}(a_{\pi l}, b_{\pi l}) \\
G_l &\sim \text{DP}(\alpha_l, G_{0l}) \\
G_{0l} &\equiv \text{N}(\mu_{Bl}, \sigma^2_{Bl}), \text{ for } l = 1, 2,
\end{aligned} \tag{2.21}$$

$\delta_0(\cdot)$ is the degenerate distribution at zero, and DP denotes the Dirichlet Process distribution. This specification offers numerous appealing properties. First, it preserves variable selection by including degenerate distributions. Second, it separates the priors assumed on the main effect and interaction coefficients resulting in a grouping of coefficients, where a group specific prior probability of exclusion can be specified via π_{0l}. Lastly, the non-parametric Bayes DP prior simultaneously provides a mechanism for within group clustering of predictors based on the magnitude of the coefficients and shrinkage. This can be illustrated by observing the DP prior within its stick-breaking representation,

$$G_l(\cdot|\boldsymbol{B}_l, \boldsymbol{\lambda}_l) = \sum_{h=1}^{\infty} \lambda_{hl}\delta_{B_{hl}}(\cdot), \tag{2.22}$$

where $\boldsymbol{\lambda}_l \sim \text{stick}(\alpha_l)$, $\boldsymbol{B}_l = \{B_{hl}\}_{h=1}^{\infty}$, and $B_{hl} \sim G_{0l}$ (Ferguson, 1973; Sethuraman, 1994). The stick breaking prior implies $\lambda_{hl} = \nu_{hl}\prod_{h'<h}(1-\nu_{h'l})$ with $\nu_{hl} \sim \text{Beta}(1, \alpha_l)$ (Sethuraman, 1994). The discreteness of the DP prior shows regression coefficients are clustered by taking the same group specific block atom value, B_{hl}, where then the atom values are shrunk toward the block specific base measure mean, μ_{Bl}. Further, the group specific precision parameter, α_l, controls the degree of regression coefficient magnitude clustering where smaller values favor clustering. The grouping and prior specification of 2.21 offers great flexibility in incorporating prior shrinkage on the regression effects and inclusion probability by effect type, which can be a way to control for hierarchy by specifying priors to favor inclusion of lower order terms. This is a sensible way to achieve hierarchy in pairwise interaction models but can become difficult to control in n-way interaction models and it does not completely guarantee hierarchy.

Bayesian methods offer the capability to simplify and incorporate shrinkage in numerous ways for assessing non-homogeneous associations between moderate to highly correlated predictors and a health outcome of interest. However, there is currently a void in Bayesian methods that can accommodate hierarchical interactions as well as simultaneous provide estimation, variable selection, grouping of predictors, and

characterization of variability. In Chapter 4, we propose and study a Bayesian semi-parametric model with hierarchical interaction selection.

2.3 Compositional Data Analysis

Compositional data analysis is defined as a multivariate random vector of size D, $\boldsymbol{y} = (y_1, \ldots, y_D)'$, with the constraint $\sum_{j=1}^{D} y_j = \kappa$. The space defined by such vectors then defines the simplex of D parts as:

$$\mathbb{S}^D = \{\boldsymbol{y} = (y_1, \ldots, y_D)', \sum_{j=1}^{D} y_j = \kappa, y_j > 0, j = 1, \ldots, D\}, \tag{2.23}$$

where κ is some positive constant. When $\kappa = 1$ then $\mathbb{S}^D$ is the unit simplex where $y_j \in \mathbb{R}_{[0,1]}$ for all j. Compositional data or compositional multivariate proportions lie in the unit simplex. The unit simplex is a specific type of multivariate scale that is bounded on the real line.

The mixture model has been a useful tool for distributional clustering in heterogeneous populations and has been adapted for compositional data in numerous ways. The mixture of Dirichlets modeling framework is a natural approach for compositional data analysis because the Dirichlet distribution is defined on the unit simplex. However, it has been seldom used in application primarily because estimation of the Dirichlet distribution parameters is not trivial atop selecting the number of components in the model (Bouguila et al., 2004; Calif et al., 2011a; Giordan and Wehrens, 2015). A class of Dirichlet Process mixture of Dirichlets models has been recently developed where estimation of the number of components is circumvented by using an infinite mixture (Barrientos et al., 2015). This Bayesian non-parametric framework has theoretical backing, can provide density estimation, and is capable of model-based clustering. Another mixture model on the simplex was defined using the principle of staying in the plane (Comas-Cufí et al., 2016). The principle of staying the plane, briefly, consists of applying a one-to-one projection from the simplex space, $\mathbb{S}^D$ to the real space, $\mathbb{R}^{D-1}$, using log ratio transformations, where then standard statistical methods are applied (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Comas-Cufí et al. (2016) defined a mixture model on the simplex using a multivariate normal kernel on the isometric log ratio coordinates. This mixture model was shown to be more flexible in density estimation over a finite mixture of Dirichlet distributions but requires mixture component selection. Current approaches for analyzing compositional data with heterogeneous populations are well suited resources for dealing with the non-conventional scale of the simplex and distributional clustering, however, because these rely on standard distributions, these are problematic at boundary elements (i.e. zero values in composition components).

Essential zeros, defined as a true absence of a component within a composition, have been traditionally problematic. For instance, the compositional mean (i.e. measure of central tendency), defined as the scaled geometric mean across all components is incapable of providing a true measure of central tendency because zeros present within a component will zero out the corresponding component in the overall estimate. In terms of modeling, methods that rely on standard distributions such as the Dirichlet or log ratio methods are also problematic because these can be zero or undefined, respectively. Thus typically, compositional data containing essential zeros is analyzed via stratification by zero pattern or imputing zeros with small positive quantities as is done for rounded zeros (Martın-Fernandez et al., 2011; Kaul et al., 2017b). Stratification complicates the interpretation of analyses without a way to jointly draw conclusions statistically, and imputation is not tenable for certain applications such as physical activity time budget proportions. Recently, a method for accounting for essential zeros via a conditional multivariate Gaussian distribution which characterizes essential zeros and estimates means and covariance matrices has been developed (Kaul et al., 2017a). This methodology is appealing because it has theoretical backing and is applicable on high dimensional data, however, this methodology while compositional may not be directly applicable for vectors in the unit simplex.

The literature on mixture models for compositional data is limited and current methods cannot handle essential zeros despite their great applicability. In environmental epidemiology, for example, it may be of interest to profile the composition of household dust but it may be possible for some households to be devoid of certain particles. Additionally, in physical activity epidemiology, it is often of interest to profile time budgets of sleep, sedentary behavior, and physical activity using accelerometer devices where many adults may not achieve moderate to vigorous levels of physical activity in a 24 hour time period. Thus, in Chapter 5, as a case study of a particular scale, we develop a joint mixture model for compositional data with essential zeros.

CHAPTER 3: JOINT MODELING OF MIXED SCALE VARIABLES USING MODULARIZED TENSOR FACTORIZATIONS

3.1 Introduction

Recent technological developments in the ability to collect and store vast amounts of information on subject units of interest have resulted in a need for development of statistical procedures that can accommodate this scenario which can span multiple variable scales. In assessing how multiple exposures, risk factors, and subject characteristics interrelate and affect human health, it may be overwhelming or infeasible to make interpretations of results in moderate to high dimensions and may present computational difficulties in certain statistical models, especially when some variables are highly correlated. Thus, simple exploratory data analysis tools that can aid in understanding complex associations among variables of different scales are necessary to aid model development and somehow summarize these. Dimension reduction tools have been successful for summarizing multivariate subject characteristics by a smaller set of functionals (e.g. principal components, clusters, latent classes, profiles, etc.) which make interpretations a bit more manageable. For example, these methods have been in particular useful for assessing the intercorrelated nature of short-term air pollution exposure effect on health (Davalos et al., 2017). Motivated by distributional approaches to clustering and generating summary multivariate mixed-scale profiles, in this paper, we study modularized tensor factorizations for assessing joint structures of multivariate mixed-scale variables.

Modeling multivariate mixed-scale distributions requires sophisticated statistical frameworks since there is not a standard parametric distribution that do so, and only a recent few facilitate clustering of similar variables (i.e. mixture modeling). Because there is no standard mixed-scale distribution, most methods have leveraged continuous and discrete distributions using different assumptions to model mixed-scale data. Some methods define a mixed scale distribution as a product of conditional and marginal distributions on the continuous and categorical block of variables and use standard multivariate distributions (e.g. multivariate Gaussian for continuous variables) on each respective homogeneous scale variable blocks (Edwards, 2012; Lauritzen and Wermuth, 1989). Other approaches define a joint distribution as a functional of a multivariate distribution of underlying continuous variables where the functional is defined by rounding (i.e. integrating out) on partitions of $\mathbb{R}$ associated with categorical variables specified by thresholds (Canale and Dunson, 2015; Kunihama et al., 2016; Poon and Lee, 1987; Zhang et al., 2015). Hybrid approaches combining both of these ideas have also

been proposed where ordinal categorical variables are assumed to have latent continuous variables and the latent multivariate continuous distribution is defined conditionally on the set of nominal variables (de Leon and Carrièrgre, 2007; Murray and Reiter, 2016). The Gaussian copula has also been leveraged to model mixed-scale distributions and ends up being related the rounding techniques (Hoff, 2007; Murray et al., 2013). Kernel density estimation has also been developed for flexible modeling of mixed scale densities (Efromovich, 2011; Li and Racine, 2003; Nagler, 2017). For a historical and detailed account of mixed-scale distribution modeling with the exception of recent developments, see (De Leon and Chough, 2013). Few of the mentioned modeling frameworks allow for an immediate leveraging of modeling components which can be used for clustering of subjects. Further, most of these methods, especially those with rounding components, are not applicable for nominal variables and those that are do not facilitate quick measures of association for nominal variables.

The product kernel mixture model is a very versatile statistical framework that has been exploited to flexibly model many different types of data including unordered categorical and mixed-scale, while providing a natural basis for multivariate distributional clustering. In mixed-scale modeling, the product kernel allows for the specification of variable specific kernels which avoids the need for a latent continuous variables and relaxes assumptions for nominal categorical variables. Dunson and Bhattacharya (2011) proposed the discrete mixtures of product kernels (DMPK) model which provided a basis for the product kernel in mixed-scale modeling but inference on the joint dependence structure of vector does not perform well (Banerjee et al., 2013). The infinite tensor mixture model (ITM) was recently proposed and defined as a tensor mixture model with product kernels which was shown to have comparable and improved prediction accuracy in certain scenarios over the DMPK (Banerjee et al., 2013). The ITM is the first of a kind framework that facilitates scale-free assessment of dependence. Another notable feature is that the tensor clustering mechanism allows marginal as well as joint clustering which can be used to summarize multivariate profiles. However, with all the highly appealing features, the ITM framework requires a sophisticated sampling algorithm and its tensor (i.e. simultaneous marginal and joint) clustering may at times hinder performance where feedback from the joint structure limits the flexibility of modeling the marginal structures.

The proposed method, simplifies the ITM by separating the modeling into two components which consequently provides a framework for addressing multiple interesting statistical elements while preserving its notable features. The ITM assumes conditional complete mutual independence among all variables where dependence among the variables is inherited through the dependence of the cluster allocations, which is exploited to construct scale-invariant measures for assessing pairwise or higher order dependence via Kullback-Leibler style measures. This provides the basis and logic behind our two component approach to mixed-scale modeling as **MO**dularized **TE**nsor **F**actorizations. In the first (marginal structures) module, we separately assume each variable has an unknown marginal distribution which is modeled with a Bayesian

non-parametric mixture model. In the second (interaction) module, the marginal cluster allocations are then treated as an unordered multivariate contingency table which is modeled via the parallel factors (PARAFAC) tensor factorization.The modularization and models chosen for each component immediately simplify the ITM sampling algorithm and the separate marginal models can be implemented in parallel for computational scalability and speed. Typical modularized approaches take a summary measure from the first module which is then fed into the second module. In epidemiologic studies, non-interwoven studies are typically used, for example, using latent factor/class analysis to estimate dietary patterns and then include indicators of pattern membership in a second stage regression model (Sotres-Alvarez et al., 2013). By construction, MOTEF allows a uni-directional sharing of information from module one to module two (i.e. interweave) which directly addresses the second limitation of ITM. PARAFAC tensor factorizations have very useful tools for assessing and quantifying pairwise and higher order dependence, but are not directly useful to mixed-scale data (Dunson and Xing, 2009). Thus, MOTEF provides an adapted framework that allows those well developed methods to be used in the mixed-scale setting, thereby preserving the appealing features of ITM.

MOTEF is a modeling framework conceptually similar to copulas but different from extended rank likelihood copula methods for mixed-scale variables, where MOTEF assumes the unifying model is a tensor as opposed to a multivariate Gaussian and the marginal structures are modeled (Hoff, 2007; Murray et al., 2013). This leads us to first statistical element to study: the effectiveness of assessing joint dependence by marginal structures. Recently, this was studied for exploring pairwise dependence and MOTEF can thought of as an extension to p-variate vectors as opposed to bi-variate vectors (Filippi et al., 2016). Secondly, MOTEF allows investigators to explore the complete mutual conditional independence data generating mechanism much similar to the DMPK, which can be useful as an exploratory method to summarize moderate to high dimensional data. This is a data generating mechanism that has seldom been assumed for mixed-scale data since most methods have used some variation of rounding, but is a basis for certain ad-hoc mixed-scale methods which categorize non-discrete variables and then assess all as multivariate discrete data (De Leon and Chough, 2013). Lastly, MOTEF provides a unified modeling framework which accounts for uncertainty in categorizing non-categorical variables by the interweaving into the interaction module. Interweaving has been used to account for uncertainty in air pollution profile clusters in assessing the association between air pollution and term low birth weight (Molitor et al., 2016). In a similar way, MOTEF accounts for variation in marginal subject cluster membership in jointly modeling mixed-scale data.

Even though the idea behind MOTEF is simple, it is a unified method which has potential to be routinely used as an exploratory data analysis tool with many attractive features. The paper proceeds as follows: Section 3.2 presents background and formulation of MOTEF, section 3.3 presents results of simulation experiments aimed at assessing joint dependence by marginal structures, section 3.4 presents an application of the proposed

method in describing associations among risk factors and confounding variables on the risk of selected birth defects from a large population-based epidemiologic study of birth defects, the National Birth Defects Prevention Study, and concludes with a discussion in section 3.5.

3.2 Modularized tensor factorizations

3.2.1 Background

Suppose we observe a multivariate mixed scale data vector, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})' \sim f$, which comes from an unknown joint density f and let $s_j \in \{1, \ldots, S\}$ index the measurement scale for variable j, for $j = 1, \ldots, p$. One of the greatest complications to modeling f is the complexity involved with specifying a multivariate distribution that can account for all the scales but a wealth of options exist for univariate distributions with parametric and non-parametric approaches. The mixture model is a sound choice for modeling an unknown marginal distribution of y_{ij},

$$f_j(y_{ij}) = \int_{\Theta_j} \mathcal{K}_j(y_{ij}|\theta_j) dP_j(\theta_j), \tag{3.24}$$

where $\mathcal{K}_j(y_{ij}|\theta_j)$ is a kernel appropriate for s_j indexed by $\theta_j \in \Theta_j$ and P_j is a probability distribution over the space Θ_j, that can "categorize" by effectively clustering observations by the indexed kernel. The Dirichlet process random measure has seen widespread use as a prior for the mixture model and the discovery of its stick breaking representation has facilitated hierarchical modeling (Ferguson, 1973; Sethuraman, 1994). A hierarchical specification is elucidated as,

$$y_{ij} \sim \mathcal{K}_j(y_{ij}|\theta^*_{x_{ij},j}), \quad \theta^*_{hj} \sim P_{0j}, \quad \Pr(x_{ij} = h) = \omega_{hj}, \tag{3.25}$$

where x_{ij} is a latent allocation variable, $\omega_{hj} = \omega^*_{hj} \prod_{h'<h}(1 - \omega^*_{h'j})$, and $\omega^*_{hj} \sim \text{beta}(1, \alpha_j)$. A short hand representation of the prior specification on the weights is $\boldsymbol{\omega}_j \sim \text{stick}(\alpha_j)$. In density estimation of continuous variables, the mixture model with DP priors and Gaussian kernel (GDPM), is a staple method among non-parametric techniques (Escobar and West, 1995). Additionally, in modeling count data, the GDPM has been successfully used with great performance as a distributional assumption on continuous latent variables where the discrete count distribution is obtained via rounding (Canale and Dunson, 2011). Given that we are motivated to bring a formalized approach to the first ad-hoc version approach to multivariate dependence, categorical variables may be left as observed (assume Dirac measure) or be assigned a multinomial kernel if we would like greater compression. This is a simple convenient framework that we adapt for the univariate

model assessment for variables with continuous and count scales. One of the highly appealing properties of the mixture modeling is its ability to effectively model data that cannot be necessarily assumed distributed according its kernel, for example the GDPM has been very successful in the density estimation of non-Gaussian distributed data.

Mixture models have also been proposed for modeling multivariate mixed scale data. We highlight two methods with distinct approaches. Canale and Dunson (2015) have recently developed theoretical properties of Bayesian non-parametric modeling of mixed scale data by assuming a multivariate Gaussian DPM on latent continuous variables where the joint distribution is induced by thresholding (i.e. rounding) the categorical variables. The only drawback to this method, is its shortcoming where the latent continuous variable assumption may not be justifiable for nominal variables thus making it difficult to interpret the degree of association between nominal and continuous variables. Dunson and Bhattacharya (2011) proposed a different mixture modeling approach by assuming conditional independence on product kernels where multivariate dependence is induced by integrating out the allocation variable, that is,

$$f(y_{i1}, \ldots, y_{ip}) = \int_{\Theta} \prod_{j=1}^{p} \mathcal{K}_j(y_{ij}|\theta_j) dP(\theta), \tag{3.26}$$

$\theta = (\theta_1, \ldots, \theta_p)'$, $P(\cdot) = \otimes_{j=1}^{p} P_j(\cdot)$, and $\Theta = \otimes_{j=1}^{p} \Theta_j$ (Dunson and Bhattacharya, 2011). This approach seemingly circumvents the nominal variable shortcoming of continuous latent variable approaches. Even though these NPB approaches offer great flexibility, in the presence of moderate to high-dimensional variables the unidirectional approach may under certain scenarios be restrictive and run into computational difficulties, especially when using a high dimensional covariance or large number of parameters.

The non-parametric Bayes product multinomial mixture mode approach proposed by Dunson and Xing (2009) has been highly successful in modeling multivariate categorical data and contingency tables applications. Under the same framework, this mixture modeling approach has performed well because the joint probability mass function is characterized as a low rank probability tensor through singular value decomposition on positive value matrices, termed non-negative PARAFAC factorization (Dunson and Xing, 2009). The multivariate p.m.f is modeled as

$$\text{pr}(x_{i1} = c_1, \ldots, x_{ip} = c_p) = \pi_{c_1 \cdots c_p} = \sum_{h=1}^{\infty} \lambda_h \prod_{j=1}^{p} \psi_{hc_j,j}, \tag{3.27}$$

where $\boldsymbol{\lambda} \sim \text{stick}(\alpha)$, $\boldsymbol{\psi}_{hj} \sim \text{Dirichlet}(\boldsymbol{a}_j)$ for $h = 1, \ldots, \infty$ and $j = 1, \ldots, p$, and $\boldsymbol{a}_j = (a_{1j}, \ldots, a_{d_j,j})'$. The brilliance behind this method is its simplicity and parsimony coupled with the DP assumption where from a set of p categorical variables the number of parameters needed to estimated the multivariate multinomial

reduces from $\prod_{j=1}^{p} d_j - 1$ to $\sum_{j=1}^{p} d_j - p$, and the DP assumption to the weights which avoids the specification of the tensor rank and who's precision parameter can be specified to favor a smaller rank values or be assigned a prior for a more data driven approach. This method however is not directly useful for mixed scale modeling but has provided a basis for other methods.

The infinite mixture of product kernels framework for mixed-scale distribution modeling seems to be powerful to tool that is simple, easy to understand and highly flexible where its only restriction is its unidirectional search along the product space of the kernel parameters. By design, the product kernel mixture model allows only a linear search along the product space of the kernel parameters, that is, the indexing of the kernel parameters is across all parameters $\boldsymbol{\theta}_h = (\theta_{h1}, \ldots, \theta_{hp})'$ which can be slow in exploring the parameters space and eventually come across computational issues. Greater flexibility can be obtained by allowing a multi-directional indexing as $\boldsymbol{\theta}_{h_1 \cdots h_p} = (\theta_{h_1,1}, \ldots, \theta_{h_p,p})'$. However, this leads to the tensor mixture model where one has to also consider how to carefully incorporate the simultaneous estimation of the multidimensional probability tensor.

The Dunson and Xing (2009) modeling of a probability tensor is one approach to defining and estimating a tensor within a mixture model setting. Banerjee et al. (2013) simultaneously generalized the tensor factorization to the infinite "arm" case (i.e. infinite tensor factorization, ITF) and developed the infinite tensor factorization mixture (ITM). The ITF is defined as

$$\boldsymbol{\pi} = \sum_{h=1}^{\infty} \lambda_h \bigotimes_{j=1}^{p} \boldsymbol{\psi}_{hj}, \tag{3.28}$$

where $\otimes$ denotes the outer product, $\boldsymbol{\lambda} \sim \text{stick}(\alpha)$, and $\boldsymbol{\psi}_{hj}$ is an infinite probability vector (otherwise known as "arms") with $\boldsymbol{\psi}_{hj} \sim \text{stick}(\beta_j)$. By introducing a latent indicator variables that denotes the location of a cell in the product infinite space such that $\text{pr}(C_1 = h_1, \ldots, C_p = h_p) = \pi_{h_1,\ldots,h_p}$, then the ITM is defined as a product kernel mixture model,

$$f(y_{i1} \in B_1, \ldots, y_{ip} \in B_p) = \sum_{h_1=1}^{\infty} \cdots \sum_{h_p=1}^{\infty} \pi_{h_1,\ldots,h_p} \prod_{j=1}^{p} \mathcal{K}_j(B_j; \theta_{h_j,j}), \tag{3.29}$$

where B_j corresponds to a Borel set from the product sigma algebra field generated from the product support space of the data vector y_i. This modeling scheme here addresses a larger problem called mixed domain modeling which accounts not only for mixed scaled data but also data of other objects such as functions, shapes, and images. The genius of this method is it achieves all desired purposes in that it is a simple straightforward extension of the simple mixture model that allows multi-directional exploration of the kernel parameter space that can be implemented via a not so simple Gibbs sampling scheme. Also, the unified approach to joint modeling here is its blessing and its curse since the joint clustering may heavily influence the information

fed into the estimation of the marginal structure, that is, $\boldsymbol{\theta}_j$ may be unduly influenced by all other variables. Additionally, even though it is an improvement upon the unidirectional mixture model, it too suffers from an inevitable breakdown point and is not readily available in software for routine implementation.

Thus, we propose a separation of the fully unified approach to the ITM via modularization where the first model captures a good estimate of each variable marginally then characterizes dependence via tensor factorization. This results in a dramatic conceptual and implementation simplification of the ITM while formalizing the current ad hoc approaches, and study how successful joint dependence is captured via marginal structures and account for uncertainty as an improvement to modularization by allowing module 1 to feed into module 2 at each iteration of the sampling which we call interweaving.

3.2.2 Data and model structure

In the previous section, we introduced the data vector of mixed scales, $\boldsymbol{y}_i$, as part of an $i.i.d$ sample of size n from an unknown joint density f and their respective scale indicators, s_j. We let $s_j = 1, 2$ correspond to categorical and count variables having support $\{0, 1, 2, \ldots, \infty\}$, respectively, with the remaining scales correspond to continuous variables with support on some subset $\mathcal{Y}_s \subset \Re$ of the real line. We proceed with the definition of the MOTEF algorithm by defining the module 1 and module 2 models.

Module 1 *(Marginal structures)*

To flexibly model each marginal variable with index $s_j \geq 2$, as stated in the previous section, we assume y_{ij} follows a mixture model as in 3.24. Note, the kernel, $\mathcal{K}_j$, is chosen appropriate for s_j. In hierarchical form, we have,

$$\begin{aligned} y_{ij} \sim \mathcal{K}_j(\theta^*_{x_{ij},j}, \tau_j), \quad \Pr(x_{ij} = h) = \omega_{hj}, \\ \boldsymbol{\omega}_j \sim \text{Dir}(\boldsymbol{a}_{j1}), \qquad\qquad \theta^*_{hj} \overset{iid}{\sim} P_{0j}, \end{aligned} \tag{3.30}$$

where $\boldsymbol{a}_{j1} = (a_{1j1}, \ldots, a_{d_j j1})'$, and τ_j is a shared global parameter that does not vary by sub-group. Note, this is similar to a DPM with the simplification that the weights do not follow a stick breaking process and are finite. We fix d_j at a large value to correspond to an over-fitted mixture prior model which has been found to favor automatic deletion of redundant components (Rousseau and Mengersen, 2011). When $s_j = 2$ so that the data are count-valued, we use the rounded Gaussian kernel mixture approach of (Canale and Dunson, 2011). When $s_j \geq 3$ we instead use mixtures of Gaussian kernels truncated to the support $\mathcal{Y}_j$. In a simple case in which none of the continuous variables have constrained support, we could have $s_j = 3$, $\mathcal{Y}_3 = \Re$ and can choose $\mathcal{K}_j$ to correspond to a Gaussian location-scale mixture (in which case $\theta^*_{hj} = (\mu_{hj}, \sigma_{hj})'$ is the location and scale and there is no τ_j). For categorical variables, $s_j = 1$, the structure of the mixture model is slightly

modified where we have:

$$y_{ij} \sim \delta_{x_{ij}}, \quad \Pr(x_{ij} = h) = w_{hj}, \quad \boldsymbol{\omega}_j \sim \text{Dir}(\boldsymbol{a}_{j1}), \tag{3.31}$$

where $\delta_B(A)$ is the Dirac measure which corresponds to assigning the value B with probability one when $A = B$, which simply passes the observed y_{ij} to x_{ij}.

Module 2 *(Interaction)*

In this module, we characterize the joint dependence among a functional of the collection of marginal allocation variables, $\boldsymbol{x}_i^* = g(\boldsymbol{x}_i) = g((x_{i1}, \ldots, x_{ip})')$, as a low rank probability tensor via an adaptation of the Dunson and Xing (2009) model presented in 3.27. The hierarchical model specification is as follows,

$$\begin{aligned} x_{ij}^* \sim \text{Multinomial}(1, \boldsymbol{\psi}_{z_i j}), \quad & \Pr(z_i = c) = \lambda_c, \\ \boldsymbol{\lambda} \sim \text{Dir}\big(\tfrac{1}{k_0}\mathbf{1}_{d_0}\big), \quad & \boldsymbol{\psi}_{hj} \overset{iid}{\sim} \text{Dir}(\boldsymbol{a}_{j2}), \end{aligned} \tag{3.32}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_0})'$, $\mathbf{1}_{d_0}$ denotes a size d_0 vector of ones, and $\boldsymbol{a}_{j2} = (a_{1j2}, \ldots, a_{d'_j j2})'$. Note, (3.32) corresponds to an over-fitted mixture of product multinomials while (3.27) is an infinite mixture of product multinomials with stick-breaking weights.

MOdularized TEnsor Factorization (MOTEF) algorithm

1. **Sample clusters independently**. Conduct posterior computation for each of the variables $j = 1, \ldots, p$ in parallel, to produce samples of x_{ij}, for $i = 1, \ldots, n$ based on the posterior distribution

$$\begin{aligned} & \pi(\boldsymbol{x}_j, \boldsymbol{\omega}_j, \boldsymbol{\theta}_j^*, \tau_j | \boldsymbol{y}_j), \\ \boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})', \quad & \boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})', \quad \boldsymbol{\theta}_j^* = (\theta_{1j}^*, \ldots, \theta_{d_j j}^*)' \end{aligned} \tag{3.33}$$

   which ignores information from the other variables $y_{i(-j)}$ and the interaction model (3.32).

2. **Sample interaction parameters**. Conduct posterior computation under model (3.30) based on the posterior distribution

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\psi}_{hj}, h = 1, \ldots, k, \boldsymbol{z} | \boldsymbol{x}_i^*, i = 1, \ldots, n), \quad \boldsymbol{z} = (z_1, \ldots, z_n)' \tag{3.34}$$

   where the functional of the latent cluster data $\boldsymbol{x}_i^* = (x_{i1}^*, \ldots, x_{ip}^*)'$ are treated as known in this step.

Precise details of the above algorithms are described in the next subsection for certain choices of kernel. In general, the modularization strategy *cuts* the dependence in conducting clustering on the individual-variable level to avoid *feedback* across the different variables, which can contribute to lack of robustness and

computational problems. Our unidirectional *interwoven* approach, that feeds module 1 allocation variables to module 2, additionally accounts for uncertainty in the marginal clustering which is in contrast to traditional modular who require a single point estimate $\hat{\boldsymbol{x}}_i$ to be fed to subsequent modules. This modularization is philosophically appealing in that it makes sense that the cluster allocation and modeling of the marginal distributions of the different variables does not depend on the other variables. In our separation of the marginals and the dependence structure, we provide a simplification to the ITM and a conceptually similar alternative to the Gaussian copula through discretization rather than using latent continuous variables (Banerjee et al., 2013; Hoff, 2007). Instead we fully specify a probability model for the marginals and accommodate richer dependence structure through our fully non-parametric interaction model in (3.27) which can accommodate most variable scales including nominal.

3.2.3 Posterior Computation

The sampling for MOTEF can be conducted in blocks. It in fact can be implemented via simple Gibbs sampling if the choice of module 1 permits. We are also interested in implementing an alternative form of MOTEF where we use stick-breaking priors in all mixture models so these modifications will be presented at the end of the section. Sampling steps are presented for the simple case where $s_j \in \{1, 2, 3\}$ for all j.

For variables with $s_j = 2$, we use the rounded Gaussian kernel:

$$\mathcal{K}_j\big(y_{ij}|\theta^*_{x_{ij},j}\big) = \int_{y_{ij}}^{y_{ij}+1} \mathrm{N}\big(y^*_{ij}|\mu_{x_{ij},j}, \sigma^2_{x_{ij},j}\big)dy^*_{ij},$$

where $\theta^*_{x_{ij},j} = (\mu_{x_{ij},j}, \sigma^2_{x_{ij},j})'$ (Canale and Dunson, 2011). Recall, for categorical variables, we set $x_{ij} = y_{ij}$ for $\{j : s_j = 1\}$. For count and continuous variables we choose P_{0j} to be the normal-inverse gamma prior, $\mathrm{N}\big(\mu_0, \frac{\sigma^2}{\kappa}\big) \times \text{Inv-Ga}(a_\sigma, b_\sigma)$.

Module 1

For each j update each component by sampling from each full conditional:

1. $\boldsymbol{\omega}_j|- \sim \text{Dir}(\tilde{\boldsymbol{a}}_{j1})$ where $\tilde{\boldsymbol{a}}_{j1} = (a_{1j1} + n_{1j1}, \ldots, a_{d_jj1} + n_{d_jj1})'$ and $n_{hj1} = \sum_{i=1}^n I(x_{ij} = h)$.

2. $\Pr\big(x_{ij} = h|-\big) \propto \omega_{hj}\mathcal{K}_j(y_{ij}|\theta^*_{hj})$ for $h = 1, \ldots, d_j$.

3. Update kernel specific parameters and latent variables

   (a) If $s_j = 3$, $(\mu_{hj}, \sigma^2_{hj})|- \sim \mathrm{N}\big(\tilde{\mu}_{hj}, \frac{\sigma^2_{hj}}{\tilde{\kappa}_{hj}}\big) \times \text{Inv-Ga}(\tilde{a}_{h\sigma_j}, \tilde{b}_{h\sigma_j})$, where $\tilde{a}_{h\sigma_j} = a_{\sigma_j} + n_{hj1}$, $\tilde{b}_{h\sigma_j} = b_{\sigma_j} + (SS_h + \tilde{\kappa}_{hj}p_{hj}(1 - p_{hj})(\bar{y}_{hj} - \mu_{0j})^2)/2$, $SS_h = \sum_{i:x_{ij}=h}(y_{ij} - \bar{y}_{hj})^2$, $p_{hj} = \frac{\kappa_j}{\kappa_j + n_{hj1}}$, $\bar{y}_{hj} = \sum_{i:x_{ij}=h} \frac{y_{ij}}{n_{hj1}}$, $\tilde{\kappa}_{hj} = \kappa_j + n_{hj1}$, and $\tilde{\mu}_{hj} = p_{hj}\mu_{0j} + (1 - p_{hj})\bar{y}_{hj}$ for $h = 1, \ldots, d_j$.

(b) If $s_j = 2$, update

i. $y^*_{ij}|- \sim \text{tN}(\mu_{hj}, \sigma^2_{hj}, y_{ij}, y_{ij}+1)$ for $i = 1, \ldots, n$ where tN denotes the truncated normal.

ii. $(\mu_{hj}, \sigma^2_{hj})|- \sim \text{N}\big(\tilde{\mu}_{hj}, \frac{\sigma^2_{hj}}{\tilde{\kappa}_{hj}}\big) \times \text{Inv-Ga}(\tilde{a}_{h\sigma_j}, \tilde{b}_{h\sigma_j})$ as in (a) with the exception that y_{ij} is replaced with y^*_{ij} in $\tilde{b}_{h\sigma_j}$ and $\tilde{\mu}_{hj}$.

Rather than pass along the latent allocation variables $\boldsymbol{x}_i$ into the interaction module, we incorporate an intermediate step that is a relabeling step. We observed greater efficiency in the dependence measures and computation time by passing along $\boldsymbol{x}^*_i$ into the interaction module instead, defined as the $\boldsymbol{x}^*_i \in 1, \ldots, d'_j$ where $d'_j = \sum_{k_j=1}^{d_j} I(n_{k_j j 1} > 0)$ for $\{j : s_j \neq 1\}$. In other words, unoccupied labels are discarded. Note, for non-nominal variables we pass through the observed data, $y_{ij} = x_{ij} = x^*_{ij}$ for $j : s_j = 1$. Also, note that $d'_j \leq d_j$ since we are relabeling and discarding unoccupied allocation labels.

Module 2

1. $\boldsymbol{\lambda}|- \sim \text{Dir}(\frac{1}{k_0} + m_1, \ldots, \frac{1}{k_0} + m_{d_0})$, where $m_l = \sum_{i=1}^n I(z_i = l)$ for $l = 1, \ldots, d_0$.

2. $\Pr(z_i = l|-) \propto \lambda_l \prod_{j=1}^p \psi_{l x^*_{ij}, j}$ for $l = 1, \ldots, d_0$.

3. $\boldsymbol{\psi}_{lj}|- \sim \text{Dir}(a_{1j2} + \sum_i I(z_i = l, x^*_{ij} = 1), \ldots, a_{d'_j j 2} + \sum_i I(z_i = l, x^*_{ij} = d'_j))$, for $l = 1, \ldots, d_0$.

In routine implementation, we recommend the following default choices for prior specifications in the simple case. For each j, set d_j to some large value that exceeds the number of suspected clusters and set $\boldsymbol{a}_{j1} = \frac{1}{k_j}\mathbf{1}_{d_j}$. We follow previously recommended convention that for j such that $s_j = 3$ variables be normalized and set $\mu_{0j} = 0$, $\kappa_j = 1$, $a_{\sigma_j} = 2$, and $b_{\sigma_j} = 4$ (Gelman et al., 2014). For count variables $s_j = 2$, we follow an empirical Bayes approach with $\mu_{0j} = \bar{y}_j$, $\kappa_j = 1/\hat{\sigma}^2_j$, $a_{\sigma_j} = 1$, and $b_{\sigma_j} = 2$ where $\bar{y}_j = \sum_{i=1}^n y_{ij}$ and $\hat{\sigma}^2_j = \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2/(n-1)$ (Gelman et al., 2014). In the second module, we let the "arm" assume a symmetric Dirichlet with $\boldsymbol{a}_{j2} = \mathbf{1}_{d'_j}$. In all of the over-fitted mixture model components we set the number of components to a large number relative to sample size, say 100 as a default, $d_j = 100$ for $j = 0, 1, \ldots, p$. And, in order to effectively zero out unnecessary components, we follow the recommended practice of setting Dirichlet parameters to $k_0 = k_j = 10^{25}$ for $\{j : s_j \neq 1\}$. This specification is equivalent to a sparse finite mixture models (Malsiner-Walli et al., 2016).

3.2.4 Inference

MOTEF does not have a unified model for defining an eMI. However, we combine each of its components to define a pseudo-joint function by taking components of each module. Module one, the marginal component, provides the cluster allocations which allow the transformation of the mixed-scale vector to become a p-way

contingency table. One appealing feature of the first module is that at each iteration a corresponding mixture model kernel can be identified for non-zero clusters. The non-zero clusters are important because for efficiency, unoccupied labels are discarded. Thus, the pseudo-joint distribution functions can be defined from the modular product kernel tensor mixture model:

$$\tilde{f}(\boldsymbol{y}_i) = \sum_{h_1=1}^{d_1'} \cdots \sum_{h_p=1}^{d_p'} \hat{\pi}_{h_1 \cdots h_p} \prod_{j=1}^{p} \mathcal{K}_j(y_{ij}|\hat{\boldsymbol{\theta}}_{h_j j}), \tag{3.35}$$

where $\hat{\pi}_{h_1 \cdots h_p}$ is the estimated tensor from the interaction module and $\hat{\boldsymbol{\theta}}_{h_j j}$ is the parameter corresponding to marginal module kernel parameters identified for the occupied labels.

The empirical mutual information measure for this model was defined as:

$$\zeta_{j_1 j_2} = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\sum_{h_{j_1}=1}^{d_{j_1}'} \sum_{h_{j_2}=1}^{d_{j_2}'} \hat{\pi}_{h_{j_1} h_{j_2}} \mathcal{K}_{j_1}(y_{ij_1}|\hat{\boldsymbol{\theta}}_{h_{j_1} j_1}) \mathcal{K}_{j_2}(y_{ij_2}|\hat{\boldsymbol{\theta}}_{h_{j_2} j_2})}{\tilde{f}_{j_1}(y_{ij_1}) \tilde{f}_{j_2}(y_{ij_2})} \right), \tag{3.36}$$

where $\tilde{f}_j(y_{ij}) = \sum_{h_j=1}^{d_j'} \hat{\pi}_{h_j} \mathcal{K}_j(y_{ij}|\hat{\boldsymbol{\theta}}_{h_j j})$.

A mutual information measure is computed at each iteration of a sampler and summarized by the mean, $\hat{\zeta}_{j_1 j_2} = \frac{1}{T} \sum_{t=1}^{T} \zeta_{j_1 j_2, t}$ for all pairs of variables. For testing, $\widehat{\Pr}(H_{1 j_1 j_2}|\boldsymbol{Y}) = \frac{1}{T} \sum_{t=1}^{T} \zeta_{j_1 j_2, t} > 0$ where $H_{1 j_1 j_2}$ is the alternative hypothesis of $H_{0 j_1 j_2} : Y_{j_1} \perp Y_{j_2}$. For defining dependence, we flag pairs of variables as dependent if $\widehat{\Pr}(H_{1 j_1 j_2}|\boldsymbol{Y}) > 1 - \alpha$.

3.3 Simulation Study

The performance of the proposed method was assessed via a simulation study. We compared the performance of MOTEF relative to the performance of the MPK. There is no standard or straightforward way of modeling a mixed-scale distribution so we defaulted to the MPK because of its robustness, simplicity in definition, and excellent computational performance. The aim of the simulation study was to compare the ability to: 1) adequately model the dependence structure; 2) convergence diagnostics; 3) the posterior number of clusters chosen.

The aims of our simulation study were rooted by our interest in determining how a collection of mixed-scale variables interrelate and jointly profiling or clustering observations. The multivariate dependence was assessed via the empirical mutual information measure defined in the previous section for MOTEF with an analogous version defined for MPK. Since posterior mutual information measures were used for determining the dependence structure, these posterior samples for all pairs were assessed for convergence with the multivariate potential scale reduction factor and effective sample size (Brooks and Gelman, 1998).

Given that we are using a distributional approach to clustering, it is important to monitor some measure of the number of classes chosen. We chose to monitor the posterior median number of clusters chosen defined as the number of occupied labels in the module 2 latent allocation variable $\boldsymbol{z}$. The MPK in difference to MOTEF only has one latent allocation variable which is used for defining the number of clusters. Additionally, we monitor the number of clusters chosen from an optimal partition of the data computed using the least-squares approach of Dahl (2006).

The simulated data consisted of $p = 20$ mixed-scale variables with 11 dichotomous 0/1 variables and three variables each for polytomous, continuous, and ordered categorical scales. 500 data sets were generated with the binary, polytomous, continuous, and ordered categorical variable blocks lined up adjacently in the specified order for variables 1 - 20 assuming dependence among variables $\{2,3,7,8,12,15,18\}$. That is, dependence was induced among four binary and one of each of the other types. We assumed data sets were composed of three subpopulations where a three-class latent subpopulations indicator was generated with probability (0.20, 0.55, 0.25). For binary and polytomous variables, the probability vector for each variable differed within each subpopulation for $j = 2, 3, 7, 8, 12$. Continuous variables were generated from two ($j = 15$) and three ($j = 16, 17$) component mixture models also with different mixture probability components within each subpopulation for $j = 15$. Similarly, ordinal variables were generated by applying the floor function on a latent continuous variable generated from a three component mixture with different mixture probabilities within each subpopulation for $j = 18$. Lastly, all data sets were of sample size 1000.

Each data set was analyzed separately using the Gibbs sampling scheme detailed in the previous section. For both the MOTEF and MPK procedures, their respective samplers were run for 5,000 iterations with a 1,000 iteration burn-in where every 4^{th} iteration was stored. Additionally, five separate chains were initialized at different starting allocation values for a total effective sample size of 5,000.

The results of the simulation study indicate that MOTEF and MPK differ slightly in elucidating the underlying dependence structure among the mixed scale variables. Figure 3.1 displays the proportion of simulations at each variable pair flagged as dependent. The range of proportions for flagging true dependent ranged between (0.42 - 1.00) and (0.73 - 1.00) for MOTEF and MPK, respectively. MOTEF notably outperforms MPK in flagging location (15,18) with this location flagged as dependent in 99% of the data sets while MPK flagged 73%. MPK notably outperforms MOTEF in flagging locations (7,18) and (8,18) with proportions 0.97 and 0.86, respectively, compared to MOTEF proportions of 0.42 and 0.74. False discovery proportions ranged from (0.00 - 0.12) for MOTEF and (0.00 - 0.08) for MPK. Overall, MPK marginally outperforms MOTEF in selecting the true dependence structure with slightly better performance at correctly flagging true dependence and lower false discovery proportions.

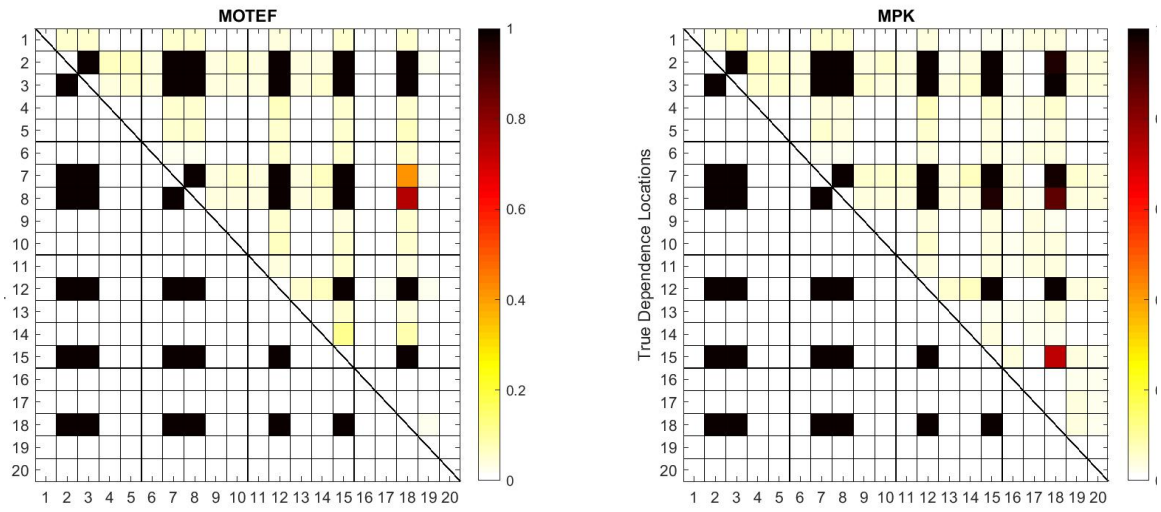

Figure 3.1: Results of simulations for MOTEF and MPK, which display percentages of simulations for each variable pair flagged as dependent, $\widehat{\Pr}(H_{1jj'} : \zeta_{jj'} > 0 | \boldsymbol{Y}) > 0.95$.

Table 3.1 displays simulation results for various diagnostics comparing MOTEF with MPK. The medians of the multivariate potential scale reduction factors for MOTEF and MPK were estimated to be 1.14 and 2.40, respectively. Generally, values less than 1.2 indicate the chains have achieved stationarity and mixed well which implies MOTEF outperforms MPK. MOTEF achieves good diagnostics on all pairwise empirical mutual information measures treated as multivariate, which are used to characterize the dependence structure of the simulated data sets. The median of the scaled multivariate effective sample size estimates also show the superiority of MOTEF over the MPK with values of 0.75 and 0.29, respectively. Scaled multivariate effective sample size values close to one indicate the multivariate posterior sample is akin to an independent identically distributed sample. Thus, the characteristics of the MOTEF posterior samples are much better than MPK.

MOTEF also outperformed the MPK in selecting a number of clusters closer to the true number of subpopulations. Table 3.1 also displays the proportion of data sets by posterior median number of clusters and the number of clusters (i.e. occupied components) from the least squares selected optimal partition for MOTEF and MPK. Across the overwhelming majority of data sets, MOTEF correctly estimated the posterior median number of clusters at three using the posterior median and the optimal partition with 98.8% and 96.6% of the 500 data sets. MPK on the other hand selected a greater number of clusters which ranged from four to seven for both the posterior median number of clusters and number of clusters from the optimal partitions. The mode of the median number of clusters for MPK was five in 58.2% data sets while the optimal number of clusters chose five clusters in 51.2% of the data sets.

Table 3.1: Simulation study results for various diagnostics comparing MOTEF with MPK; median (IQR) for continuous variables, percentages for integer valued diagnostics (out of 500 data sets)

Diagnostic	**MOTEF**	**MPK**
PSRF[1]	1.14	2.40
	(1.05 – 1.29)	(1.35 – 3.56)
Eff. Sample Size[2]	0.75	0.29
	(0.70 – 0.79)	(0.19 – 0.43)
Median no. clusters		
2	1.2 %	0 %
3	98.8 %	0 %
4	0.0 %	21.2 %
5	0.0 %	58.2 %
6	0.0 %	20.4 %
7	0.0 %	0.2 %
No. clusters - opt. part.		
2	3.4 %	0 %
3	96.6 %	0 %
4	0 %	18.8 %
5	0 %	51.2 %
6	0 %	26.8 %
7	0 %	3.2 %

[1] Multivariate potential scale reduction factor for empirical mutual information measures from all pairs of variables

[2] Scaled multivariate effective sample size by 5000 (1000 samples per chain, 5 chains).

Based on our simulation scenario, MOTEF has better performance over the MPK in efficiency and subpopulation estimation with somewhat comparable performance in elucidation the multivariate dependence structure. The proposed method has the ability to mostly characterize multivariate dependence correctly, and offers the ability to identify latent classes for profiling. Additionally, MOTEF has the potential for being computationally feasible for moderate to large data sets when coupled with parallelization for the first module components. Lastly, in addition to providing a smaller number of subclasses, which is useful for investigators seeking to jointly profile multiple mixed-scale variables, the modularization has the added feature providing information about the marginal clustering of the non-nominal categorical variables. This feature is not unique to MOTEF since it is also inherent in the ITM but MOTEF avoids the potential for undue feedback from the multivariate clustering mechanism as can be observed in ITM.

3.4 Application to Birth Defects Data

The National Birth Defects Prevention Study (NBDPS) consists of cases of birth defects ascertained through surveillance systems in participating sites: Arkansas, California, Georgia, Iowa, Massachusetts, New York, North Carolina, Texas, and Utah. Control infants (i.e. children without birth defects) are matched by geographic region through birth certificates or hospital records to cases. The study interviews mothers of cases and controls to obtain information on maternal demographic characteristics, medical history, lifestyle, occupational and environmental exposures, diet, and medication use before and during pregnancy. Paternal characteristics and exposures were also obtained. Motivated by jointly profiling variables of mixed scale, we take an exposome approach to jointly characterizing as many risk factors of select adverse outcomes as an application of our method. By the mixed scale nature of maternal/paternal risk factors observed by the NBDPS, we use this as a starting point for application of MOTEF and comparison to MPK.

In this application, we implement a stratified analysis of conotruncal heart and gastroschisis defects where an encompassing list of risk factors is available in NBDPS for both outcomes. Conotruncal heart defects encompass multiple heart defects which include: truncus arteriosus, transposition of the great arteries, double outlet right ventricle and tetralogy of Fallot. Gastroschisis is a defect in the anterior abdominal wall. We model the joint distribution of a set of risk factors on cases and controls. We include the case-control indicator as means of constructing supervised risk factor profiles. The set of risk factors included 46 mixed scale variables composed of maternal characteristics, medication use, lifestyle behaviors, proxys for fertility, and illness; and paternal characteristics and behaviors. The set of risk factors includes those identified by recent reviews of risk factors for congenital heart and gastroschisis defects which have some overlap and some outcome specific factors (Feng et al., 2014; Frolov et al., 2010; Patel and Burns, 2013; Rasmussen and Frías, 2008). MOTEF

and MPK are implemented for jointly profiling (i.e. identify clusters) the list of risk factors by defect. Analysis was restricted to complete cases.

We ran both methods for 25,000 iterations with the first 10,000 discarded as burn-in samples, storing every 15^{th} sample on five separate chains each. MOTEF was implemented with 100 mixture modeling components and 30 components for each of the continuous and ordinal variable mixture modeling components. MPK was implemented with 100 mixture modeling components.

The optimal partition for MOTEF indicated there were a smaller number of clusters than the optimal partition for MPK. MPK observed 42 clusters in the gatroschisis and conotruncal analyses. MOTEF, on the other hand, identified 9 clusters for both analyses. In terms of cluster size, MPK selected clusters with sizes varying from 7.5% to $<$1% while MOTEF selected clusters with sizes between 28.0% and 4.6%. See Table 3.2 for a summary of the dependence clusters for both methods.

Table 3.2: Summary of optimal partition for dependence allocation variables for MOTEF and MPK by outcome

	Gastroschisis				**Conotruncal**			
	MOTEF		**MPK**		**MOTEF**		**MPK**	
Cluster	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
1	2268	23.2	733	7.5	2984	28.0	788	7.4
2	1377	14.1	573	5.9	2043	19.1	569	5.3
3	1129	11.6	504	5.2	1265	11.9	539	5.1
4	1081	11.1	483	4.9	1073	10.1	469	4.4
5	994	10.2	462	4.7	974	9.1	454	4.3
6	937	9.6	450	4.6	801	7.5	453	4.2
7	889	9.1	415	4.3	532	5.0	444	4.2
8	584	6.0	401	4.1	503	4.7	416	3.9
9	505	5.2	381	3.9	496	4.6	370	3.5
10	0	0.0	337	3.5	0	0.0	361	3.4
11	0	0.0	318	3.3	0	0.0	360	3.4
12	0	0.0	312	3.2	0	0.0	353	3.3
13	0	0.0	309	3.2	0	0.0	351	3.3
14	0	0.0	299	3.1	0	0.0	301	2.8
15	0	0.0	277	2.8	0	0.0	296	2.8
16	0	0.0	266	2.7	0	0.0	291	2.7
17	0	0.0	242	2.5	0	0.0	267	2.5

Table 3.2: Summary of optimal partition for dependence allocation variables for MOTEF and MPK by outcome

	Gastroschisis				Conotruncal			
	MOTEF		MPK		MOTEF		MPK	
Cluster	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
18	0	0.0	226	2.3	0	0.0	254	2.4
19	0	0.0	200	2.0	0	0.0	254	2.4
20	0	0.0	190	1.9	0	0.0	251	2.4
21	0	0.0	188	1.9	0	0.0	240	2.2
22	0	0.0	180	1.8	0	0.0	210	2.0
23	0	0.0	180	1.8	0	0.0	210	2.0
24	0	0.0	167	1.7	0	0.0	199	1.9
25	0	0.0	166	1.7	0	0.0	194	1.8
26	0	0.0	144	1.5	0	0.0	168	1.6
27	0	0.0	126	1.3	0	0.0	157	1.5
28	0	0.0	126	1.3	0	0.0	134	1.3
29	0	0.0	125	1.3	0	0.0	133	1.2
30	0	0.0	120	1.2	0	0.0	129	1.2
31	0	0.0	120	1.2	0	0.0	117	1.1
32	0	0.0	106	1.1	0	0.0	113	1.1
33	0	0.0	100	1.0	0	0.0	109	1.0
34	0	0.0	96	1.0	0	0.0	105	1.0
35	0	0.0	95	1.0	0	0.0	93	0.9
36	0	0.0	90	0.9	0	0.0	92	0.9
37	0	0.0	79	0.8	0	0.0	87	0.8
38	0	0.0	76	0.8	0	0.0	86	0.8
39	0	0.0	54	0.6	0	0.0	81	0.8
40	0	0.0	45	0.5	0	0.0	80	0.7
41	0	0.0	2	0.0	0	0.0	53	0.5
42	0	0.0	1	0.0	0	0.0	40	0.4
All	9764	100.0	9764	100.0	10671	100.0	10671	100.0

One of the more appealing properties of MOTEF, or more generally tensor mixture models of product kernels, is its ability to generate a smaller number of clusters relative to MPK. A large number of clusters often implies the presence of several clusters with minor differences, or small and singleton class sizes. These pose small sample difficulties for parameter precision and post-hoc analyses. Investigator input is then required to collapse certain groups together, which can defeat the utility of latent class analyses as methods that reveal underlying subclasses with distributional criteria. Additionally, when attempting to profile and interpret clusters, a smaller set is easier to manage and reduces the number of comparisons in post-hoc pairwise comparisons. The tensor mixture methodology has two layers of dependence, where the first induces dependence among a set of mixed-scale variables via a set of marginal (i.e. variable specific) latent classes and the second induces dependence among the set of marginal latent variables via a joint latent class. The set of marginal latent variables are assumed mutually independent within each latent class and as such the distribution of each marginal latent class can be summarized within each cluster. A smaller number of joint latent classes implies the probability tensor (i.e multivariate categorical variables) has a simpler structure and a single joint cluster implies mutual independence among all variables, conversely, a larger number of joint latent classes implies a more complex dependence structure.

Cluster profile summaries can aid investigators in comparing the composition of clusters. Profile summaries are constructed by taking all cross tabulations between each marginal and joint latent classes and displaying cell specific cells. For instance the proportion of defect cases, can be computed and compared by cluster by scaling the number of cases within a cluster relative to its total cluster size. Once proportions are computed, then other measures such as the odds ratios can be computed for comparing clusters.

Figures 3.2 and 3.3 display optimal cluster risk factor profiles using the MOTEF dependence partitions (i.e. joint latent classes) which are heat map plots of cross-tabulations between each risk factor latent class variable and the optimal joint cluster for the gastroschisis and conotruncal defect analyses, respectively. Each column represents one joint cluster. Risk factors are represented across rows corresponding to each class level and the proportions displayed sum to one within each joint class. Marginal risk factor class levels are represented by their class mean and level for non-categorical and categorical risk factors, respectively. Binary categorical risk factors however display only one level. For instance, all medication use is binary and only displays yes proportions, while site is displayed across 10 rows with each site labeled accordingly. The optimal clusters are labeled in ascending order by descending sample size. That is, cluster 1 is the largest group while cluster 9 is the smallest. Also, each profile panel is ordered in descending order by proportion of defect cases in each group from left to right. In what follows, cluster profiles are summarized and compared separately by birth defect.

Table 3.3: Summary of cases by defect type and MOTEF clusters.

	Gastroschisis		Conotruncal	
Cluster	**n**	**%**	**n**	**%**
1	34	1.5	540	18.1
2	42	3.1	411	20.1
3	233	20.6	172	13.6
4	42	3.9	182	17.0
5	325	32.7	163	16.7
6	93	9.9	139	17.4
7	70	7.9	114	21.4
8	4	0.7	116	23.1
9	173	34.3	86	17.3

The derived gastroschisis latent classes seem to uphold the evidence of many of the established risk factors. The proportion of cases varied greatly by cluster from 32.7% for cluster 5 to <1% for cluster 8, see Table 3.3. In what follows we highlight some of the more obvious patterns in the profile compositions and how these relate to risk factors of gastroschisis such as: parental age, maternal BMI, substance abuse, smoking, income, birth weight, gestational age, and infection (Frolov et al., 2010; Rasmussen and Frías, 2008). One of the immediately striking observations is the association between case status and parental age. Mother's age was found to have three marginal latent classes: late teens - early 20s (median: 19, IQR: (18 - 21)), mid 20s (median: 26, IQR: (23 - 28)), and late 20s - mid 30s (median: 31, IQR: (28 - 34)), while father's age was found to have four marginal latent classes: late teens - early 20s (median: 21, IQR: (19 - 23)), mid - late 20s (median: 27, IQR: (24 - 30)), late 20s - mid 30s (median: 32, IQR: (28 - 36)), and late 30s - mid 40s (median: 42, IQR: (37 - 46)). The clusters with the three largest proportions of gastroschisis cases, clusters 9, 5, and 3, were composed of 53.1%, 66.1%, and 47.7% of mothers marginally classified as late teens - early 20s, respectively. By contrast, the clusters with the smallest proportion of gastroschisis cases, clusters 2, 1, and 8, were composed of 51.3%, 73.2%, and 67.0% of mothers classified as late 20s - mid 30s, respectively. Father's age correlates with the maternal age, as visible in clusters 2, 1, and 8 which were composed of 58.5%, 69.8%, and 67.5% of fathers classified as late 30s - mid 40s, respectively. Furthermore, the odds ratios of gastroschisis for clusters 9 and 5 relative to cluster 2 were 16.6 (95% CI: 15.5 - 17.7) and 15.4 (95% CI: 14.6 - 16.3), respectively. The profile of cluster 9 supports associations with substance abuse, smoking, and lower income, since this cluster contains large proportions of such risk factors (maternal SA: 77.0%, paternal SA: 84.6%, maternal smoke: 77.6%, household smoke: 52.5%, income<10K: 33.1%). For comparison, the odds of gastroschisis defect for cluster 9 was estimated to be 2.0 (95% CI: 1.9 - 2.1) times the odds for cluster 3, a cluster composed of mostly young parents with little substance abuse and smoking. The protective effect of advanced parental age and higher BMI can be observed by comparing cluster 5 with cluster 2 which has similar characteristics except for larger percentage of parents classified as non- late teens - early 20s (2: 95.8% vs 5:

53.6%) and overweight/obese mothers (2: 56.2% vs 5: 26.8%). The odds ratio of gastroschisis comparing cluster 5 to cluster 2 was estimated to be 15.4 (95% CI: 14.6 - 16.3).

Some of the derived latent class profiles have similarities with demographic and exposure profiles observed in previous NBDPS analyses of gastroschisis defects which include decreased risk with greater household income and increased risk with younger women, smoking, and alcohol consumption (Mac Bird et al., 2009). While our results generally overlap with these findings, our approach to deriving profiles was dependent on the discovery of latent classes. As such, it is not possible to disentangle the effect of particular risk factors for direct comparison with previous findings.

Three marginal latent classes were identified for both birth weight and gestational age. Birth weight was classified as infants with class means at near very low (mean: 3.9 lbs, IQR: (2.7 - 5.1)), near average (mean: 6.9 lbs, IQR: (5.9 - 7.9)), and just above average (mean: 7.6 lbs, IQR: (7.0 - 8.2)) US birth weight. Similarly, gestational age was classified as infants with class means at very preterm (mean: 31.2 weeks, IQR: (28 - 35)), preterm (36.8 weeks, IQR: (35 - 38)), and full term (39.1 weeks, IQR: (38 - 40)). The four subclasses with greatest proportion of cases had proportions of low birth weight and very preterm at $>3.9\%$ and $>3.6\%$, respectively, compared to $<2.4\%$ and $<2.2\%$ in the rest of the classes, which can be used to support associations between gastroschisis and low birth weight and low gestational age. Additionally, the proportions of kidney/bladder/UTI infection by latent class correlate almost perfectly with the proportions of cases as they observe an almost monotonic decreasing trend (9: 17.8%, 5: 14.7%, 3: 10.3%, 6: 9.4%, 7: 7.9%, 4: 6.1%, 2: 6.5%, 1: 3.2%, 8: 3.1%).

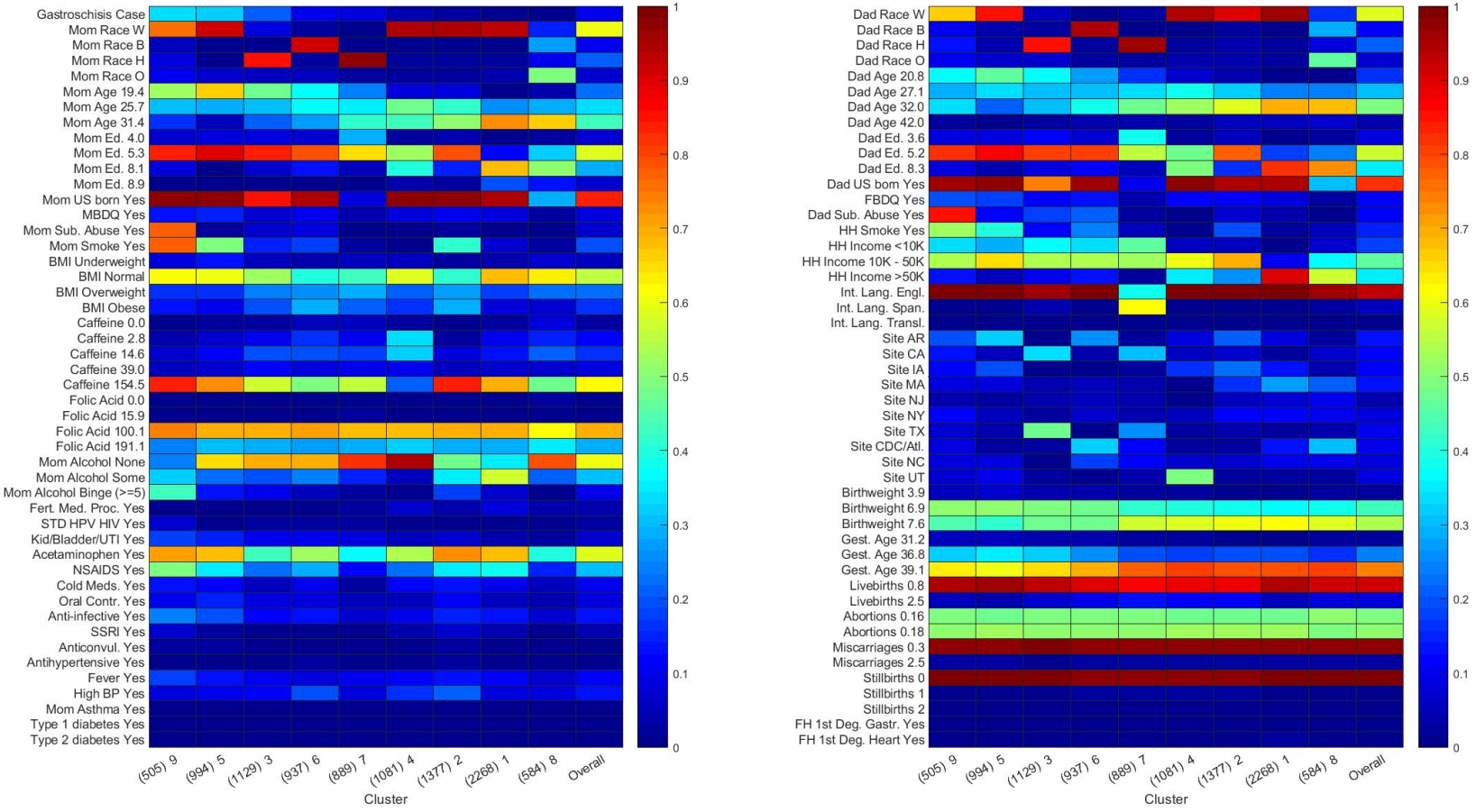

Figure 3.2: Gastroschisis optimal cluster risk factor profiles using MOTEF dependence partitions.

As expected, the profiles for conotruncal heart defects differ significantly from those for gastroschisis, and in some ways the associations appear to be more complex (Figure 3.3). Because it is difficult to tease out general observations, we describe the profiles with the greatest proportion of cases. Clusters 8 and 7 had the greatest proportion of cases at 23.1% and 21.4%, respectively. By comparison, cluster 3 had the smallest proportion at 13.6% and the overall was 18.0% (n=1,923). Cluster 8 surprisingly had a very healthy profile with a large proportion mothers with a class mean age of 30.9 (73.2%), and the greatest proportion of mothers with normal BMI (71.6%) and the smallest proportion of overweight/obese mothers (20.3%). Cluster 7, on the other hand, had highest proportions of parental substance abuse (maternal: 66.7%, paternal: 82.7%), smoking (maternal: 77.8%, household: 47.4%), and indications of illness (STD/HPV/HIV: 6.0%, Kid/Bladder/UTI: 15.4%, Fever: 17.1%). The odds of conotruncal defect for clusters 8 and 7 were increased by 90% and 70%, respectively, relative to the odds for cluster 3. Some characteristics such as: low and high BMI, advanced parental age, diabetes, febrile illness, and smoking, have been identified as risk factors for congenital heart defects in general and specifically for conotruncal heart defects in review and previous NBDPS studies (Adams et al., 1989; Botto et al., 2014; Feng et al., 2014; Gilboa et al., 2010; Green et al., 2010; Malik et al., 2008; Patel and Burns, 2013). It is reassuring to see our method capture some of these characteristics in the derived subclasses.

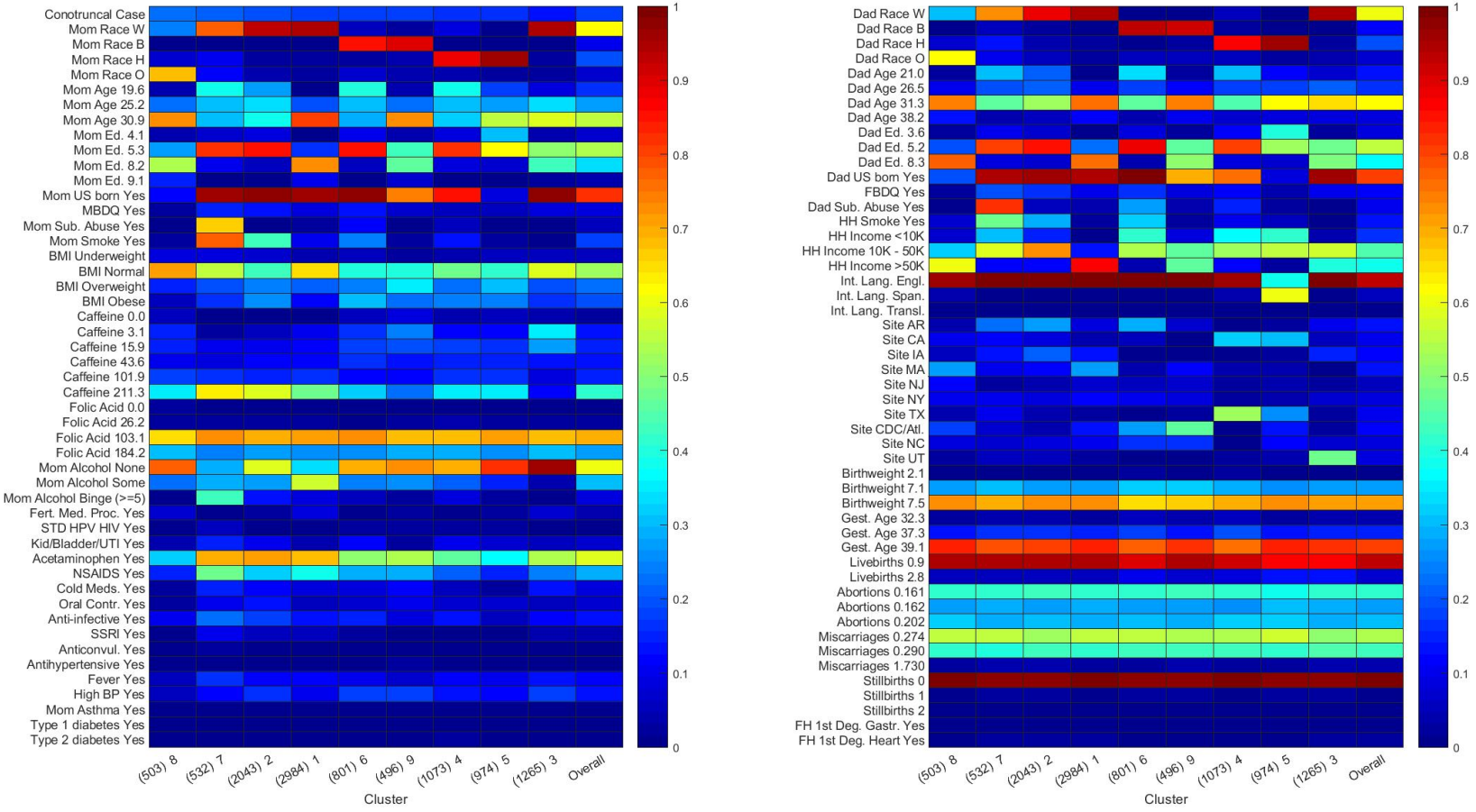

Figure 3.3: Conotroncal optimal cluster risk factor profiles using MOTEF dependence partitions.

3.5 Discussion

In this article, we have proposed a simpler form of the tensor mixture model with product kernels via modularization where the complexity of analyzing mixed-scale data is simplified by essentially discretizing non-nominal categorical variables to analyze the entire set of variables as a multivariate contingency table. This is conceptually similar to the copula model where assumed marginal structures are combined with a unifying copula distribution model. Further, it can be thought of as an extension to the exploration of pairwise dependence via Dirichlet process mixtures of Filippi et al. (2016) albeit we assess dependence with different measures. Similar to the method proposed by Hoff (2007) our model assumes unknown marginal distributions but unified via a PARAFAC tensor model. It was shown that under certain scenarios it can appropriately elucidate the underlying joint dependence structure as well as offer greater compression of latent subclasses and improved MCMC diagnostics. The greater compression of latent subclasses can be particularly useful in elucidating subclass variable mixture profiles in comparison to Dirichlet process mixtures which is known for over selecting number of subclasses, which at times may require investigator input to collapse certain groups (Stephenson et al., 2017). MOTEF can be useful as an exploratory data analysis tool for describing multivariate mixed-scale profiles in moderate to large sets of variables. This framework can be scalable due to its modularization since all of the first modules are independent of all other component and as such be computed in parallel.

In spite of the appealing performance of MOTEF relative to MPK, there are some notable opportunities for improvement. First, it was our experience that there was variability in the elucidation of the dependence structure with respect to the measure used. One of our initial attempts at elucidating the dependence of MOTEF was with the model based Cramer's V statistic proposed by Dunson and Xing (2009), however, we opted to use the empirical mutual information measure since there is no analogous Cramer's V measure available for the MPK. Additionally, we found the empirical mutual information measure with the pseudo-joint tensor mixture defined in this paper to outperform the Cramer's V measure. There is a possibility that this measure supports the idea that it is beneficial to use the observed data when assessing dependence in the mixed-scale setting. Filippi et al. (2016) used two different approaches than ours where: 1) included using the latent marginal allocation variables to construct contingency tables for testing, and 2) included a modeling approach where pairs of variables are modeled as a two-component mixture of marginal independent mixture models and an MPK type model. Our approach is somewhat of a mixture between both of their approaches, however, more investigation is required to determine which method or whether an alternative may perform better under various scenarios. Second, there may be a more favorable label switching move that may help the efficiency of MOTEF. We incorporated a simple label collapsing move that at each step of the algorithm occupied labels are

relabeled from 1 up to the number of non-zero classes, $\boldsymbol{x}_i^*$, into the dependence module because improved efficiency was observed over directly inputting $\boldsymbol{x}_i$ into module two. The third point of possible improvement, is one that generally plagues all mixture modeling with product kernels, the special case of uni-modal correlated distributions. As a special case, MPK and MOTEF will presumably perform poorly if data is generated from a bi-variate normal with non-zero correlation. We intend to extend the tensor mixture of product kernels formulation to the tensor mixture model with dependent kernels which can accommodate this special case.

CHAPTER 4: BAYESIAN SEMI-PARAMETRIC MODELING WITH VARIABLE SELECTION FOR HIERARCHICAL INTERACTIONS

4.1 Introduction

The difficulties that arise when estimating the joint or simultaneous effect of correlated predictors are well known. Recent calls for the need to develop methodology that can account for estimation instability resulted in the development of statistical methods for addressing this issue (Davalos et al., 2017; Hamra and Buckley, 2018). Statistical methods used for overcoming difficulties with highly correlated predictors are necessary tools for assessing health effects of exposure to complex chemical mixtures (e.g. air pollution, phthalates, PBCs, etc.), and joint effects of multiple exposures in the exposome setting. In epidemiologic studies, in addition to estimating effects, it is also of interest in clustering effects and identifying those associated with a response (Dunson et al., 2008).

There are few statistical approaches that can address estimation issues with collinearity as well as simultaneously incorporate variable selection and effect clustering. MacLehose et al. (2007) introduced a Bayesian semi-parametric model where regression coefficients of correlated predictors are assumed to have a Dirichlet process (DP) prior. The DP prior base measure is assumed to have spike-slab base measure with a point mass at zero which facilitates variable selection. The DP prior as an almost sure discrete random measure achieves clustering and greater flexibility over normal priors (Ferguson, 1973). Additionally, as a Bayesian variable selection method, this semi-parametric framework simultaneously accounts for multiple comparison and incorporates model averaging.

The Bayesian nature of the MacLehose et al. (2007) modeling framework offers appealing ways to incorporate prior information. The non-parametric Bayes prior allows incorporating prior information on the probability of exclusion via a specification of the hyperprior on the weight associated with the point mass at zero in the DP prior. The inclusion of the point mass at zero frees up the hyperparameters on the slab portion of the DP prior for incorporating prior information on the magnitude of the effect estimates. Without a spike, then effect priors are often centered around zero for shrinking and variable selection. Lastly, the hyperprior on the precision parameter of the DP prior allows informing the degree of clustering the regression coefficients which can be used to favor or discourage a priori a high degree of clustering in the predictor effects.

Herring (2010) extended this methodology by grouping main effects and pairwise interaction effects separately via assigning different independent DP priors on each coefficient block of terms, thereby incorporating separate degrees of shrinkage by block or group. This methodology, albeit presented for pairwise interactions, is more general by including different blocks in the model and assigning independent DP priors on each block. For illustrative purposes and without loss of generality, we present our framework in terms of the pairwise interaction generalized linear model. Let y_i be an outcome of interest and the objective of the analysis is to quantify the possibly non-homogeneous joint effect of a set of predictors, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, while accounting for an encompassing set of confounding variables, $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{ip_0})'$. Consider the mean regression model with pairwise interaction terms:

$$g(\mu_i) = \mathbf{z}_i'\boldsymbol{\kappa} + \boldsymbol{x}_i'\boldsymbol{\beta}_1 + \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \beta_{j_1 j_2} x_{ij_1} x_{ij_2}, \tag{4.37}$$

where $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)'$, and $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \ldots, \kappa_{p_0})'$. This modeling specification allows investigators to separately incorporate different degrees shrinkage for the main and interaction term effects. However, the specification of independent priors for each of the coefficient blocks does not account for hierarchical interaction terms. That is, if a main effect coefficient drops out of the model, it does not force all of the corresponding interaction terms to drop out. Hierarchical interaction terms are of interest to investigators because these facilitate interpretation and may positively influence estimation.

Hierarchical interactions impose a restriction on all lower order terms. For instance, in the case of model (4.37), hierarchical interactions imply:

$$\beta_j = 0, \textit{ only if } \beta j_1 j_2 = 0 \text{ for } j_1, j_2 = j. \tag{4.38}$$

One way to incorporate this condition is by building a joint prior for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_2 = \{\beta_{j_1 j_2}, j_1 = 1, \ldots, p-1, j_2 = j_1 + 1, \ldots, p\}$ such that

$$\text{Pr}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \text{Pr}(\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2) \text{Pr}(\boldsymbol{\beta}_2), \tag{4.39}$$

where

$$\text{Pr}(\beta_j | \boldsymbol{\beta}_2) = \text{Pr}_1(\beta_j)\delta(A) + \text{Pr}_2(\beta_j)\delta(A^c), \tag{4.40}$$

and $A = \{\beta_{j_1 j_2} = 0, \text{ for all } j_1, j_2 = j\}$. For the pairwise interaction model, this is certainly achievable where $\text{Pr}(\boldsymbol{\beta}_2) = \pi_0\delta_0 + (1-\pi_0)G_0$ and $G_0 = \text{DP}$, $\text{Pr}_2(\beta_j) = G_1$, and $\text{Pr}_1(\beta_j) = \pi_1\delta_0 + (1-\pi_1)G_1$ and $G_1 = \text{DP}$. However, there are certainly computational and tractability issues that should arise when incorporating higher

order interactions. Thus, our aim is to develop a less expensive framework which can achieve the incorporation of higher order interactions circumventing the hierarchical iterated conditioning by taking a decomposition of the regression coefficients.

Regression coefficients are often decomposed to achieve difference purposes. Suh et al. (2011) used a decomposition logistic regression coefficients of air pollutants within a two-step analyses where these were decomposed by pollutant and chemical properties. We define a decomposition of regression coefficients into independent components that when taken together achieve condition (4.38).

There are dual objectives with this chapter, to develop: 1) and study the performance of a Bayesian semi-parametric model with hierarchical interaction selection (BHIS) thereby extending the Herring (2010) framework, and; 2) a Matlab package for Bayesian semi-parametric modeling with blocked variable selection (BBVS) which can be used for BHIS. The BBVS Matlab package will be able to handle n-way interaction selection. The chapter proceeds as follows: Section 4.2 motivates and defines BHIS, Section 4.3 present results from simulations, Section 4.4 applies our methodology to the Mount Sinai Children's Environmental Health Study, and concludes with a brief discussion.

4.2 Bayesian Hierarchical Interaction Selection

In the derivation that follows below, for illustrative purposes, consider model (4.37) where $p = 2$. The framework below is applicable without loss of generality to the n-way interactions case.

4.2.1 Derivation

In a decomposition of effects, taking the sum of independent effects on lower order terms facilitates the elucidation of priors within a Bayesian analysis which reduce to zero only if each of the component effects is zero. Consider the case where $p = 2$ and for purposes of notation let the interaction effect equal an independent effect, $\gamma_{12} = \beta_{12}$. The main effects thus become a sum of a main effect specific component and all corresponding higher order effects,

$$\begin{aligned} g(\mu_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} \\ &= \beta_0 + (\gamma_1 + \gamma_{12}) x_{i1} + (\gamma_2 + \gamma_{12}) x_{i2} + \gamma_{12} x_{i1} x_{i2} \\ &= \beta_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_{12}(x_{i1} + x_{i2} + x^*_{i1} x_{i2}) \\ &= \beta_0 + \gamma_1 x^*_{i1} + \gamma_2 x^*_{i2} + \gamma_{12} x^*_{i12}, \end{aligned}$$

where $x^*_{ij} = x_{ij}$ and $x^*_{i12} = (x_{i1}+1)(x_{i2}+1)-1$. From this, we observe that under our formulation we can incorporate ordered variable selection for hierarchical interactions by simply modeling a different a model with a different design matrix. In the propositions that follow, we detail the structure of the design matrix required for BHIS for (4.37) and the three-way interaction model.

Proposition 1: Under the pairwise interaction effects model with independent coefficient effects, the corresponding BHIS model is

$$g(\mu_i) = \gamma_0 + \boldsymbol{\gamma}_1' \boldsymbol{x}_i + \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \gamma_{j_1 j_2}((x_{ij_1}+1)(x_{ij_2}+1)-1), \tag{4.41}$$

where $\boldsymbol{\gamma}_1 = (\gamma_1, \ldots, \gamma_p)'$.

Proposition 2: Under the three-way interaction effects model with independent coefficient effects ($p \geq 3$), the corresponding BHIS model is

$$g(\mu_i) = \gamma_0 + \boldsymbol{\gamma}_1' \boldsymbol{x}_i + \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \gamma_{j_1 j_2} \left(\prod_{l=1}^{2} (x_{ij_l}+1) - 1 \right) + \sum_{j_1=1}^{p-2} \sum_{j_2=j_1+1}^{p-1} \sum_{j_3=j_2+1}^{p} \gamma_{j_1 j_2 j_3} \left(\prod_{l=1}^{3} (x_{ij_l}+1) - 1 \right), \tag{4.42}$$

where $\boldsymbol{\gamma}_1 = (\gamma_1, \ldots, \gamma_p)'$.

Proposition 3: Under the n-way interaction effects model with independent coefficient effects ($3 < n \leq p$), the corresponding BHIS model is

$$\begin{aligned} g(\mu_i) = \gamma_0 + \boldsymbol{\gamma}_1' \boldsymbol{x}_i &+ \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \gamma_{j_1 j_2} \left(\prod_{l=1}^{2} (x_{ij_l}+1) - 1 \right) + \\ &\sum_{j_1=1}^{p-2} \sum_{j_2=j_1+1}^{p-1} \sum_{j_3=j_2+1}^{p} \gamma_{j_1 j_2 j_3} \left(\prod_{l=1}^{3} (x_{ij_l}+1) - 1 \right) + \cdots + \\ &\sum_{j_1=1}^{p-(n-1)} \sum_{j_2=j_1+1}^{p-(n-2)} \cdots \sum_{j_n=j_{n-1}+1}^{p} \gamma_{j_1 j_2 \cdots j_n} \left(\prod_{l=1}^{n} (x_{ij_l}+1) - 1 \right), \end{aligned} \tag{4.43}$$

where $\boldsymbol{\gamma}_1 = (\gamma_1, \ldots, \gamma_p)'$.

By proposition 3, we can incorporate any number of interactions (i.e. the n-way interaction model) by simply modeling an alternative model with a design matrix of equivalent rank. Thereby, not increasing the complexity or the number of terms needed to perform the analysis.

4.2.2 Hierarchical Model

In this section, we develop a model corresponding to *Proposition 1*. The above framework presented a framework for the corresponding model when lower order regression coefficients have a decomposition which can facilitate variable selection for hierarchical interactions. The novelty in the decomposition is that it allows the elucidation of independent priors by different blocks where then lower order terms are only set to zero if their corresponding higher order terms also take the values zero. As such we specify the priors for the coefficient components as:

$$\gamma_j \sim \pi_{01}\delta_0 + (1-\pi_{01})G_1,\ j = 1, \ldots, p$$

$$\gamma_{j_1 j_2} \sim \pi_{02}\delta_0 + (1-\pi_{02})G_2,\ 1 \leq j_1 < j_2 \leq p$$

$$G_l = DP(\alpha_l, G_{0l}),\ G_{0l} = N(\mu_{Bl}, \sigma^2_{Bl}),$$

$$\pi_{0l} \sim \text{Beta}(a_{\pi l}, b_{\pi l})$$

$$\alpha_l \sim \text{Gamma}(a_{\alpha l}, b_{\alpha l})$$

$$(\mu_{Bl}, \sigma^2_{Bl}) \sim \text{N}(\cdot|\mu_{0Bl}, \tau^{-1}_{Bl}\sigma^2_{Bl})\text{Inv-Gamma}(\cdot|a_{\sigma Bl}, b_{\sigma Bl}),$$

for $l = 1, 2$. Note, that as the magnitude of the interactions grows, the index of the component coefficients γ grows and the inclusion of an additional interaction will define another block of components distributed as $\pi_{0l}\delta_0 + (1-\pi_{0l})G_l$.

4.2.3 Posterior Computation

For purposes of notation, assume the joint distribution of the outcomes is denoted by $\mathcal{K}\big(\mathbf{y}|\{\mathbf{X}_l\}_{l=1}^L, \mathbf{Z}, \{\boldsymbol{\beta}_l\}_{l=1}^L, \boldsymbol{\kappa}, \boldsymbol{\theta}\big)$ where $\boldsymbol{\theta}$ can be a global parameter across all subjects or a subject specific or both depending on the need, $\mathbf{X}_l = [\mathbf{x}_{1l}, \mathbf{x}_{2l}, \ldots, \mathbf{x}_{nl}]'$, and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]'$. Note, the Dirichlet Process prior assumed on each block of the regression coefficients requires the introduction of L pairs of infinite parameters that characterize the probabilities and the domain values of the DP and when combined with the spike for simultaneous variable selection requires the introduction of some notation for the distribution G_l. This is visible as follows:

$$G_l(\cdot|\boldsymbol{\psi}_l, \boldsymbol{B}_l) = \sum_{h=0}^{\infty} \psi_{hl}\delta_{B_{hl}}(\cdot),$$

where $\boldsymbol{\psi}_l = \{\psi_{hl}\}_{h=0}^{\infty}$, $B_{0l} = \{0\}$, $B_{hl} \overset{iid}{\sim} G_{0l}$, $\boldsymbol{\psi}_l = \{\pi_{0l}, (1-\pi_{0l})\boldsymbol{\lambda}_l\}$, $\boldsymbol{\lambda}_l = \{\lambda_{hl}\}_{h=1}^{\infty}$, $\lambda_{hl} = \nu_{hl}\prod_{h'<h}(1-\nu_{h'l})$, and $\nu_{hl} \overset{iid}{\sim} \text{Beta}(1, \alpha_l)$ for $l = 1, \ldots, L$ and $h = 1, \ldots, \infty$.

The full joint distribution is thus:

$$f\big(\boldsymbol{y},\boldsymbol{\kappa},\boldsymbol{\theta},\{\boldsymbol{\beta}_l\}_{l=1}^L,\boldsymbol{\pi}_0,\{\boldsymbol{B}_l\}_{l=1}^L,\{\boldsymbol{\nu}_l\}_{l=1}^L,\{\boldsymbol{\theta}_{Bl}\}_{l=1}^L,\boldsymbol{\theta}_\kappa,\boldsymbol{\alpha}\big) =$$
$$\mathcal{K}\big(\mathbf{y}|\{\mathbf{X}_l\}_{l=1}^L,\mathbf{Z},\{\boldsymbol{\beta}_l\}_{l=1}^L,\boldsymbol{\kappa},\boldsymbol{\theta}\big)\pi(\boldsymbol{\kappa}|\boldsymbol{\theta}_\kappa)\pi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_\kappa)$$
$$\times\prod_{l=1}^L\Big\{\prod_{j=1}^{p_l}G_l\big(\beta_{jl}|\boldsymbol{\psi}_l,\mathbf{B}_l\big)\pi(\pi_{0l})\prod_{h=1}^{\infty}\pi\big(\nu_{hl}|\alpha_l\big)G_{0l}\big(B_{hl}|\boldsymbol{\theta}_{Bl}\big)\Big\}\pi(\boldsymbol{\theta}_{Bl})\pi(\alpha_l),$$

where $\boldsymbol{\pi}_0=\{\pi_{0l}\}_{l=1}^L$ and $\boldsymbol{\alpha}=\{\alpha_l\}_{l=1}^L$.

The Dirichlet process assumption on the regression coefficient blocks requires specialized techniques to make sampling tractable given the infinite stick breaking parametrization. As previously stated, the Polya-urn sampler has been successfully used in modeling the GLM semi-parametric model (Dunson et al., 2008). This sampling technique however, is not completely straightforward especially for updating the precision parameter. A straightforward alternative could be the blocked Gibbs sampler by Ishwaran and James (2001) but can potentially be computationally intensive for large truncation. We consider slice sampling as an alternative that combines the efficiency of the Polya-urn and the ease of updating precision parameters via the stick-breaking representation as would be used in a truncation (Ishwaran and James, 2001; Mena and Walker, 2015). In the following we focus on incorporating the slice sampling on the prior of a single block of coefficients since we will capitalize on the independence assumption for block updating and potential speed-ups through parallelization in computation. Consider the prior for a single block of coefficients, β_{jl}:

$$\prod_{j=1}^{p_l}G_l\big(\beta_{jl}|\boldsymbol{\psi}_l,\mathbf{B}_l\big)=\prod_{j=1}^{p_l}\sum_{h=0}^{\infty}\psi_{hl}\delta_{B_{hl}}(\beta_{jl})$$
$$=\sum_{h_1=0}^{\infty}\cdots\sum_{h_{p_l}=0}^{\infty}\prod_{j=1}^{p_l}\psi_{h_jl}\delta_{B_{h_jl}}(\beta_{jl}),$$

where the equivalence is established by iterated distributive law and noting the change in indexing the second equivalence. To make the iterated infinite sum tractable, at this point we incorporate slice sampling as was done in Mena and Walker (2015) by introducing a regression block specific latent uniform variable and a deterministic parameter $\boldsymbol{\xi}_l$. This results in a joint prior as:

$$\pi\big(\boldsymbol{\beta}_l,u_l|\boldsymbol{\psi}_l,\boldsymbol{B}_l\big)=\sum_{h_1=0}^{\infty}\cdots\sum_{h_{p_l}=0}^{\infty}1(u_l<\xi_{h_1\cdots h_{p_l},l})\xi^{-1}_{h_1\cdots h_{p_l},l}\prod_{j=1}^{p_l}\psi_{h_jl}\delta_{B_{h_jl}}(\beta_{jl}),$$

where $\xi_{h_1\cdots h_{p_l},l}=\prod_{j=1}^{p_l}(1-\omega_l)\omega_l^{h_j}$. The introduction of the latent uniform has the effect of reducing the dimensionality of the labeling space from infinite to finite and the independent geometric assumption on the

deterministic parameter has the effect of jointly favoring smaller labels with almost no gaps (Mena and Walker, 2015). We introduce an allocation variable from the finite set that indicates if $\beta_{jl} = B_{hl}$ then $d_{jl} = h$. This implies:

$$\pi\big(\boldsymbol{\beta}_l, \boldsymbol{d}_l, u_l | \boldsymbol{\psi}_l, \boldsymbol{B}_l\big) = 1(u_l < \xi_{d_{1l}\cdots d_{p_l l}, l}) \xi^{-1}_{d_{1l}\cdots d_{p_l l}, l} \prod_{j=1}^{p_l} \psi_{d_{jl}l} \delta_{B_{d_{jl}l}}(\beta_{jl}).$$

We augment the full joint distribution with uniform variables and latent allocation variables:

$$\begin{aligned} f\big(\boldsymbol{y}, \boldsymbol{\kappa}, \boldsymbol{\theta}, \{\boldsymbol{\beta}_l\}_{l=1}^L, \boldsymbol{\pi}_0, \{\boldsymbol{B}_l\}_{l=1}^L, \{\boldsymbol{\nu}_l\}_{l=1}^L, \{\boldsymbol{\theta}_{Bl}\}_{l=1}^L, \boldsymbol{\theta}_\kappa, \boldsymbol{\alpha}, \boldsymbol{u}, \{\boldsymbol{d}_l\}_{l=1}^L\big) = \\ \mathcal{K}\big(\mathbf{y} | \{\mathbf{X}_l\}_{l=1}^L, \mathbf{Z}, \{\boldsymbol{\beta}_l\}_{l=1}^L, \boldsymbol{\kappa}, \boldsymbol{\theta}\big) \pi(\boldsymbol{\kappa}|\boldsymbol{\theta}_\kappa) \pi(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_\kappa) \\ \times \prod_{l=1}^L \Bigg\{ \prod_{j=1}^{p_l} 1(u_l < \xi_{d_{1l}\cdots d_{p_l l}, l}) \xi^{-1}_{d_{1l}\cdots d_{p_l l}, l} \prod_{j=1}^{p_l} \psi_{d_{jl}l} \delta_{B_{d_{jl}l}}(\beta_{jl}) \pi(\pi_{0l}) \\ \times \prod_{h=1}^{\infty} \pi\big(\nu_{hl} | \alpha_l\big) G_{0l}\big(B_{hl} | \boldsymbol{\theta}_{Bl}\big) \Bigg\} \pi(\boldsymbol{\theta}_{Bl}) \pi(\alpha_l), \end{aligned}$$

where $\boldsymbol{u} = (u_1, \ldots, u_L)'$ and $\boldsymbol{d}_l = \{d_{jl}\}_{j=1}^{p_l}$.

The posterior sampling will depend on the underlying assumption of the outcome y_i. Consider the linear regression setting:

$$\boldsymbol{y} = \sum_{l=1}^L \boldsymbol{X}_l \boldsymbol{\beta}_l + \boldsymbol{Z}\boldsymbol{\kappa} + \boldsymbol{\epsilon}, \tag{4.44}$$

where $\boldsymbol{y} = \{y_i\}_{i=1}^n$ is an $(n \times 1)$ vector, $\boldsymbol{X}_l$ is a $(n \times p_l)$ matrix for $l = 1, \ldots, p_l$, $\boldsymbol{\kappa} = \{\kappa_j\}_{j=1}^{p_0}$ is a $(p_0 \times 1)$ vector, $\boldsymbol{\beta}_l = \{\beta_{jl}\}_{j=1}^{p_l}$ is a $(p_l \times 1)$ vector for $l = 1, \ldots, p_l$, and $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ with $\boldsymbol{\Sigma}_\epsilon = \sigma^2_\epsilon \boldsymbol{I}_n$. We further assume $\boldsymbol{\kappa} \sim \mathrm{N}_{p_0}(\boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa)$, $\boldsymbol{\mu}_\kappa = \mu_\kappa \boldsymbol{J}_{p_0}$, $\boldsymbol{\Sigma}_\kappa = \sigma^2_\kappa \boldsymbol{I}_n$, and $(\mu_\kappa, \sigma^2_\kappa) \sim \mathrm{N}(\mu_{0\kappa}, \tau^{-1}_\kappa \sigma^2_\kappa)\text{Inv-G}(a_{\sigma\kappa}, b_{\sigma\kappa})$. Below we inspect the posterior distributions for certain updates and detail a Gibbs sampling algorithm for the linear regression setting:

1. Coefficient block specific slice variable: $\pi(u_l|-) = 1(u_l < \xi_{d_{1l}\cdots d_{p_l l}, l})$

   - For $l = 1, \ldots, L$ sample $u_l \sim \mathrm{U}(0, \xi_{d_{1l}\cdots d_{p_l l}, l})$.

2. Coefficient block specific measure weights:

   (a) Point mass at zero weight: $\pi(\pi_{0l}|-) \propto \pi(\pi_{0l}) \prod_{j=1}^{p_l} \psi_{d_{jl}l}$.

   - For $l = 1, \ldots, L$ sample $\pi_{0l} \sim \text{beta}(\hat{a}_{\pi l}, \hat{b}_{\pi l})$ where $\hat{a}_{\pi l} = a_{\pi l} + n_{0l}$, $\hat{b}_{\pi l} = b_{\pi l} + m_{0l}$, $n_{hl} = \sum_{j=1}^{p_l} 1(d_{jl} = h)$ for $h = 0, \ldots, \tilde{d}_l$, $\tilde{d}_l = \max\{\boldsymbol{d}_l\}$, and $m_{hl} = \sum_{h'>h} n_{h'l}$.

   (b) DP weights components: $\pi(\{\nu_{hl}\}_{h=1}^{K_l}|-) \propto \prod_{j=1}^{p_l} \psi_{d_{jl}l} \prod_{h=1}^{K_l} \pi\big(\nu_{hl}|\alpha_l\big)$, where $K_l = \max\{A_{1l}, \ldots, A_{p_l l}\}$.

- For $h = 1, \ldots, \tilde{d}_l$ and $l = 1, \ldots, L$ sample $\nu_{hl} \sim \text{beta}(1 + n_{hl}, \alpha_l + m_{hl})$.

3. DP atoms and confounder coefficients:

$$\pi(\boldsymbol{\kappa}, \{\boldsymbol{B}_l\}_{l=1}^L|-) \propto \mathcal{K}\big(\mathbf{y}|\{\mathbf{X}_l\}_{l=1}^L, \mathbf{Z}, \{\boldsymbol{\beta}_l\}_{l=1}^L, \boldsymbol{\kappa}, \boldsymbol{\theta}\big)\pi(\boldsymbol{\kappa}|\boldsymbol{\theta}_\kappa)\prod_{l=1}^{L}\prod_{h=1}^{K_l}\pi\big(B_{hl}|\boldsymbol{\theta}_{Bl}\big)$$

   - Sample $\boldsymbol{B^*} \sim \text{N}_{p^*}(\hat{\boldsymbol{\mu}}_{B^*}, \hat{\boldsymbol{\Sigma}}_{B^*})$ where $\boldsymbol{B^*} = \{\{B_{hl}, h = 1, \ldots, \tilde{d}_l\}_{l=1}^L, \boldsymbol{\kappa}\}$ is a $(p^* \times 1)$ vector, $\hat{\boldsymbol{\Sigma}}_{B^*} = (\boldsymbol{X^{*\prime}}\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X^*} + \boldsymbol{\Sigma}_{B^*}^{-1})^{-1}$, $\hat{\boldsymbol{\mu}}_{B^*} = \hat{\boldsymbol{\Sigma}}_{B^*}(\boldsymbol{X^{*\prime}}\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}_{B^*}^{-1}\boldsymbol{\mu}_{B^*})$, $\boldsymbol{X^*} = \big[\boldsymbol{X}_1^*, \cdots, \boldsymbol{X}_L^*, \boldsymbol{Z}\big]$, $\boldsymbol{X}_l^* = \boldsymbol{X}_l\boldsymbol{W}_l$, $\boldsymbol{W}_l = \{1(d_{jl} = h)\}_{jh}$ is a $(p_l \times \tilde{d}_l)$ indicator matrix, $\boldsymbol{\mu}_{B^*} = \{\{\boldsymbol{\mu}_{Bl}\}_{l=1}^L, \boldsymbol{\mu}_\kappa\}$ is a $(p^* \times 1)$ vector, $p^* = p_0 + \sum_{l=1}^L \tilde{d}_l$, $\boldsymbol{\mu}_{Bl} = \mu_{Bl}\boldsymbol{J}_{\tilde{d}_l}$, $\boldsymbol{\Sigma}_{B^*} = \text{blkdiag}(\boldsymbol{\Sigma}_{B1}, \ldots, \boldsymbol{\Sigma}_{BL}, \boldsymbol{\Sigma}_\kappa)$, and $\boldsymbol{\Sigma}_{Bl} = \sigma_{Bl}^2\boldsymbol{I}_n$.

4. Global kernel parameter: $\pi(\boldsymbol{\theta}|-) \propto \pi(\boldsymbol{\theta})\mathcal{K}\big(\mathbf{y}|\{\mathbf{X}_l\}_{l=1}^L, \mathbf{Z}, \{\boldsymbol{\beta}_l\}_{l=1}^L, \boldsymbol{\kappa}, \boldsymbol{\theta}\big)$

   - Sample $\sigma_\epsilon^2 \sim \text{Inv-Gamma}(\hat{a}_\epsilon, \hat{b}_\epsilon)$ where $\hat{a}_\epsilon = a_\epsilon + n/2$, $\hat{b}_\epsilon = b_\epsilon + \boldsymbol{e}'\boldsymbol{e}/2$, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{B^*}$

5. Coefficient block specific DP base measure parameters: $\pi(\boldsymbol{\theta}_{Bl}|-) \propto \pi(\boldsymbol{\theta}_{Bl})\prod_{h=1}^{K_l}\pi\big(B_{hl}|\boldsymbol{\theta}_{Bl}\big)$.

   - For $l = 1, \ldots, L$ sample:

     a. $\sigma_{Bl}^2 \sim \text{Inv-Gamma}(\tilde{a}_{\sigma Bl}, \tilde{b}_{\sigma Bl})$ where $\tilde{a}_{\sigma Bl} = a_{\sigma Bl} + \tilde{d}_l$, $\tilde{b}_{\sigma Bl} = b_{\sigma Bl} + \sum_{h=1}^{\tilde{d}_l}(B_{hl} - \bar{B}_l)^2/2 + \dfrac{\tau_{Bl}\tilde{d}_l}{\tau_{Bl} + \tilde{d}_l}(\bar{B}_l - \mu_{0Bl})^2/2$, and $\bar{B}_l = \sum_{h=1}^{\tilde{d}_l} B_{hl}$.

     b. $\mu_{Bl} \sim \text{N}(\tilde{\mu}_{Bl}, \tilde{\sigma}_{Bl}^2)$ where $\tilde{\sigma}_{Bl}^2 = \sigma_{Bl}^2/(\tau_{Bl} + \tilde{d}_l)$, and $\tilde{\mu}_{Bl} = \dfrac{\tau_{Bl}}{\tau_{Bl} + \tilde{d}_l}\mu_{0Bl} + \dfrac{\tilde{d}_l}{\tau_{Bl} + \tilde{d}_l}\bar{B}_l$.

6. Confounder coefficient hyper-parameters: $\pi(\boldsymbol{\theta}_\kappa|-) \propto \pi(\boldsymbol{\theta}_\kappa)\pi\big(\boldsymbol{\kappa}|\boldsymbol{\theta}_\kappa\big)$.

   - Sample:

     a. $\sigma_\kappa^2 \sim \text{Inv-Gamma}(\tilde{a}_{\sigma\kappa}, \tilde{b}_{\sigma\kappa})$ where $\tilde{a}_{\sigma\kappa} = a_{\sigma\kappa} + p_0$, $\tilde{b}_{\sigma\kappa} = b_{\sigma\kappa} + \sum_{h=1}^{p_0}(\kappa_h - \bar{\kappa})^2/2 + \dfrac{\tau_\kappa p_0}{\tau_\kappa + p_0}(\bar{\kappa} - \mu_{0\kappa})^2/2$, and $\bar{\kappa} = \sum_{h=1}^{p_0}\kappa_h$.

     b. $\mu_\kappa \sim \text{N}(\tilde{\mu}_\kappa, \tilde{\sigma}_\kappa^2)$ where $\tilde{\sigma}_\kappa^2 = \sigma_\kappa^2/(\tau_\kappa + p_0)$, and $\tilde{\mu}_\kappa = \dfrac{\tau_\kappa}{\tau_\kappa + p_0}\mu_{0Bl} + \dfrac{p_0}{\tau_\kappa + p_0}\bar{\kappa}$.

7. Coefficient latent allocation variables: $\pi(d_{jl} = h|-) \propto \psi_{hl}\omega_l^{-h}\mathcal{K}\big(\mathbf{y}|\{\mathbf{X}_l\}_{l=1}^L, \mathbf{Z}, \{\boldsymbol{\beta}_l\}_{l=1}^L, \boldsymbol{\kappa}, \boldsymbol{\theta}\big)$ for $h = 1, \ldots, A_{jl}$, $A_{jl} = \max\{h : u_l < \xi_{d_{1l}\cdots d_{j-1,l}hd_{j+1,l}\cdots d_{p_l l},l}\}$, $j = 1, \ldots, p_l$, and $l = 1, \ldots, L$. Note, $\boldsymbol{\beta}_{l'} = (\beta_{1l'}, \ldots, \beta_{p_{l'}l'})' = (B_{d_{1l'}l'}, \ldots, B_{d_{p_{l'}l'}l'})'$ then the corresponding coefficient is updated as $\beta_{jl} = B_{d_{jl}l}$.

   - For $j = 1, \ldots, p_l$, and $l = 1, \ldots, L$ sample d_{jl} such that:

$$\Pr(d_{jl} = h) \propto \frac{\psi_{hl}}{\omega_l^h}\text{N}_n\big(\boldsymbol{e}_{jl}|\boldsymbol{X}_{jl}B_{hl}, \boldsymbol{\Sigma}_\epsilon\big),$$

where $\boldsymbol{X}_{jl}$ is the j^{th} column of $\boldsymbol{X}_l$, $\boldsymbol{e}_{jl} = \boldsymbol{y} - \sum_{l' \neq l} \boldsymbol{X}_{l'}\boldsymbol{\beta}_{l'} - \boldsymbol{X}_{-jl}\boldsymbol{\beta}_{-jl}$, $\boldsymbol{X}_{-jl}$ is the $\boldsymbol{X}_l$ matrix without the j^{th} column, $\boldsymbol{\beta}_{-jl}$ is the l^{th} coefficient block without the j^{th} coefficient, and $h = 0, \ldots, d_{jl} + \left\lfloor \dfrac{\log u_l^*}{\log \omega_l} \right\rfloor$. Note, $\lfloor \cdot \rfloor$ is the floor function.

8. Coefficient block specific precision parameters: $\pi(\alpha_l|-) \propto \pi(\alpha_l) \prod_{h=1}^{K_l} \pi\big(\nu_{hl}|\alpha_l\big)$.

   - For $l = 1, \ldots, L$ sample $\alpha_l \sim$ gamma$(\hat{a}_{\alpha l}, \hat{b}_{\alpha l})$ where $\hat{a}_{\alpha l} = a_{\alpha l} + \tilde{d}_l$ and $\hat{b}_{\alpha l} = b_{\alpha l} - \sum_{h=1}^{\tilde{d}_l} \log(1 - \nu_{hl})$.

The linear regression case is straightforward since we have conjugate priors for the atom and confounder coefficient parameters. Other types such as binomial, negative binomial, and Poisson may require data augmentation to preserve the Gibbs sampling scheme.

4.2.4 Testing

Under the proposed method, we can test for the exclusion of any term within hierarchical interactions models. In the case of the pairwise interaction model, we may test for exclusion of the interaction effects: $H_{0j_1j_2} : \beta_{j_1j_2} = 0$ which is equivalent to $H_{0j_1j_2} : \gamma_{j_1j_2} = 0$ for $1 \leq j_1 < j_2 \leq p$. We may also test for the exclusion of main effects: $H_{0j} : \beta_j = 0$ which is equivalent to $H_{0j} : \gamma_j = \gamma_{j_1j} = \gamma_{jj_2} = 0$ for $1 \leq j_1 < j < j_2 \leq p$. From the sampling output, we may compute posterior probabilities $\widehat{\Pr}(H_{0j_1j_2}|\boldsymbol{Y}, \boldsymbol{X})$ and $\widehat{\Pr}(H_{0j}|\boldsymbol{Y}, \boldsymbol{X})$ by simply counting the number of times the respective conditions are met across the sampler and dividing by the total number of iterations. That is,

$$\widehat{\Pr}(H_{0j_1j_2}|\boldsymbol{Y}, \boldsymbol{X}) = \frac{\sum_{t=b}^{B} I(\gamma_{j_1j_2,t} = 0)}{B - b} \tag{4.45}$$

$$\widehat{\Pr}(H_{0j}|\boldsymbol{Y}, \boldsymbol{X}) = \frac{\sum_{t=b}^{B} I(\gamma_{j,t} = 0, \gamma_{1j,t} = 0, \ldots, \gamma_{j-1j,t} = 0, \gamma_{jj+1,t} = 0, \ldots, \gamma_{jp,t} = 0)}{B - b}, \tag{4.46}$$

where b corresponds to a value of a burn in period and B corresponds to the total number of iterations a chain MCMC is run.

4.3 Simulation Experiments

The purpose of the simulation experiments was to assess the performance of our proposed method in various scenarios within the n-way interaction linear regression setting. We attempt to simulate scenarios arising in environmental epidemiology, where the aim is to assess the simultaneous effect of chemical mixtures (i.e. five continuous predictors) while adjusting for mixed-scale covariates (i.e. confounding variables) which

are not subject to variable selection. The primary focus of the experiments was on variable selection with a secondary focus on estimation.

4.3.1 Data structure

All generated data had the structure of the three-way interaction regression model. Letting y_i be the outcome of interest each data set, we specified:

$$g(\mu_i) = \kappa_0 + \mathbf{z}_i'\boldsymbol{\kappa} + \boldsymbol{x}_i'\boldsymbol{\beta}_1 + \sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\beta_{j_1j_2}x_{ij_1}x_{ij_2} + \sum_{j_1=1}^{p-2}\sum_{j_2=j_1+1}^{p-1}\sum_{j_3=j_2+1}^{p}\beta_{j_1j_2j_3}x_{ij_1}x_{ij_2}x_{ij_3}, \quad (4.47)$$

where $E(y_i) = \mu_i$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip_0})'$, $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)'$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{p_0})'$, and $y_i \sim \mathrm{N}(\mu_i, \sigma^2)$. Note, the similarities between equation 4.47 and 4.44, where the $\boldsymbol{z}$ covariates (confounders) are elucidated as such because these are variables which will not be subject to selection.

The set of predictors, $\boldsymbol{x}$, and confounding variables, $\boldsymbol{z}$, are jointly generated. Each data set consisted of five continuous predictors, $p = 5$, along with 10 mixed-scale confounding variables, $p_0 = 10$. Details of the data generating mechanism for these variables are presented in Appendix A. In brief, the predictors and confounding variables were jointly generated from a rounded induced density with varying correlation among all variables (Canale and Dunson, 2015; Kunihama et al., 2016). High correlation was induced among all predictors with correlations varied from 0.52 to 0.92. Correlation among confounding variables varied from -0.1 to 0.2. Lastly, cross correlation between predictors and confounders varied between -0.3 and 0.1.

The coefficients for covariates not subject to selection were set to either -1 or 1 across all data sets. Specific values were:

$$(\kappa_0, \boldsymbol{\kappa}) = (1, 1, 1, -1, 1, 1, -1, 1, 1, 1, -1).$$

The specification of the true values for the predictor coefficients (i.e. variables subject to selection) varied according to the simulation scenario. However, the following remained fixed across all scenarios $\beta_1 = \beta_3 = \beta_4 = \beta_5 = 1$.

4.3.2 Simulation Scenarios

Eight simulation scenarios were considered to assess performance when the true data generating mechanism contains strong and combined (i.e. strong and weak) hierarchical interactions, as well as, assess the effect of incorporating high correlation between one confounding variable (always confounding variable 4) and one predictor (either x_2 or x_5). Within each simulation scenario, 500 data sets were generated each with sample size 500.

The simulation scenarios are summarized in Table 4.4. Scenario 1 was the main effects simulation scenario where $\beta_2 = 1$. Scenarios 2 to 7 had the following pairwise interaction effects set to one, $\beta_{1,3} = \beta_{1,4} = \beta_{1,5} = \beta_{3,4} = 1$, thereby inducing strong hierarchy among predictors 1,3, and 4. Scenarios 5 to 7 additionally had $\beta_{2,3} = 1$ which induced weak hierarchy among predictors 2 and 3. Scenario 8 had pairwise interactions specified in scenarios 2 - 4 and the three-way interaction between predictors 1, 3, and 4 set to one, $\beta_{1,3,4} = 1$. Among all scenarios if coefficients are not specified then they were set to zero. Scenarios 3 and 6 additionally had high correlation induced between predictor 5 and a confounding variable. Scenarios 4 and 7 had high correlation induced between predictor 2 and a confounding variable. Scenario 4 should help to assess whether the correlation of among a confounding variable confounds the null predictor 2.

Table 4.4: Summary of simulation scenarios

	Hierarchy type	β_2	$\beta_{2,3}$	$\beta_{1,3,4}$	**Predictor with high conf. corr.**
Scenario 1	main effects	1	0	NA	NA
Scenario 2	strong	0	0	NA	NA
Scenario 3	strong	0	0	NA	x_5
Scenario 4	strong	0	0	NA	x_2
Scenario 5	combined	0	1	NA	NA
Scenario 6	combined	0	1	NA	x_5
Scenario 7	combined	0	1	NA	x_2
Scenario 8	strong	0	0	1	NA

4.3.3 Methods

The BHIS was implemented using Proposition 1 for scenarios 1 to 7, while Proposition 2 was used in assessing scenario 8. These implementations used different block specifications for the main effect, pairwise, and three-way interactions where applicable, using model 4.44. That is, each block had an independent set specific probability of exclusion built into the model.

We additionally applied a BBVS model where, similar to the BHIS, separate non-parametric Bayes priors were assumed on regression coefficients by main effects and interaction degree as in Herring (2010). Recall, this specification will allow the estimation procedure to consider models interaction effect only models.

Standard Bayesian procedures were followed in the implementation of BBVS and BHIS. For each model (i.e. model within each data set within each scenario), five MCMC chains were run each for 15,000 iterations discarding the first 5,000 as a burn-in and storing every 10^{th} iteration for a total effective posterior sample size of 5,000. Weakly informative priors were used for most parameters with the exception of interaction terms where prior specification favoring exclusion was used. That is, the expected prior probability of exclusion for each interaction term was set 0.9.

We compared the performance of the above with two gold standard frequentist methods, least absolute shrinkage and selection operator (LASSO) and one of its hierarchical interactions selection extension group-lasso interaction network (GLINTERNET). The ubiquitously used LASSO was chosen because of it is a gold standard for analysis in the presence of correlated predictors and the R package glmnet allows the implementation of LASSO with a penalty on certain variables (Friedman et al., 2009; Tibshirani, 1996). GLINTERNET was chosen as a frequentist alternative that can implement hierarchical interactions, however, its R implementation requires a specification of a model with all pairwise interactions (Lim and Hastie, 2013, 2015). As such, GLINTERNET was implemented on all pairwise interaction models in all scenarios with the predictors and confounding variables as candidates for exclusion. GLINTERNET was not considered in scenario 8 because it cannot incorporate three-way interactions.

4.3.4 Results

Figure 4.4 displays average sensitivity and false discovery proportion and relative model across all data sets and scenarios by method. Sensitivity proportion was computed as the average of proportions of terms in the true model included in the fitted model across all data sets. For LASSO and GLINTERNET, the sensitivity proportion for a data set is determined by the number of true term estimates not equal to zero. The sensitivity proportion for BBVS and BHIS is determined by the number of true terms whose posterior probability for inclusion exceed 0.5, or equivalently, whose posterior log odds for inclusion greater than zero. False discovery proportion was computed as the average of proportions of terms not in the true model included in the fitted model across all data sets. Relative model size was computed as average of the number of terms in the fitted model relative to the number of terms in the true model across all data sets.

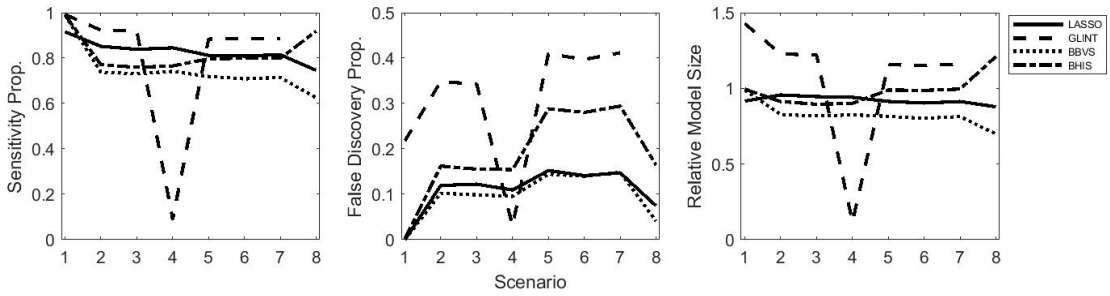

Figure 4.4: Results of all simulation scenarios displaying mean of summary measures across all 500 data sets.

All methods mostly performed comparably in terms of sensitivity across all scenarios with the exception of a few scenarios. All methods except for LASSO had perfect performance in terms of sensitivity in the main effects only scenario, scenario 1. In scenarios 2 through 7, LASSO, BBVS, and BHIS had similar sensitivity performance with average proportions around 80%. GLINTERNET seemed slightly outperform all

other methods with the highest sensitivity proportion averages except in scenario 4, where its performance really suffered. This may indicate that GLINTERNET is confounded by having high correlation between confounding variables and non-significant predictors. In scenario 8, the proposed BHIS slightly outperformed all other methods an average sensitivity proportion of approximately 90% while LASSO and BBVS observed averages less than 80%.

The false discovery proportion averages revealed that the higher performing methods in terms of sensitivity also came with trade offs in false discovery. For instance, GLINTERNET in the main effects scenario performed perfectly in choosing significant predictors but observed a false discovery proportion of approximately 20%. The same holds for GLINTERNET across all scenarios with the exception of scenario 4. The same is reflected in terms of relative model size. In terms of false discovery overall, LASSO and BBVS had improved performance over GLINTERNET and the proposed BHIS across all scenarios with average false discovery proportions less than 15%.

The proposed BHIS was the best performer in scenario 8. Table 4.5 displays coefficient median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean or estimate across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion for BBVS and BHIS. Although BHIS had greater false discoveries, in this scenario: BBVS failed to identify the only true three-way interaction as significant and LASSO identified it only 19% of the time, while the proposed identified it in 72%. Additionally, the median estimate of $\beta_{1,3,4}$ across all data sets in LASSO was 0.00 with an associated $2.5^{th} - -97.5^{th}$ percentiles of 0.00 – 0.94 while the corresponding estimates for BHIS were 0.41 with percentiles straddling the true value 1, (0.05 – 1.12).

For details of each simulation scenario performance by method, see Appendix B and C.

Table 4.5: Simulation scenario 8 results summary comparing LASSO, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		**BBVS**		**BHIS**	
	true	**Median**	**Incl Prop**	**Median**		**Median**	
Coeff.	**value**	**Estimate**	**Sel flag**	**Estimate**	**Incl Prop**	**Estimate**	**Incl Prop**
β_1	1	1.13 (0.21 – 2.00)	0.99	1.07 (0.62 – 1.58)	0.99	1.17 (0.64 – 1.34)	0.99
β_2	0	0.00 (0.00 – 0.54)	0.20	0.28 (0.01 – 0.88)	0.40	0.15 (0.00 – 0.61)	0.29

Table 4.5: Simulation scenario 8 results summary comparing LASSO, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		**BBVS**		**BHIS**	
	true	**Median**	**Incl Prop**	**Median**		**Median**	
Coeff.	**value**	**Estimate**	**Sel flag**	**Estimate**	**Incl Prop**	**Estimate**	**Incl Prop**
β_3	1	0.80 (0.00 – 1.62)	0.96	0.93 (0.41 – 1.29)	0.97	0.84 (0.32 – 1.24)	0.98
β_4	1	0.96 (0.10 – 1.85)	0.99	1.04 (0.55 – 1.47)	0.99	1.00 (0.42 – 1.32)	0.99
β_5	1	0.68 (0.00 – 1.53)	0.94	1.01 (0.34 – 1.33)	0.96	0.70 (0.07 – 1.19)	0.93
$\beta_{1,2}$	0	0.00 (0.00 – 0.41)	0.06	0.14 (0.01 – 0.81)	0.06	0.12 (0.01 – 0.54)	0.18
$\beta_{1,3}$	1	0.27 (0.00 – 1.94)	0.63	0.51 (0.05 – 2.43)	0.39	0.68 (0.13 – 1.19)	0.92
$\beta_{1,4}$	1	0.77 (0.00 – 2.25)	0.78	0.81 (0.10 – 2.89)	0.62	0.84 (0.24 – 1.24)	0.95
$\beta_{1,5}$	1	0.20 (0.00 – 1.32)	0.63	0.45 (0.02 – 1.66)	0.44	0.60 (0.04 – 0.98)	0.88
$\beta_{2,3}$	0	0.00 (0.00 – 0.55)	0.13	0.04 (0.00 – 0.73)	0.03	0.04 (0.00 – 0.24)	0.02
$\beta_{2,4}$	0	0.00 (0.00 – 0.35)	0.06	0.10 (0.00 – 0.63)	0.04	0.08 (0.00 – 0.46)	0.09
$\beta_{2,5}$	0	0.00 (0.00 – 0.00)	0.02	0.04 (0.00 – 0.40)	0.01	0.04 (0.00 – 0.28)	0.03
$\beta_{3,4}$	1	0.21 (0.00 – 1.37)	0.61	0.30 (0.02 – 1.31)	0.28	0.53 (0.15 – 1.15)	0.91
$\beta_{3,5}$	0	0.00 (0.00 – 0.62)	0.11	0.11 (0.01 – 0.71)	0.04	0.28 (0.01 – 0.92)	0.48
$\beta_{4,5}$	0	0.00 (0.00 – 0.89)	0.15	0.14 (0.01 – 0.94)	0.08	0.42 (0.03 – 1.02)	0.65

Table 4.5: Simulation scenario 8 results summary comparing LASSO, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		**BBVS**		**BHIS**	
	true	**Median**	**Incl Prop**	**Median**		**Median**	
Coeff.	**value**	**Estimate**	**Sel flag**	**Estimate**	**Incl Prop**	**Estimate**	**Incl Prop**
$\beta_{1,2,3}$	0	0.00 (0.00 – 0.00)	0.02	0.00 (-0.00 – 0.04)	0.00	0.02 (0.00 – 0.17)	0.00
$\beta_{1,2,4}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.00 – 0.07)	0.00	0.05 (0.00 – 0.33)	0.03
$\beta_{1,2,5}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.00 – 0.02)	0.00	0.02 (0.00 – 0.19)	0.00
$\beta_{1,3,4}$	1	0.00 (0.00 – 0.94)	0.19	0.00 (-0.00 – 0.20)	0.00	0.41 (0.05 – 1.12)	0.72
$\beta_{1,3,5}$	0	0.00 (0.00 – 0.30)	0.04	0.00 (-0.00 – 0.05)	0.00	0.17 (0.01 – 0.74)	0.29
$\beta_{1,4,5}$	0	0.00 (0.00 – 0.52)	0.12	0.00 (-0.00 – 0.05)	0.00	0.27 (0.01 – 0.72)	0.50
$\beta_{2,3,4}$	0	0.00 (0.00 – 0.77)	0.26	0.00 (-0.00 – 0.04)	0.00	0.00 (-0.00 – 0.12)	0.00
$\beta_{2,3,5}$	0	0.00 (0.00 – 0.00)	0.02	0.00 (-0.00 – 0.02)	0.00	0.00 (-0.00 – 0.05)	0.00
$\beta_{2,4,5}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.00 – 0.02)	0.00	0.01 (-0.00 – 0.11)	0.00
$\beta_{3,4,5}$	0	0.00 (0.00 – 0.00)	0.01	0.00 (-0.00 – 0.05)	0.00	0.04 (0.00 – 0.58)	0.07

4.4 Application

The objective of the application was to compare the performance of the BHIS to BBVS and showcase the pros ad cons of each. The BBVS is slightly more flexible in allowing specific terms to drop out while the

BHIS allows for hierarchical interaction selection at the expense of more terms in the model. The modeling to be presented was motivated by an assessment of the association between prenatal urinary phthalate metabolite concentrations and childhood fat mass in a New York City cohort.

4.4.1 Cohort Analysis Sample

The Mount Sinai Children's Environmental Health Study is a prospective study which followed mother-infant pairs at Mount Sinai Hospital and two adjacent facilities in New York City. The study enrolled 479 primiparous women who delivered at Mount Sinai Medical Center between 1998 and 2002. Mother-infant pairs in the cohort were invited to three follow-up visits at approximately ages 4–5.5, 6, and 7–9 years. Maternal and infant demographic and lifestyle characteristics were obtained via questionnaires administered during the third trimester and at follow-up visits. The cohort analysis sample consists of 144 infants from the Mount Sinai Children's Environmental Health Study with prenatal phthalate metabolite concentrations measured in maternal urine and at least one outcome measurement at follow-up.

4.4.2 Phthalate Exposures

Spot maternal urine samples were analyzed for the following phthalate metabolites: monoethyl phthalate (MEP), mono-n-butyl phthalate (MnBP), monoisobutyl phthalate (MiBP), mono(3-carboxypropyl) phthalate (MCPP), monobenzyl phthalate (MBzP), and four metabolites of DEHP: mono(2-ethylhexyl) phthalate (MEHP), mono(2-ethyl-5-hydroxyhexyl) phthalate (MEHHP), mono(2-ethyl-5-oxohexyl) phthalate (MEOHP), and mono(2-ethyl-5-carboxypentyl) phthalate (MECPP) between 25 and 40 weeks' gestation. Due to the high correlation among the four metabolites of DEHP, the molar sum of the metabolite concentrations was used in all analyses.

4.4.3 Outcome

Percent fat mass defined as (fat mass / weight) $\times 100$ was the outcome of interest. Weight and body composition were measured in children at each follow-up visit using a pediatric Tanita scale (model TBF-300; Tanita Corporation of America).

4.4.4 Covariates

All modeling included adjustment for confounding variables similar to those previously identified via directed acyclic diagrams (Buckley et al., 2015). Covariates included: maternal race/ethnicity (non-Hispanic

white/non-Hispanic black/Hispanic), age (years), education (less than college/college degree or more), employment status during pregnancy (employed/unemployed), prepregnancy BMI, gestational weight gain, smoking during pregnancy (yes/no), breastfeeding (ever/never), year of urine collection, natural log creatinine, and child physical activity at follow-up (active/inactive). Child sex and age in months at follow-up were also covariates included, however, their inclusion in models differed by method implemented. All continuous variables were standardized by the sample mean and two times the standard deviation (Gelman et al., 2013).

4.4.5 Modeling Specifics

The methodology presented in section 4.2 was adapted to account for multiple observations per child by including random-intercepts using a Bayesian hierarchical model. The subject specific intercepts were assumed i.i.d from a normal distribution with a population intercepts. To account for uncertainty in the population intercept, a conjugate normal-inverse gamma hyper prior was assumed on the population intercept and variance.

The associations of maternal urinary phthalate metabolite concentrations with fat mass were assessed with hierarchical models including BBVS and BHIS, separately. Following similar to previous analyses, each model was adjusted for factors listed in (4.4.4) which included cubic maternal age and weight gain; and quadratic prepregnancy BMI and log creatinine polynomials (Buckley et al., 2015). Two modeling strategies were considered. Model 1 included variable selection on all pairwise interactions among phthalate concentrations, and interactions between child sex and phthalates. Model 2 included variable selection on all pairwise interaction among phthalate concentrations, child sex, and child age. The BBVS was implemented using models 1 and 2, while BHIS only using model 2. Variables not considered for selection were included in the model by using a normal prior on the these regression coefficients (i.e. not assumed DP; forced into the model). For instance, model 1 includes an interaction term between child sex and age at follow-up and their main effect terms assuming the corresponding regression coefficients have a normal prior. Additionally, the BBVS modeling includes separate variable selection blocks for the main and interaction effects, that is, there is a distinct probability of inclusion parameter for the interaction and main effects. The BHIS similarly has a blocking scheme albeit on the re-parametrized interaction components.

Sensitivity analyses were conducted to assess the effect of using informative priors on the probability of exclusion (i.e. β=0) of main effect terms. The informative priors were specified to favor inclusion of main effect terms in the model by specifying $\pi_{01} \sim \text{beta}(0.001, 1)$, that is, our prior belief is that probability of main effects are excluded from the model is on average approximately 0.001.

The MCMC for each model was run on five separate chains each run for 50,000 iterations after discarding the first 10,000 as a burn-in and stored every 50^{th} iterate. Potential scale reduction factors and trace plots were used to assess model convergence (Gelman et al., 2013). Weakly informative prior and hyperprior specifications were used. For instance, consider a normal-inverse gamma prior with parameters μ, τ, a, b then hyper parameters can be set as: $\mu = 0$ and $\tau = a = b = 0.001$.

4.4.6 Results

The analysis presented here consists of complete case analysis discarding missing observations in the percent fat mass outcome and confounding variables. There were 144 mother-child pairs with complete case data in the analysis cohort and total $n = 285$ observations.

In general, convergence diagnostics among all regression coefficients were satisfactory using the potential scale reduction factor score (PSRF) below 1.10, with the exception a few coefficients in two different models. In BBVS model 2, convergence on the MiBP:MCPP interaction term coefficient was less than optimal with an observed PSRF of 1.12 while four regression coefficients were less than optimal in the BHIS model (MiBP: 1.20; Age (months): 1.16; MiBP:Age (months): 1.24; Sex (female):Age (months): 1.12). All PSRF factors for models with informative priors on probability of coefficient exclusion were less than 1.10.

Across all scenarios there was insufficient evidence to support the inclusion of any pairwise phthalate interaction nor main effect terms in model since all corresponding coefficient probabilities of inclusion were less than 0.5 using weakly informative priors. However, both variable selection procedures in model 2 supported inclusion of child sex by age interactions with estimate posterior probabilities of inclusion at 1.00 and 0.96 for BBVS (mean: 6.11: 80% CI 4.76, 7.83) and BHIS (mean: 3.17; 80% CI 1.29, 4.05), respectively. By construction of the BHIS, the inclusion of the sex by age interaction implies inclusion of the corresponding main effects which is observed in the corresponding estimated posterior inclusion probabilities of 0.96 for sex (mean: 2.58; 80% CI 0.00, 3.95) and 1.00 for age (mean: 3.55; 80% CI 2.67, 4.09). For a summary of regression coefficient variable selection, see Table 4.6.

Table 4.6: Summary of variable selection regression coefficients displaying posterior mean estimate (80% Cred. Int.), probability of inclusion, and potential scale reduction factor for models with weakly informative priors on block probability of exclusion by selection and model type.

	BBVS						BHIS		
	Model 1			Model 2			Model 2		
Term	**Estimate**	**Pr. Incl.**	**PSRF**	**Estimate**	**Pr. Incl.**	**PSRF**	**Estimate**	**Pr. Incl.**	**PSRF**
MnBP	-0.00 (0.00, 0.00)	0.0012	1.00	0.00 (0.00, 0.00)	0.0096	1.02	0.00 (0.00, 0.00)	0.0130	1.01
MiBP	0.00 (0.00, 0.00)	0.0012	1.00	0.12 (0.00, 0.00)	0.0700	1.04	0.19 (0.00, 1.23)	0.1596	1.20
MEP	0.00 (0.00, 0.00)	0.0008	1.00	0.01 (0.00, 0.00)	0.0116	1.03	0.00 (0.00, 0.00)	0.0078	1.00
MBzP	-0.00 (0.00, 0.00)	0.0004	1.00	0.06 (0.00, 0.00)	0.0414	1.04	0.01 (0.00, 0.00)	0.0224	1.01
MCPP	-0.00 (0.00, 0.00)	0.0006	1.00	0.11 (0.00, 0.00)	0.0694	1.07	0.07 (0.00, 0.00)	0.0632	1.08
Sex (F)				0.01 (0.00, 0.00)	0.0144	1.00	2.58 (0.00, 3.95)	0.9572	1.10
Age (mos.)				0.46 (0.00, 2.09)	0.2258	1.09	3.55 (2.67, 4.09)	1.0000	1.16
ΣDEHP	-0.00 (0.00, 0.00)	0.0020	1.00	0.00 (0.00, 0.00)	0.0062	1.00	0.00 (0.00, 0.00)	0.0062	1.00
MnBP:MiBP	0.00 (0.00, 0.00)	0.0002	1.00	0.01 (0.00, 0.00)	0.0320	1.02	0.00 (0.00, 0.00)	0.0036	1.01
MnBP:MEP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0024	1.00	0.00 (0.00, 0.00)	0.0006	1.00
MiBP:MEP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0088	1.01	0.00 (0.00, 0.00)	0.0012	1.00
MnBP:MBzP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0086	1.02	0.00 (0.00, 0.00)	0.0010	1.00
MiBP:MBzP	0.00 (0.00, 0.00)	0.0000		0.03 (0.00, 0.00)	0.0708	1.03	0.00 (0.00, 0.00)	0.0010	1.00
MEP:MBzP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0128	1.04	0.00 (0.00, 0.00)	0.0008	1.00
MnBP:MCPP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0162	1.01	0.00 (0.00, 0.00)	0.0014	1.00
MiBP:MCPP	0.00 (0.00, 0.00)	0.0000		0.17 (0.00, 0.69)	0.2658	1.12	0.00 (0.00, 0.00)	0.0022	1.01
MEP:MCPP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0110	1.00	0.00 (0.00, 0.00)	0.0004	1.00
MBzP:MCPP	0.00 (0.00, 0.00)	0.0002	1.00	0.06 (0.00, 0.33)	0.1306	1.03	0.00 (0.00, 0.00)	0.0012	1.00
MnBP:Sex (F)	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0210	1.00	-0.00 (0.00, 0.00)	0.0024	1.01
MiBP:Sex (F)	0.00 (0.00, 0.00)	0.0012	1.00	0.06 (0.00, 0.16)	0.1062	1.01	0.01 (0.00, 0.00)	0.0122	1.06
MEP:Sex (F)	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0118	1.00	-0.00 (0.00, 0.00)	0.0012	1.00
MBzP:Sex (F)	0.00 (0.00, 0.00)	0.0000		0.01 (0.00, 0.00)	0.0322	1.00	0.00 (0.00, 0.00)	0.0022	1.00
MCPP:Sex (F)	0.00 (0.00, 0.00)	0.0006	1.00	0.06 (0.00, 0.00)	0.0908	1.00	0.00 (0.00, 0.00)	0.0022	1.00
MnBP:Age (mos.)				0.04 (0.00, 0.00)	0.0970	1.01	0.00 (0.00, 0.00)	0.0030	1.00
MiBP:Age (mos.)				0.10 (0.00, 0.40)	0.1578	1.01	0.17 (0.00, 1.20)	0.1354	1.24
MEP:Age (mos.)				0.04 (0.00, 0.00)	0.1020	1.01	0.00 (0.00, 0.00)	0.0008	1.00
MBzP:Age (mos.)				0.14 (0.00, 0.72)	0.2136	1.02	0.01 (0.00, 0.00)	0.0130	1.02
MCPP:Age (mos.)				0.06 (0.00, 0.29)	0.1294	1.00	0.07 (0.00, 0.00)	0.0514	1.08
Sex (F):Age (mos.)				6.11 (4.76, 7.83)	1.0000	1.02	3.17 (1.29, 4.05)	0.9572	1.12
MnBP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		-0.00 (0.00, 0.00)	0.0032	1.00	-0.00 (0.00, 0.00)	0.0010	1.00
MiBP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0060	1.01	0.00 (0.00, 0.00)	0.0006	1.00
MEP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		-0.00 (0.00, 0.00)	0.0024	1.00	-0.00 (0.00, 0.00)	0.0004	1.00
MBzP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0070	1.00	-0.00 (0.00, 0.00)	0.0008	1.00
MCPP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0082	1.00	0.00 (0.00, 0.00)	0.0010	1.00
Sex (F):ΣDEHP	0.00 (0.00, 0.00)	0.0000		-0.00 (0.00, 0.00)	0.0104	1.00	-0.00 (0.00, 0.00)	0.0006	1.00
Age (mos.):ΣDEHP				0.17 (0.00, 0.61)	0.3244	1.02	-0.00 (0.00, 0.00)	0.0010	1.00

The analysis with informative priors on the exclusion probability of main effect regression coefficients yielded similar results. All posterior probabilities of inclusion on pairwise interactions including phthalate metabolite concentrations were estimated to be less than 0.14, and the child sex by age interaction had probability of inclusion of 1.00 in both the BBVS and BHIS model 2 models. In different results, there was not sufficient evidence to influence excluding main effects from each model since all estimated posterior probabilities of exclusion were greater than 0.50 in the presence of informative exclusion probability priors (see Table 4.7). In BBVS model 1, phthalate main effects varied from -0.05 to -0.03 and all corresponding 80% credible intervals straddled zero (MnBP: -0.04 (-0.20, 0.10); MiBP: -0.03 (-0.20, 0.11); MEP: -0.03 (-0.20, 0.10); MBzP: -0.03 (-0.19, 0.11); MCPP: -0.03 (-0.20, 0.11); ΣDEHP: -0.05 (-0.21, 0.09)).

By favoring the inclusion of main effects in BBVS and BHIS model 2 models, the magnitude of the posterior mean effects increased in phthalate metabolite coefficients, child sex and age. For example, the BBVS model 2 observed the greatest magnitude increase in the posterior mean effect estimate of age increasing from 0.46 (80% CI: 0.00, 2.09) to 0.71 (80% CI: 0.00, 2.26) in the weakly informative and informative priors, respectively. The BHIS observe more modest increases in the phthalate concentration effects with MBzP observing the greatest increase from 0.01 (80% CI: 0.00, 0.00) to 0.14 (80% CI: 0.00, 0.35), while effects for child sex and age remained similar.

Table 4.7: Summary of variable selection regression coefficients displaying posterior mean estimate (80% Cred. Int.), probability of inclusion, and potential scale reduction factor for models with informative priors on main effects favoring inclusion on block probability of exclusion by selection and model type.

	BBVS						BHIS		
	Model 1			Model 2			Model 2		
Term	**Estimate**	**Pr. Incl.**	**PSRF**	**Estimate**	**Pr. Incl.**	**PSRF**	**Estimate**	**Pr. Incl.**	**PSRF**
MnBP	-0.04 (-0.20, 0.10)	0.6436	1.01	0.16 (-0.11, 0.43)	0.8496	1.01	0.09 (0.00, 0.32)	0.6310	1.03
MiBP	-0.03 (-0.20, 0.11)	0.6430	1.01	0.29 (0.00, 0.55)	0.8638	1.01	0.21 (0.00, 0.52)	0.6758	1.03
MEP	-0.03 (-0.20, 0.10)	0.6430	1.01	0.21 (-0.03, 0.45)	0.8506	1.03	0.11 (0.00, 0.32)	0.6308	1.05
MBzP	-0.03 (-0.19, 0.11)	0.6428	1.01	0.26 (0.00, 0.49)	0.8514	1.03	0.14 (0.00, 0.35)	0.6444	1.03
MCPP	-0.03 (-0.20, 0.11)	0.6432	1.01	0.25 (-0.00, 0.50)	0.8542	1.02	0.15 (0.00, 0.37)	0.6552	1.02
Sex (F)				0.17 (-0.12, 0.44)	0.8494	1.01	2.74 (-0.14, 3.98)	1.0000	1.04
Age (mos.)				0.71 (0.00, 2.26)	0.8742	1.02	3.59 (3.01, 4.07)	1.0000	1.01
ΣDEHP	-0.05 (-0.21, 0.09)	0.6434	1.02	0.14 (-0.15, 0.42)	0.8492	1.01	0.06 (-0.04, 0.29)	0.6342	1.03
MnBP:MiBP	0.00 (0.00, 0.00)	0.0004	1.00	0.00 (0.00, 0.00)	0.0124	1.01	0.00 (0.00, 0.00)	0.0072	1.03
MnBP:MEP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0030	1.01	0.00 (0.00, 0.00)	0.0068	1.03
MiBP:MEP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0038	1.01	0.00 (0.00, 0.00)	0.0048	1.00
MnBP:MBzP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0100	1.04	0.00 (0.00, 0.00)	0.0070	1.03
MiBP:MBzP	-0.00 (0.00, 0.00)	0.0002	1.00	0.01 (0.00, 0.00)	0.0210	1.02	0.00 (0.00, 0.00)	0.0124	1.03
MEP:MBzP	-0.00 (0.00, 0.00)	0.0002	1.00	0.00 (0.00, 0.00)	0.0016	1.00	0.00 (0.00, 0.00)	0.0082	1.03
MnBP:MCPP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0062	1.01	0.00 (0.00, 0.00)	0.0068	1.03
MiBP:MCPP	0.00 (0.00, 0.00)	0.0000		0.04 (0.00, 0.00)	0.0652	1.01	0.00 (0.00, 0.00)	0.0086	1.00
MEP:MCPP	0.00 (0.00, 0.00)	0.0002	1.00	0.00 (0.00, 0.00)	0.0034	1.00	0.00 (0.00, 0.00)	0.0062	1.03
MBzP:MCPP	0.00 (0.00, 0.00)	0.0004	1.00	0.02 (0.00, 0.00)	0.0402	1.03	0.00 (0.00, 0.00)	0.0100	1.01
MnBP:Sex (F)	0.00 (0.00, 0.00)	0.0002	1.00	0.00 (0.00, 0.00)	0.0074	1.00	-0.00 (0.00, 0.00)	0.0060	1.01
MiBP:Sex (F)	0.01 (0.00, 0.00)	0.0070	1.01	0.03 (0.00, 0.00)	0.0402	1.01	0.01 (0.00, 0.00)	0.0170	1.05
MEP:Sex (F)	0.00 (0.00, 0.00)	0.0002	1.00	-0.00 (0.00, 0.00)	0.0026	1.00	-0.00 (0.00, 0.00)	0.0058	1.00
MBzP:Sex (F)	0.00 (0.00, 0.00)	0.0004	1.00	0.00 (0.00, 0.00)	0.0102	1.00	0.00 (0.00, 0.00)	0.0062	1.00
MCPP:Sex (F)	0.00 (0.00, 0.00)	0.0008	1.00	0.02 (0.00, 0.00)	0.0366	1.00	0.00 (0.00, 0.00)	0.0072	1.01
MnBP:Age (mos.)				0.02 (0.00, 0.00)	0.0336	1.00	0.00 (0.00, 0.00)	0.0056	1.01
MiBP:Age (mos.)				0.10 (0.00, 0.25)	0.1056	1.01	0.07 (0.00, 0.00)	0.0682	1.02
MEP:Age (mos.)				0.01 (0.00, 0.00)	0.0340	1.01	0.00 (0.00, 0.00)	0.0060	1.00
MBzP:Age (mos.)				0.07 (0.00, 0.00)	0.0974	1.02	0.01 (0.00, 0.00)	0.0180	1.01
MCPP:Age (mos.)				0.04 (0.00, 0.00)	0.0566	1.00	0.03 (0.00, 0.00)	0.0338	1.01
Sex (F):Age (mos.)				6.73 (4.99, 8.45)	1.0000	1.01	3.46 (2.92, 4.02)	1.0000	1.01
MnBP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		-0.00 (0.00, 0.00)	0.0032	1.00	0.00 (0.00, 0.00)	0.0034	1.00
MiBP:ΣDEHP	-0.00 (0.00, 0.00)	0.0002	1.00	0.00 (0.00, 0.00)	0.0030	1.01	0.00 (0.00, 0.00)	0.0066	1.03
MEP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		0.00 (0.00, 0.00)	0.0008	1.00	0.00 (0.00, 0.00)	0.0074	1.03
MBzP:ΣDEHP	0.00 (0.00, 0.00)	0.0000		-0.00 (0.00, 0.00)	0.0008	1.00	0.00 (0.00, 0.00)	0.0076	1.02
MCPP:ΣDEHP	0.00 (0.00, 0.00)	0.0006	1.00	0.00 (0.00, 0.00)	0.0034	1.01	0.00 (0.00, 0.00)	0.0072	1.01
Sex (F):ΣDEHP	-0.00 (0.00, 0.00)	0.0002	1.00	-0.00 (0.00, 0.00)	0.0054	1.00	-0.02 (0.00, 0.00)	0.0318	1.06
Age (mos.):ΣDEHP				0.07 (0.00, 0.41)	0.1336	1.01	0.00 (0.00, 0.00)	0.0062	1.01

In terms of confounding variables, there was great consistency across priors and variable selection procedures. These findings were not surprising since varying the prior on the selection probability should have minimal effect on non-selection variables. This is observed by noting that effect estimates for the confounding variables vary minimally across prior selection specification within each selection and model type (see Tables 4.8 & 4.9). The coefficients for BBVS and BHIS in model 2, were also similar across variable selection type even though these observed different variable selection results by exclusion probability priors. The magnitude of effect estimates for common coefficients between models 1 and 2 were different, with model 1 coefficients observing greater magnitudes in general with the exception of the cubic maternal age terms.

Table 4.8: Summary of confounder regression coefficients displaying posterior mean estimate (80% Cred. Int.), probability of inclusion, and potential scale reduction factor for models with weakly informative priors on block probability of exclusion by selection and model type.

	BBVS				BHIS	
	Model 1		Model 2		Model 2	
Term	**Estimate**	**PSRF**	**Estimate**	**PSRF**	**Estimate**	**PSRF**
Intercept (Pop.)	1.19 (-11.15, 13.90)	1.02	13.98 (8.48, 19.39)	1.01	13.52 (7.99, 19.05)	1.01
Creatinine	3.03 (1.56, 4.47)	1.00	0.10 (-0.05, 0.19)	1.00	0.11 (-0.05, 0.23)	1.01
Creatinine^2	2.13 (0.53, 3.79)	1.00	0.07 (-0.05, 0.16)	1.00	0.08 (-0.06, 0.20)	1.00
Maternal age	1.54 (0.13, 2.92)	1.01	0.07 (-0.05, 0.17)	1.01	0.08 (-0.05, 0.19)	1.01
Maternal age^2	-0.05 (-0.11, 0.01)	1.01	0.01 (-0.01, 0.03)	1.00	0.01 (-0.01, 0.03)	1.00
Maternal age^3	0.00 (-0.00, 0.00)	1.01	-0.00 (-0.00, 0.00)	1.00	-0.00 (-0.00, 0.00)	1.00
Race (black)	1.06 (-0.83, 2.93)	1.00	0.05 (-0.06, 0.14)	1.00	0.06 (-0.06, 0.17)	1.00
Race (hispanic/other)	2.55 (0.77, 4.41)	1.00	0.08 (-0.05, 0.16)	1.00	0.09 (-0.05, 0.19)	1.00
Sex (F)	-0.80 (-2.14, 0.52)	1.00				
Sex (F):Age (mos.)	5.23 (4.20, 6.29)	1.00				
Maternal smoking	2.37 (0.73, 4.00)	1.00	0.09 (-0.05, 0.19)	1.00	0.09 (-0.05, 0.22)	1.01
Year urine collection	0.60 (-0.74, 1.96)	1.00	0.03 (-0.06, 0.13)	1.00	0.05 (-0.06, 0.16)	1.00
Age (mos.)	2.74 (2.02, 3.47)	1.00				
Breastfed (ever)	-1.23 (-2.82, 0.32)	1.00	-0.01 (-0.08, 0.10)	1.00	0.00 (-0.08, 0.12)	1.00
Education ($\geq$college)	-0.06 (-1.97, 1.83)	1.00	0.02 (-0.06, 0.11)	1.00	0.02 (-0.07, 0.12)	1.00
Mother unemployed	0.59 (-1.10, 2.31)	1.00	0.04 (-0.06, 0.14)	1.00	0.05 (-0.06, 0.18)	1.00
Phys. activity (active)	0.52 (-0.09, 1.14)	1.00	0.05 (-0.05, 0.17)	1.00	0.06 (-0.05, 0.20)	1.00
Maternal BMI	3.59 (1.86, 5.42)	1.00	0.10 (-0.05, 0.18)	1.00	0.11 (-0.05, 0.23)	1.01
Maternal BMI^2	-2.16 (-3.53, -0.79)	1.00	-0.00 (-0.07, 0.11)	1.00	0.00 (-0.07, 0.13)	1.00
Gest. weight gain	1.33 (-0.41, 3.02)	1.00	0.06 (-0.05, 0.17)	1.00	0.07 (-0.05, 0.19)	1.00
Gest. weight gain^2	0.09 (-1.95, 2.09)	1.00	0.04 (-0.06, 0.14)	1.00	0.04 (-0.06, 0.16)	1.00
Gest. weight gain^3	-0.80 (-2.57, 0.89)	1.00	0.03 (-0.06, 0.13)	1.00	0.04 (-0.06, 0.17)	1.00

Additionally, the population intercepts differed greatly by model type. The population intercept for model 1 was observed to be much lower than (Weak: 1.19; Inf. ME: 1.58) those observed in BBVS (Weak: 13.98; Inf. ME: 11.20) and BHIS (Weak: 13.52; Inf. ME: 12.23) model 2 models, and the corresponding 80% credible intervals straddled zero (Weak: -11.15, 13.90; Inf. ME: -11.38, 14.62). This may be an artifact of the larger effects of confounding variables observed in model 1, and it is worrisome that the range of percent fat mass estimates can be negative. The population intercept 80% credible intervals for model 2 models, on the other

hand, all spanned sensible ranges (BBVS Weak: 8.48, 19.39/Inf. ME: 5.11, 17.69; BHIS Weak: 7.99, 19.05/Inf ME: 6.32, 18.13).

Table 4.9: Summary of confounder regression coefficients displaying posterior mean estimate (80% Cred. Int.), probability of inclusion, and potential scale reduction factor for models with informative priors on main effects favoring inclusion on block probability of exclusion by selection and model type.

	BBVS				BHIS	
	Model 1		Model 2		Model 2	
Term	**Estimate**	**PSRF**	**Estimate**	**PSRF**	**Estimate**	**PSRF**
Intercept (Pop.)	1.58 (-11.38, 14.62)	1.01	11.20 (5.11, 17.69)	1.01	12.23 (6.32, 18.13)	1.02
Creatinine	3.22 (1.48, 5.02)	1.01	0.07 (-0.05, 0.14)	1.03	0.07 (-0.05, 0.15)	1.01
Creatinine^2	2.19 (0.50, 3.88)	1.00	0.05 (-0.06, 0.13)	1.01	0.06 (-0.05, 0.14)	1.01
Maternal age	1.55 (0.07, 3.06)	1.01	0.06 (-0.05, 0.13)	1.02	0.05 (-0.06, 0.13)	1.01
Maternal age^2	-0.05 (-0.11, 0.01)	1.00	0.01 (-0.01, 0.03)	1.00	0.01 (-0.01, 0.03)	1.00
Maternal age^3	0.00 (-0.00, 0.00)	1.00	-0.00 (-0.00, 0.00)	1.00	-0.00 (-0.00, 0.00)	1.00
Race (black)	1.17 (-0.76, 3.11)	1.00	0.04 (-0.06, 0.12)	1.01	0.04 (-0.06, 0.13)	1.01
Race (hispanic/other)	2.62 (0.80, 4.50)	1.00	0.06 (-0.05, 0.14)	1.02	0.06 (-0.06, 0.14)	1.01
Sex (F)	-0.80 (-2.17, 0.55)	1.00				
Sex (F):Age (mos.)	5.24 (4.19, 6.34)	1.00				
Maternal smoking	2.46 (0.81, 4.10)	1.00	0.07 (-0.05, 0.14)	1.02	0.06 (-0.05, 0.15)	1.01
Year urine collection	0.66 (-0.68, 1.96)	1.00	0.02 (-0.06, 0.12)	1.00	0.03 (-0.06, 0.12)	1.00
Age (mos.)	2.74 (1.98, 3.49)	1.00				
Breastfed (ever)	-1.22 (-2.81, 0.31)	1.00	-0.01 (-0.08, 0.09)	1.00	0.00 (-0.07, 0.10)	1.00
Education ($\geq$college)	-0.13 (-2.09, 1.82)	1.00	0.02 (-0.07, 0.10)	1.00	0.02 (-0.07, 0.11)	1.00
Mother unemployed	0.62 (-1.07, 2.32)	1.00	0.03 (-0.06, 0.12)	1.00	0.04 (-0.06, 0.12)	1.00
Phys. activity (active)	0.52 (-0.07, 1.14)	1.00	0.05 (-0.05, 0.14)	1.01	0.04 (-0.05, 0.14)	1.01
Maternal BMI	3.68 (1.85, 5.62)	1.00	0.08 (-0.05, 0.15)	1.02	0.07 (-0.05, 0.16)	1.01
Maternal BMI^2	-2.21 (-3.63, -0.82)	1.00	-0.00 (-0.07, 0.10)	1.01	0.01 (-0.07, 0.11)	1.00
Gest. weight gain	1.28 (-0.44, 3.02)	1.00	0.05 (-0.05, 0.14)	1.01	0.05 (-0.05, 0.15)	1.01
Gest. weight gain^2	0.11 (-1.93, 2.14)	1.00	0.03 (-0.06, 0.12)	1.00	0.03 (-0.06, 0.12)	1.00
Gest. weight gain^3	-0.80 (-2.59, 1.01)	1.00	0.02 (-0.06, 0.12)	1.00	0.03 (-0.06, 0.13)	1.00

The findings we observed were similar to previous findings. The credible interval estimates straddling zero in model 1 estimates, was also observed in phthalate metabolize concentration effects by Buckley et al. (2015), albeit their modeling strategy did not include variable selection in their analyses, and it included effect modification by child sex. In either case, their 80% credible intervals for overall concentration effects also included zero. However, the estimates we observed seemed to be highly influenced by shrinkage relative to their estimates. Also, our modeling although influenced by their analysis was not an exact replication. Their analyses included a larger list of confounding variables and accounted for missing covariate observations and imputation for concentrations below limits of detections (LOD). Our analysis may be limited and influenced by the reduced sample size as a complete case analysis and use of LOD imputation for concentrations below LOD. However, the similarities with our model 1 and previous analyses is encouraging.

Our decision to compare associations by using an informative prior favoring inclusion of main effects was guided by previous findings. A recent study assessing simultaneous associations between urinary phthalate concentrations and childhood BMI and obesity, used variable selection on phthalate concentrations through

Bayesian Kernel Machine Regression (BKMR), and found mostly linear positive associations with MEP, MnBP, MBzP, and ΣDEHP with BMI z-scores in children at age 12 while MCPP observed a negative relationship (Harley et al., 2017). Our observations of positive simultaneous associations between all phthalate metabolite concentrations encouragingly coincides with their findings except for MCPP.

4.5 Discussion

In this chapter, we have studied a Bayesian semi-parametric modeling framework with variable selection for hierarchical interactions. We have found that the proposed method performs as well as LASSO in most simulation scenarios and found its flexibility by design is advantageous in screening for significant variables in three-way interaction models with highly correlated predictors while adjusting for mixed-scale confounding variables. This modeling framework is useful for epidemiologic research as it allows simultaneous estimation, model averaging, variable selection, effect clustering, automatically accounts for hierarchical interactions; and preserves grouping, incorporation of prior information on block specific exclusion probability, and effect magnitude as in Herring (2010). We have additionally developed a Matlab package that can implement Bayesian hierarchical semi-parametric with blocked (i.e. grouped) variable selection which can account for repeated measures (i.e. as analogous random intercept linear mixed models) and can accommodate hierarchical interactions. This Matlab package will also have functionality to account for variables subject to observations below limits of quantification by leveraging its Bayesian nature.

The methodology proposed in this chapter is computationally burdened in the presence of increasing number of predictors and greater degree of n-way interactions. LASSO and its variants are a great resource for their scalability to screening variables in the high dimensional setting with feasible computational time (Tibshirani, 1996; Lim and Hastie, 2015). BHIS estimation, however, relies on an MCMC sampler where it may advantageous to consider alternative estimation procedures for it.

CHAPTER 5: A JOINT MIXTURE MODEL FOR COMPOSITIONAL DATA WITH ESSENTIAL ZEROS: PROFILES OF PHYSICAL ACTIVITY AND HEALTH RISK

5.1 Introduction

Jointly modeling time spent in sedentary behavior and physical activity has been a difficult problem to approach because of the compositional nature of the data, that is the sum of time is constrained by the observed time interval (e.g. 24 hours in a day, wake time, etc.), especially when the behaviors are multicomponent (e.g. intensity levels or by type of activity). Accelerometry-assessed sedentary behavior and physical activity are often reported as a four-part component of time (e.g., a day of wear time) spent in behaviors of incremental intensity: sedentary, light, moderate, and vigorous activities. This implies that the variables are constrained to the wear time and thus it is important to ensure these data are analyzed as a composition. Latent class (also known as latent profile) analysis and clustering techniques have been used for patterning, but each have been implemented with some shortcomings (Carson et al., 2015; Evenson et al., 2015, 2016; Huh et al., 2011; Liu et al., 2010; Patnode et al., 2011).

Latent class analysis for assessing sedentary behavior and physical activity include separate analyses on each component or joint analyses ignoring the compositional nature of the data (Carson et al., 2015; Evenson et al., 2015, 2016; Huh et al., 2011; Liu et al., 2010; Patnode et al., 2011). Separate analyses often have the advantage of simplicity. However, current approaches do not make the most efficient use of the data and do not yield information about how all the components interrelate. Joint analyses, ignoring the compositional nature of the data may yield inconsistent results (Aitchison, 1986; Dumuid et al., 2017b; Gupta et al., 2018; Leech et al., 2014; Fernández et al., 2015). Furthermore, it has been shown that when summarizing a sample of compositions with arithmetic means, these over-estimate proportions for non-sedentary times (Chastin et al., 2015). For compositional data, the normalized geometric means of each component are more appropriate measures of central tendency (Aitchison, 2005; Pawlowsky-Glahn and Egozcue, 2002).

Compositional data analysis is an area of active research, and with recent developments, researchers have begun applying these methods to analyzing 24-hour time budgets of sleep, sedentary, and physical activity behaviors. Current approaches have relied on implementing appropriate transformations (i.e. isometric log ratio transformations) of the data that achieve desirable properties of linear-like behaviors and have applied k-means clustering for patterning of behaviors (Dumuid et al., 2017a). These transformations preserve good

mathematical properties in terms of distances between compositions, but they rely on log-ratios of components, which are not directly usable in the presence of zeros (Martın-Fernandez et al., 2011).

Many people do not engage in vigorous activity, which introduces the essential zero difficulty into compositional data. Essential zeros are defined as true absence of a composition element, which is different from rounded zeros, where levels of a composition are below limits of detection. Current methods for accounting for zeros in compositional data include stratification, data imputation, or conditional distribution methods (Kaul et al., 2017b,a; Martın-Fernandez et al., 2011). In physical activity, imputation is not a tenable approach, as it would impute data where data does not exist. On the other hand, stratification, defined as separate analyses by zero pattern configuration, does not make the most effective use of the data and complicates interpretation. Conditional distribution methods model the distribution of a composition (e.g. using a multivariate Gaussian) conditional on a zero patter configuration and can discern group differences via discriminant analysis (Kaul et al., 2017a). However, it may be difficult to discover profiles present in the data given that it requires prior knowledge of how many subclasses exist. Thus, there is currently a need for latent class methods which can account for essential zeros in compositional data. Additionally, essential zeros pose a unique difficulty in estimation of some measures of central tendency because zero values imply zero summaries (i.e. zeros in geometric means).

A general form of the latent class model for compositional data has been previously defined; however, it does not incorporate essential zeros (Comas-Cufí et al., 2016). In this paper, we will develop a new latent class method specifically designed for analyzing compositional data with essential zeros. To our knowledge, this will be the first unified latent class modeling approach for such data.

This paper is organized as follows: Section 5.2 describes the motivation for the proposed joint mixture model in the physical activity setting and how it relates to compositional data. Section 5.3 defines the proposed joint mixture model and presents properties of the model, the algorithm for posterior inference, and centroids for compositional data with essential zeros. Section 5.4 presents results from simulation experiments. Section 5.5 applies our method to the Hispanic Community Health Study / Study of Latinos (HCHS/SOL) for characterizing latent profile of accelerometry-assessed sedentary behaviors and physical activity and describing associations with abdominal obesity (adiposity). Lastly, section 5.6 provides a summary of our work with a discussion of future work and potential applications.

5.2 Motivation

5.2.1 Notation and Data Structure

Compositional data, mathematically, is defined as multivariate random vectors of size D, $\boldsymbol{y} = (y_1, \ldots, y_D)'$, which lie within the simplex of size D:

$$\mathbb{S}^D = \{\boldsymbol{y} = (y_1, \ldots, y_D)', \sum_{j=1}^{D} y_j = \kappa, y_j > 0, j = 1, \ldots, D\}, \tag{5.48}$$

where κ is some positive constant. Let physical activity time use variables, $\boldsymbol{t}_i = (t_{i1}, t_{i2}, t_{i3}, t_{i4})'$, represent the amount of time spent in sedentary, light, moderate, and vigorous activity for the i^{th} participant, respectively. When all study participants are observed over the same time period, $\kappa = \sum_{j=1}^{4} t_{ij}$ for $i = 1, \ldots, n$, then each study participant's time use variables are four part compositions, $\boldsymbol{t}_i \in \mathbb{S}^4$.

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL) is a community based prospective cohort study of 16,415 self-identified Hispanic/Latino adults (aged 18 - 74 year) from randomly selected households in four U.S. field centers (Chicago, IL; Miami, FL; Bronx, NY; San Diego, CA), whose goals are to describe the prevalence of risk and protective factors of certain chronic conditions and quantify certain health outcomes of interest over time (LaVange et al., 2010). At baseline, study participants were instructed to wear accelerometers on their right hip for seven days during wake hours. In HCHS/SOL, accelerometry-assessed time use variables do not have uniform wear time across participants, κ but subject specific wear time, $\kappa_i = \sum_{j=1}^{4} t_{ij}$. To circumvent this issue and project these time use variables to a four component simplex space, time use variables can be scaled by wear time which defines time budget proportions, $\boldsymbol{y}_i = \boldsymbol{t}_i/\kappa_i$. Time budget proportions are contained within the four component unit simplex, $\boldsymbol{y}_i \in \mathbb{S}^4$ which implies $y_{ij} \in \mathbb{R}_{[0,1]}$ for all j and $\kappa = 1$. This restriction can be thought of as focusing on scaled compositional data or compositional multivariate proportions. In the rest of the paper, we index the compositional data to include the i^{th} observation in a sample of size n and l^{th} repeated measure, $\boldsymbol{y}_{il} = (y_{il1}, \ldots, y_{ilD})'$.

5.2.2 Mixtures on the simplex

Mixture models have been an important tool for analyzing multivariate non-homogeneous data, however, these have been limited for compositional data. The mixture of Dirichlet distributions model has been used for analyzing non-homogeneous compositional data, however, the Dirichlet distribution is very restrictive, for example, only allowing a particular negative dependence in the elements (Barrientos et al., 2015; Calif et al., 2011b; Comas-Cufí et al., 2016).

Log ratio methods have contributed greatly to the development of more flexible and interpretable modeling of compositional data (Pawlowsky-Glahn et al., 2015; Mateu-Figueras et al., 2013). Log ratio methods rely on a mathematical framework which induces a measure on the simplex via a one-to-one mapping, $h : \mathbb{S}^D \to \mathbb{R}^{D-1}$, from the simplex to the real space (Mateu-Figueras et al., 2011, 2013). This is called the principle of working on the coordinates and is appealing because all standard statistical methods and properties can be applied on and transfered to the simplex (Mateu-Figueras et al., 2013). The function h is called a coordinate function and often chosen to be a log ratio. The isometric log ratio coordinate function,

$$h(\boldsymbol{y}) = \text{ilr}(\boldsymbol{y}) = \left(\sqrt{\frac{j}{j+1}} \ln \left(\frac{\sqrt[j]{\prod_{j^*=1}^{j} y_{j^*}}}{y_{j+1}} \right) \right)_j, j = 1, \ldots, D-1, \tag{5.49}$$

is an orthonormal log ratio coordinate system which has recently been used for analyzing compositional data (Chastin et al., 2015; Dumuid et al., 2017b; Rivera-Pinto et al., 2018). Other log ratio coordinate functions have been used as well, such as, the additive log ratio (alr) and the centered log ratio (clr) (Aitchison, 2005). The additive log ratio coordinate function is defined as,

$$h(\boldsymbol{y}) = \text{alr}(\boldsymbol{y}) = \left(\ln \left(\frac{y_j}{y_D} \right) \right)_j, j = 1, \ldots, D-1. \tag{5.50}$$

The normal distribution on the simplex was defined using this framework which allows for improved flexibility over the Dirichlet distribution (Mateu-Figueras et al., 2013).

5.2.2.1 The mixture model on the simplex

The principle of working on the coordinates enabled the definition of mixture models on the simplex. Comas-Cufí et al. (2016) specify, however, note that the mixture methodology requires mixtures to be defined on basis invariant distributions and most common distributions are basis invariant (e.g. multivariate normal). The mixture model for compositional data is defined as:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{K}(\boldsymbol{y}^*|\boldsymbol{\theta}_k), \tag{5.51}$$

where $\boldsymbol{y}^* = h(\boldsymbol{y})$, h is a coordinate function, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$, and $\mathcal{K}$ is a multivariate kernel (i.e. distribution) (Comas-Cufí et al., 2016). The mixtures of multivariate Gaussian (BMMG) and skewed-normals on the simplex are defined by choosing the respective kernel. The BMMG is a versatile method for modeling non-homogeneous compositional data that can capture more general forms of dependence over the mixture of Dirichlet distributions.

5.2.2.2 The mixture of product kernels on the simplex

The mixture of product kernels methodology can be adapted for compositional data using the principle of working on the coordinates. The mixture of product kernels methodology assumes all variables are mutually independent conditional on a latent class and is highly flexible (Dunson and Bhattacharya, 2011). When this framework is combined with the principle of working on the coordinates, the mixture of product kernels (BMPK) on the simplex is defined as:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D-1} \mathcal{K}_j(y_j^*|\boldsymbol{\theta}_k), \tag{5.52}$$

$\boldsymbol{y}^* = h(\boldsymbol{y})$, h is a coordinate function, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$, and $\mathcal{K}_j$ is a coordinate component specific kernel. We hereafter refer to BMPK for the case when the coordinate function is the ilr with component specific normal kernels. This specification is a special case of the BMMG with zero correlations.

The mixture of product kernels is similar to latent profile analysis (LPA) and equivalent to latent class analysis (LCA) when analyzing n-way contingency tables using a mixture of product multinomials model (Dunson and Xing, 2009). An appealing feature of this methodology is that BMPK can be used to build densities for mixed-scale variables by simply augmenting compositional data with other variables and selecting an appropriate kernel (Davalos et al., 2019; Dunson and Bhattacharya, 2011). This idea can be leveraged to jointly profile physical activity and a health outcome of interest thereby producing supervised latent profiles.

5.2.2.3 Limitations

The Center for Disease Control and Prevention (CDC) estimates that approximately 22.9% of U.S. adults aged 18–64 meet the recommended physical activity guidelines of at least 150 minutes of MPA, 75 minutes of VPA, or a combination of both per week. In HCHS/SOL, using objectively measure SBPA, approximately 60% and 70% of the target population do not achieve any vigorous physical activity during the week and weekend, respectively. In attempting to characterize objectively measure SBPA from a time use perspective, these figures highlight the need for compositional data analysis methods that can account for essential zeros.

The mixture modeling methods presented in the previous sections cannot account for essential zeros because of limitations in the coordinate functions and kernels. The log ratio coordinate functions can yield infinite or undefined values in the presence of essential zeros. Consider the ilr coordinates of time budget proportions on participants who do not achieve MVPA, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, 0, 0)$, then the third coordinate is undefined as $\propto \ln(0/0)$. The third ilr coordinates for participants with no VPA, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3}, 0)$, are

infinite, $\propto \ln(\prod_{j=1}^{3} y_{ij}/0)$. Additionally, it is not possible to evaluate common distribution kernels (e.g. multivariate normal, Dirichlet) at these values.

5.2.2.4 Essential zeros adjustment strategies

Essential zeros can be adjusted for in mixture models on the simplex by imputing zeros or stratification. There are many imputation methods for rounded zeros (i.e. zeros below limit of detection) which include simple methods of replacing zero values with a small arbitrary non-zero value to highly sophisticated robust methods (Martín-Fernández et al., 2012). However, treating essential zeros as rounded zeros may not be a tenable approach for SBPA since there can be a large amount of zeros and the non-zero values that do inform the imputation process may unduly bias the time budget proportions of participants with no MVPA. For profiling SBPA, it is possible imputed zero values may obscure some of the latent classes.

Stratification, or stratified analyses by essential zero configuration, may be an alternative to circumventing essential zeros in compositional data. For objectively measured SBPA, essential zero configurations are typically of the form: all sedentary $(1, 0, 0, 0)$, no MVPA $(y_{i1}, y_{i2}, 0, 0)$, no VPA $(y_{i1}, y_{i2}, y_{i3}, 0)$, and no zeros $(y_{i1}, y_{i2}, y_{i3}, y_{i4})$. A stratified analysis would entail modeling the no MVPA, no VPA, and no zeros configurations separately which can pose modeling difficulties if there are few participants in some of the configurations. Additionally, two sets of stratified analyses will have to be considered when profiling weekday and weekend SBPA time budget proportions. The difficulty with stratification by essential zero configuration is that profiles are defined by the zero configurations and it may be useful to, for instance, collapse profiles of individuals with high MPA and no VPA with individuals with moderate MPA and low VPA. Lastly, stratification complicates interpretation when latent profiles are derived within each configuration in attempting to draw unified conclusions.

5.3 Tensor mixture model on the simplex

The principle of working on the coordinates and product kernels modeling can be jointly leveraged to account for essential zeros im mixture modeling. As detailed in the previous section, the ilr coordinates are not useful in the presence of zeros, however, the alr coordinates can allow accounting for zeros in all but one compositional component. The alr coordinates together with the product kernels allows the use of zero inflated-like kernels for accounting for essential zeros. Additionally, the tensor mixture model is used to conveniently enumerate all zero pattern configurations. In the subsections that follow, we derive our proposed methodology.

5.3.0.1 Derivation

Definition 5.1. (Tensor mixture model on the simplex) Let $\boldsymbol{\pi} = \{\pi_{k_1 \cdots k_{D-1}}, k_j = 1, \ldots, K_j, j = 1, \ldots, D-1\} \in \Pi_{K_1 \cdots K_{D-1}}$ a probability tensor in the space of all probability tensors of size $K_1 \times \cdots \times K_{D-1}$ and $\boldsymbol{y} \in \mathbb{S}^D$. The density is said to be tensor mixture on the simplex if:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k_1=1}^{K_1} \cdots \sum_{k_{D-1}=1}^{K_{D-1}} \pi_{k_1 \cdots k_{D-1}} f_{\boldsymbol{Y}, k_1 \cdots k_{D-1}}(\boldsymbol{y}|\boldsymbol{\theta}_{k_1 \cdots k_{D-1}}), \tag{5.53}$$

where each $f_{\boldsymbol{Y}, k_1 \cdots k_{D-1}}(\cdot|\boldsymbol{\theta}_{k_1 \cdots k_{D-1}})$ is density defined on the simplex.

Following Comas-Cufí et al. (2016), we extend the mixture model on the simplex. We define the tensor mixture model on the simplex by defining a tensor mixture model of product kernels on the coordinates.

Definition 5.2. (Tensor mixture of product kernels model on the simplex (BTMPK)) Let $h : \mathbb{S}^D \to \mathbb{R}^{D-1}$ be a coordinate function, $\boldsymbol{y} \in \mathbb{S}^D$, and $\boldsymbol{y}^* = h(\boldsymbol{y})$. The density $f_{\boldsymbol{Y}}$ is a tensor mixture of product kernels on the simplex if:

$$f_{\boldsymbol{Y}, k_1 \cdots k_{D-1}}(\boldsymbol{y}|\boldsymbol{\theta}_{k_1 \cdots k_{D-1}}) = f^*_{k_1 \cdots k_{D-1}}(h(\boldsymbol{y})|\boldsymbol{\theta}_{k_1 \cdots k_{D-1}}) \tag{5.54}$$

$$= \prod_{j=1}^{D-1} \mathcal{K}_{k_j j}(y^*_j|\boldsymbol{\theta}_{k_j j}), \tag{5.55}$$

where $\boldsymbol{y}^* = (h_1(\boldsymbol{y}), \ldots, h_{D-1}(\boldsymbol{y}))'$ and $\mathcal{K}_{k_j j}$ is a kernel.

The tensor mixture modeling framework allows analysts great flexibility in building multivariate models because of how it separates a kernel for each distributional component. It is this flexibility that we leverage to build a distribution that can account for the presence of essential zeros. We emphasize that in difference to other specifications of a tensor mixture of product kernels, the indexing of the kernel not just by j but also by the marginal component k_j. As presented in the next section, we use this to define the zero inflated mixture of Gaussians kernel.

5.3.0.2 Zero inflated mixture of Gaussians kernel

In order to gain insight into the derivation of our proposed model for modeling compositional data with essential zeros, we present a base case working with a three part composition and go through an algebraic exercise of different parameterizations of the tensor mixture model. As a starting point consider the alr coordinate function which requires a referent component to construct the log ratio. We choose to work with this coordinate system since it facilitates accounting for essential zeros in the physical activity time budget

data scenario where it is tenable to assume the sedentary time budget will be non-zero for all participants. Assume the referent component is free from essential zeros but other components can have zeros, $y_D \neq 0$ while $y_j \in [0, 1)$. When $D = 2$ then the presence of essential zeros can be handled by using a zero inflated Gaussian distribution:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = \pi_1\delta_0(y_1) + \pi_2\mathrm{N}(h_1(\boldsymbol{y})|\mu, \sigma^2), \tag{5.56}$$

where $\boldsymbol{\theta} = (\mu, \sigma^2)$, $h(\boldsymbol{y}) = \log(y_1/y_2)$, $\boldsymbol{\pi} \in \mathbb{S}^2$, $y_2 = 1 - y_1$, and $\delta_0(\cdot)$ is the degenerate distribution at 0.

Extending the this framework to higher dimensions is more involved but we increase to one more dimension to reveal greater generalization. Consider $\boldsymbol{y} \in \mathbb{S}^3$ and assume $y_3 \neq 0$ and the additive log ratio coordinate function. Under this scenario, essential zeros can be present in the following combinations: $(0, y_2, y_3), (y_1, 0, y_3), (0, 0, y_3)$ which implies that we have four different configurations. Let any $\boldsymbol{\pi} \in \Pi_{2,2}$ be a probability tensor in the space of all probability tensors of size 2×2, a model can be formulated as a tensor mixture of product kernels on the simplex in the following manner:

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = \pi_{11}\delta_0(y_1)\delta_0(y_2) + \pi_{12}\delta_0(y_1)\mathrm{N}(y_2^*|\mu_2, \sigma_2^2) \\ + \pi_{21}\mathrm{N}(y_1^*|\mu_1, \sigma_1^2)\delta_0(y_2) + \pi_{22}\mathrm{N}(y_1^*|\mu_1, \sigma_1^2)\mathrm{N}(y_2^*|\mu_2, \sigma_2^2), \end{aligned} \tag{5.57}$$

where $\boldsymbol{y}^* = (y_1^*, y_2^*)' = (\log(y_1/y_3), \log(y_2/y_3))'$. This is a unified way of modeling the simplex distribution with essential zeros in difference to a stratified approach by data with essential zero pattern type. By theorem 1 of Dunson and Xing (2009), the probability tensor $\boldsymbol{\pi}$ can be characterized by the PARAFAC decomposition where $\pi_{k_1k_2} = \sum_{k=1}^K \lambda_k\psi_{k_1 1k}\psi_{k_2 2k}$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)' \in \mathbb{S}^K$, and $\psi_{jk} = (\psi_{1jk}, \psi_{2jk})' \in \mathbb{S}^2$ for $k = 1, \ldots, K$ and $j = 1, 2$. With this, 5.57 can be reparametrized as:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^K \lambda_k \prod_{j=1}^2 \Big(\psi_{1jk}\delta_0(y_j) + \psi_{2jk}\mathrm{N}(y_j^*|\mu_j, \sigma_j^2)\Big). \tag{5.58}$$

This reparameterization under the PARAFAC assumption on the tensor shows an equivalence between the tensor mixture of product kernels with the mixture of product mixtures model. With the latter, we can see that the specification we started with yields the mixture of product zero-inflated Gaussians model. Greater flexibility can be built in if the zero-inflated Gaussians mixture is extended to the zero-inflated mixture of Gaussians.

Definition 5.3. (Zero-inflated mixture of Gaussians) Let $\boldsymbol{y} \in \mathbb{S}^D$, $\boldsymbol{\psi}_j \in \mathbb{S}^K$, and $y_j^* = \log(y_j/y_D)$ with $y_D \neq 0$. The following density is said to be a zero-inflated mixture of Gaussians if,

$$f_j^*(y_j^*|\boldsymbol{\theta}_j) = \psi_{1j}\delta_0(e^{y_j^*}) + \sum_{k_j=2}^{K_j} \psi_{k_j j}\mathrm{N}(y_j^*|\mu_{k_j-1,j}, \sigma^2_{k_j-1,j}). \tag{5.59}$$

The structure of the tensor mixture coupled with the form of the zero-inflated mixture of Gaussian allows one to see how the mixture model can be leveraged to account for essential zeros in D component compositions. This can be done by simply increasing the number of variables in the tensor mixture of product kernels and using a referent component as a component absent of essential zeros. Note, this is not an issue in certain applications such as our motivating example in SBPA where everyone has some sedentary time, but it can be have limited use in other applications such as dietary intake or microbiome data.

5.3.0.3 Hierarchical Model

In this section, we present the unified extension to stratifying the composition by type with respect to essential zero configuration to compositions of size D as a tensor mixture and in hierarchical form. We define the distribution on the simplex as by defining a tensor mixture on the alr transformed compositions.

Let $\boldsymbol{y} \in \mathbb{S}^D$ with $\boldsymbol{y}^* = h(\boldsymbol{y})$ and $\boldsymbol{\pi} \in \Pi_{K_1 \cdots K_{D-1}}$ a probability tensor. Our proposed tensor mixture model on the simplex (TMS) is defined as:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k_1=1}^{K_1} \cdots \sum_{k_{D-1}=1}^{K_{D-1}} \pi_{k_1 \cdots k_{D-1}} \prod_{j=1}^{D-1} \mathcal{K}_{k_j j}(y_j^*|\boldsymbol{\theta}_{k_j j}), \tag{5.60}$$

where $\mathcal{K}_{1j}(y_j^*|\boldsymbol{\theta}_{1j}) = \delta_0(y_j)$, $\mathcal{K}_{k_j j}(y_j^*|\boldsymbol{\theta}_{k_j j}) = \mathrm{N}(y_j^*|\mu_{k_j-1,j}, \sigma^2_{k_j-1,j})$ for $k_j = 2, \ldots, K_j$, $\boldsymbol{y}^* = h(\boldsymbol{y})$, and h is the alr coordinate function $h_j(\boldsymbol{y}) = \log(y_j/y_D)$. We assume a PARAFAC decomposition on the probability tensor where we augment the data with latent joint and marginal indicator variables similar to the specification of the infinite tensor mixture model (ITM) of Banerjee et al. (2013). We simplify the ITM by taking an sparse finite mixture model approach rather than using a non-parametric Bayes prior (Malsiner-Walli et al., 2016; Rousseau and Mengersen, 2011). Additionally, by fixing the first component of each component to the Dirac measure and augmenting the data with a latent indicator, we inherently augment data similar in spirit to Dunson and Herring (2005). Suppose $\boldsymbol{x} = (x_1, \ldots, x_{D-1})'$ is the collection of marginal indicator variables then $x_j = 1$ if $y_j = 0$ and $x_j = 2, \ldots, K_j$ otherwise.

Supposing f_Y is BTMPK and $\boldsymbol{y}_i \sim f_Y$ with $\boldsymbol{y}_i \in \mathbb{S}^D$ for $i = 1, \ldots, n$, our proposed model can be expressed in hierarchical form as:

$$y^*_{ij}|x_{ij} \sim \mathcal{K}_{x_{ij}j}(\boldsymbol{\theta}_{k_jj}) \tag{5.61}$$

$$\Pr(x_{ij} = k_j|z_i) = \psi_{k_jjz_i} \tag{5.62}$$

$$\Pr(z_i = k) = \lambda_k \tag{5.63}$$

$$\boldsymbol{\psi}_{jk} = (\psi_{1jk}, \ldots, \psi_{K_jjk})' \sim \text{Dir}(\boldsymbol{a}_j), \text{ for } k = 1, \ldots, K \tag{5.64}$$

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)' \sim \text{Dir}(\boldsymbol{\alpha}) \tag{5.65}$$

$$\mu_{k_jj}|\sigma^2_{k_jj} \sim \text{N}(\mu_j, \sigma^2_{k_jj}/\tau_j) \tag{5.66}$$

$$\sigma^2_{k_jj} \sim \text{IG}(a_{\sigma j}, b_{\sigma j}), \text{ for } k_j = 1, \ldots, K_j - 1 \tag{5.67}$$

$$\mu_j \sim \text{N}(\mu_{0j}, \sigma^2_{0j}), \text{ for } j = 1, \ldots, D-1, \tag{5.68}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$, $\boldsymbol{a}_j = (a_{1j}, \ldots, a_{K_jj})'$, $\boldsymbol{\theta}_{1j} = \emptyset$, and $\boldsymbol{\theta}_{k_jj} = (\mu_{k_j-1,j}, \sigma^2_{k_j-1,j})'$ for $k_j = 2, \ldots, K_j$ and $j = 1, \ldots, D-1$.

The joint likelihood with the augmented latent indicator variables can be expressed as:

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_{\boldsymbol{x}_i} f_{\boldsymbol{Y},\boldsymbol{x}_i}(\boldsymbol{y}_i|\boldsymbol{\theta}_{\boldsymbol{x}_i}) \tag{5.69}$$

$$= \prod_{i=1}^{n} \pi_{\boldsymbol{x}_i} f^*_{\boldsymbol{Y}^*,\boldsymbol{x}_i}(\boldsymbol{y}^*_i|\boldsymbol{\theta}_{\boldsymbol{x}_i}) \tag{5.70}$$

$$= \prod_{i=1}^{n} \lambda_{z_i} \prod_{j=1}^{D-1} \psi_{x_{ij}jz_i} \mathcal{K}_{x_{ij}j}(y^*_{ij}|\boldsymbol{\theta}_{x_{ij}j}) \tag{5.71}$$

$$= \left\{ \prod_{i=1}^{n} \lambda_{z_i} \right\} \left\{ \prod_{j=1}^{D-1} \prod_{k=1}^{K} \psi_{1jk}^{n_{1jk}} \right\} \left\{ \prod_{j=1}^{D-1} \prod_{i:x_{ij}>1} \psi_{x_{ij}jz_i} \text{N}(y^*_{ij}|\mu_{x_{ij}-1,j}, \sigma^2_{x_{ij}-1,j}) \right\}, \tag{5.72}$$

where $n_{k_jjk} = \sum_{i=1}^{n} I(x_{ij} = k_j) I(z_i = k)$ for $k_j = 1, \ldots, K_j$, $j = 1, \ldots, D$, and $k = 1, \ldots, K$.

5.3.0.4 Gibbs sampling algorithm

The Gibbs sampling algorithm proceeds as follows:

1. Update the probability tensor $\boldsymbol{\pi}$ by update all components:

   (a) Sample $\boldsymbol{\lambda}|- \sim \text{Dir}(\boldsymbol{n} + \boldsymbol{\alpha})$ where $\boldsymbol{n} = (n_1, \ldots, n_K)'$ and $n_k = \sum_{i=1}^{n} I(z_i = k)$.

   (b) Sample $\boldsymbol{\psi}_{jk}|- \sim \text{Dir}(\boldsymbol{n}_{jk} + \boldsymbol{a}_j)$ where $\boldsymbol{n}_{jk} = (n_{1jk}, \ldots, n_{K_jjk})'$, $n_{k_jjk} = \sum_{i=1}^{n} I(x_{ij} = k_j) I(z_i = k)$, and $k = 1, \ldots, K$.

2. Update joint allocation variables for $i = 1, \ldots, n$ by sampling z_i such that $\Pr(z_i = k|-) \propto \lambda_k \prod_{j=1}^{D-1} \psi_{x_{ij}jk}$.

3. Update marginal allocation variables for $j = 1, \ldots, D-1$ and $i = 1, \ldots, n$ by sampling x_{ij} such that $\Pr(x_{ij}|-) \propto \psi_{k_{jj}z_i} \mathcal{K}_{k_{jj}}(y^*_{ij}|\boldsymbol{\theta}_{k_{jj}})$. Equivalently, $\Pr(x_{ij} = k_j|-) \propto \psi_{k_{jj}z_i} \text{N}(y^*_{ij}|\mu_{k_j-1,j}, \sigma^2_{k_j-1,j})$ if $y_{ij} \neq 0$ otherwise set $x_{ij} = 1$.

4. Update base measure atoms for $k_j = 1, \ldots, K_j - 1$ and $j = 1, \ldots, D-1$ by sampling $\sigma^2_{k_j,j}|- \sim \text{IG}(\hat{a}_{\sigma k_{jj}}, \hat{b}_{\sigma k_{jj}})$ and $\mu_{k_{jj}}|- \sim \text{N}(\hat{\mu}_{k_{jj}}, \sigma^2_{k_{jj}}/(\tau_j + n_{k_{jj}}))$ where $n_{k_{jj}} = \sum_{i=1}^{n} I(x_{ij} = k_j + 1)$, $\hat{a}_{\sigma k_{jj}} = a_{\sigma j} + n_{k_{jj}}/2$, $\hat{b}_{\sigma k_{jj}} = b_{\sigma j} + \frac{1}{2}\big[SS_{k_{jj}} + (n_{k_{jj}} + \tau_j)(1 - p_{k_{jj}})p_{k_{jj}}(\bar{y}^*_{k_{jj}} - \mu_j)^2\big]$, $SS_{k_{jj}} = \sum_{i:x_{ij}=k_j+1}(y^*_{ij} - \bar{y}^*_{k_{jj}})^2$, $p_{k_{jj}} = \tau_j/(\tau_j + n_{k_{jj}})$, $\bar{y}^*_{k_{jj}} = \sum_{i:x_{ij}=k_j+1} y^*_{ij}/n_{k_{jj}}$, and $\hat{\mu}_{k_{jj}} = p_{k_{jj}}\mu_j + (1 - p_{k_{jj}})\bar{y}^*_{k_{jj}}$.

5. Update base measure hyperparameters for $j = 1, \ldots, D-1$ by sampling $\mu_j|- \sim \text{N}(\hat{\mu}_{0j}, \hat{\sigma}^2_{0j})$, where $\hat{\sigma}^2_{0j} = (\tau_j(\sum_{k_j=1}^{K_j} \sigma^{-2}_{k_j,j}) + \sigma^{-2}_{0j})^{-1}$ and $\hat{\mu}_{0j} = \hat{\sigma}^2_{0j}(\tau_j \sum_{k_j=1}^{K_j}(\mu_{k_{jj}}/\sigma^2_{k_j,j}) + (\mu_{0j}/\sigma^2_{0j}))$.

By construction, the BTMPK maximizes the use of the data. Stratified approaches do not allow use of all the data because these require partitioning of the data conditional on essential zero configuration type.

We can incorporate additional days of data by increasing the multivariate index to include more repeated measures. Given that the BTMPK is defined on the alr coordinates, the general framework of the tensor mixture of product kernels can allow us to jointly model the distribution of repeated measure by simply including more tensor components.

5.3.0.5 Clustering for subclass identification

Researchers often wish to classify study participants by activity patterns (e.g. sedentary, weekend warriors). A single joint allocation variable is used to derive the profiles of the clusters. We select an optimal joint allocation variable as the allocation variable minimum mean squared error of all pairwise participant cluster matching (Dahl, 2006). This can be thought of as selecting the joint allocation variable that maintains greatest consistency in the cluster make-up.

5.3.0.6 Centroids

When staying in the simplex plane, the usual definition of expectation or center is updated since densities are defined with respect the simplicial Lesbegue measure and not the usual Lesbegue measure on $\mathbb{R}^{D-1}$. The compositional expectation is also referred to as the center or centroid. The following definition is adapted

the simplex from Mateu-Figueras et al. (2013) where they present a more general definition of any subset of $E \subset \mathbb{R}^D$.

Definition 5.4. (Compositional Expection) Let $\boldsymbol{y} \in \mathbb{S}^D$ and $h : \mathbb{S}^D \to \mathbb{R}^{D-1}$ a coordinate function on $\mathbb{S}^D$. The expectation on the simplex is:

$$\text{Cen}(\boldsymbol{y}) = E_{\mathbb{S}^D}(\boldsymbol{y}) = h^{-1}\big(\int \boldsymbol{y}^* f_{\boldsymbol{Y}^*}(\boldsymbol{y}^*) d\boldsymbol{y}^* \big) \tag{5.73}$$

$$= h^{-1}(E(\boldsymbol{y}^*)), \tag{5.74}$$

where $\boldsymbol{y}^* = h(\boldsymbol{y})$ and $E(\boldsymbol{y}^*)$ exists (Mateu-Figueras et al., 2013).

The closed geometric mean from a sample of compositions is the compositional linear and unbiased estimator of the centroid (Pawlowsky-Glahn and Egozcue, 2002). The closed geometric mean is also equivalent to the inverse alr of the sample mean of alr transformed compositions. The sample mean being the MLE of a normal distribution coupled with the alr assumption on the composition implies the centroid estimator can be deduced via.

The closure of the geometric mean is a problematic in the presence of essential zeros. The presence of an essential zero is problematic because if present the geometric mean for a component containing the zero is also zeroed producing the effect of zeroing out a component. Conversely, if one ignores zeros in a component then the resulting estimator can over-inflate other components leading to misleading results. The problems with the closure of the geometric mean in the presence of essential zeros can be circumvented by defining alternative estimators.

Definition 5.5. (Convex combination centroids) Let $\boldsymbol{Y}$ be D-part compositions with essential zeros then the convex combination centroid is:

$$\text{Cen}(Y) = \sum_{k_1=0}^{1} \cdots \sum_{k_D=0}^{1} \pi_{k_1 \cdots k_D} \boldsymbol{\zeta}_{k_1 \cdots k_D}, \tag{5.75}$$

where $\pi_{k_1 \cdots k_D} = 0$ when $k_j = 0$ for all j, $\boldsymbol{\zeta}_{k_1 \cdots k_D} = (\zeta_{1,\boldsymbol{k}}, \ldots, \zeta_{D,\boldsymbol{k}})' \in \mathbb{S}^D$ and $\zeta_{j,\boldsymbol{k}} = 0$ if $k_j = 0$.

Consider the case where $D = 3$, then the convex combination of centroids becomes:

$$\text{Cen}(\boldsymbol{Y}) = \pi_{100}\boldsymbol{\zeta}_{100} + \pi_{010}\zeta_{010} + \pi_{001}\zeta_{001} + \pi_{011}\zeta_{011} + \pi_{101}\zeta_{101} + \pi_{110}\zeta_{110} + \pi_{111}\zeta_{111} \tag{5.76}$$

where sub-centroids $\boldsymbol{\zeta}_{k_1 k_2 k_3}$ with two indices set to zero correspond to a corner composition. For instance, $\boldsymbol{\zeta}_{001} = (0, 0, 1)'$. By construction of BTMPK, then we can summarize the sub-centroid for each essential

zero configuration. In physical activity data, for instance, when time budgets are multivariate compositional proportions of sedentary, light, and moderate to vigorous activity then the corresponding sub-centroids would be centroids for participants with all sedentary time ($\boldsymbol{\zeta}_{100}$), no MV ($\boldsymbol{\zeta}_{110}$), and some MV ($\boldsymbol{\zeta}_{111}$). All other subcentroids can be ignored as it is unlikely there may be participants whose entire waking hours are composed of moderate to vigorous behaviors. As such, we may restrict using the zero-inflated mixture of Gaussian kernels only to higher end of physical activity intensities as these may be the only ones with essential zeros. As a result, our methodology can be used to estimate the essential zero probability configurations, the probability tensor $\boldsymbol{\pi}$.

5.4 Simulation Experiments

The objective of the simulation experiments was to assess the performance of our proposed tensor mixture model on the simplex versus three other latent class methods. We focused our attention on the determining if the methods identified the correct number of true underlying subpopulation and whether these underlying subpopulations produced centroids representative of true subpopulation centroids. A secondary objective was to assess whether jointly modeling the compositions with a dichotomous outcome improved performance, which is hereafter referred to as supervised.

Three simulation scenarios were considered, two of which had similar data generating mechanism while the third a different. The first two scenarios were generated using a three component mixture model on the unit simplex, each with three subclasses which correspond to different centroids (Comas-Cufí et al., 2016). The first scenario within this framework (scenario 1) was generated with each subclass having positive correlation (AllPosCorr). The second scenario within this framework (scenario 2) generated data with two subclasses having positive correlation and one subclass with negative correlation (2Pos1NegCorr), see Figure 5.5 for ternary plots of example data sets in each scenario. Rounded zeros were induced in two of the three subclasses by rounding down to zero all values below the 10^{th} percentile in the corresponding subclass. Within each scenario, 500 data sets were generated each with a sample size of 1000.

The third simulation scenario (scenario 3) generated data according to a four component tensor mixture model (TMPK 4 comp) on the simplex similar the hierarchical model presented in section 5.3.0.3 with three latent subclasses. This framework has the capability to truly model essential zeros without artificially inducing rounded zeros as in the other scenarios. Essential zeros in this scenario were included in the third and fourth components only. Within this scenario, 500 data sets were generated each with a sample size of 3000. See Figure 5.6 for all combinations of three component sub-composition ternary plots of an example data set in

this scenario. Because all simulated data sets contain zeros, all cluster-specific centroids are summarized by the proposed convex-combination centroid.

The proposed method was implemented on each simulated data set within each scenario along with three other latent class methods for comparison. The other methods included: a modularized version of the proposed method (MOTEF), mixture of product kernels (BMPK), and a mixture model on the simplex (BMMG) (Comas-Cufí et al., 2016; Dunson and Bhattacharya, 2011). The mixture of product kernels can be thought of as a special case of the mixture model on the simplex since the mixture of product kernels assumed a multivariate Gaussian kernel with zero correlation among all components. BMPK and BMMG methods cannot handle zeros. Thus, in order to implement these two approaches, zero values were treated as rounded and in a pre-implementation step were imputed for each data set separately using the robCompositions package in R. Additionally, the BMMG method is the only method not implemented using supervision. Table 5.10 displays a summary of all latent class method features.

Table 5.10: Summary of latent class methods features.

Method	**Supervision**	**Essential Zeros**
BTMPK	Yes	Yes
MOTEF	Yes	Yes
BMPK	Yes	No
BMMG	No	No

All data sets were generated with an additional dichotomous variable with different probabilities conditional on a subclass. This variable was used for assessing the effect of supervision on the methods where applicable.

All methods were implemented within a Bayesian framework with the same cluster selection procedure. Cluster selection included: 1) specifying symmetric Dirichlet distribution with precision parameter set to $1e-25$ and 2) selecting an optimal clustering by minimizing Binder's loss (Dahl, 2006). For the AllPosCorr and 2Pos1NegCorr scenarios, each method was run for 3,000 iterations dropping the first 1,000 as a burn-in and storing every other iterate with five chains initialized with the number of joint clusters ranging from 50 to 100. For the TMPK 4 component scenario, each method was run for 10,000 iterations dropping the first 5,000 as a burn-in and storing every fifth iterate, with five chains initialized with the number of joint clusters ranging from 50 to 100.

Table 5.11 displays a summary of the unsupervised simulation experiments by method and scenario. The proposed method tended to select one more cluster than the three true underlying subclasses in the AllPosCorr and 2Pos1NegCorr by selecting four clusters in 74.4% and 72.3% of the data sets. In the TMPK 4 component scenario, the proposed method tended to select 5+ latent classes in 85.6% of the data sets. MOTEF produced

the smallest number of clusters with little variability. In the AllPosCorr and 2Pos1NegCorr scenarios, MOTEF selected two clusters in 97.8% of the data sets in each scenario, while correctly selecting three subclasses in the TMPK 4 component scenario in all 500 data sets. As expected, the BMPK method selected the most clusters across all scenarios with modes at 7 in the AllPosCorr (71.0%) and 2Pos1NegCorr (60.4%) scenarios, and 10 clusters for the TMPK 4 component scenario (73.4%). Lastly, the BMMG method performed similar to MOTEF in the AllPosCorr and 2Pos1NegCorr scenarios with modes at two selected clusters, but mostly selected four clusters (96.6%) in the TMPK 4 component scenario.

Supervision seemed to be helpful for MOTEF, ineffective for BMPK, and harmful for BTMPK in terms adequately selecting the true number of underlying subclasses. For MOTEF, supervision improved the number of selected clusters from two to three in the AllPosCorr and 2Pos1NegCorr scenarios by selecting three in 57.4% and 96.4% of the data sets, respectively. The BTMPK method, on the other hand, had increased modes of number of selected clusters from four to five in 79.6% and 87.8% of the data sets in the AllPosCorr and 2Pos1NegCorr scenarios, respectively. For BMPK, supervision seemed largely ineffective in reducing the number of selected clusters where the modes were preserved at 7 in the first scenarios and 10 in the TMPK 4 component scenarios.

In what follows, we describe the quality of the centroids by method and simulation scenario. For methods with supervision, we describe the effect of supervision as well. Figure 5.7 displays ternary plots of estimated cluster-specific convex combination centroids for each data set in the AllPosCorr simulation scenario by supervision status and method. The proposed method, when unsupervised, selected too many clusters in the AllPosCorr scenario and disappointingly only appropriately concentrated centroids across all the data sets in one subclass. However, the concentration of identified subclasses improved for two of the three true centroids under supervision, albeit with a greater number of identified subclasses. MOTEF seemed to concentrate centroids near the true but only when the selected number of subclasses was found to be three (2.2% of data sets). Under supervision, MOTEF concentrations were near the true improved when the number of selected clusters was three (96.4%). The BMPK method seemed largely unaffected by supervision and as expected the method had to select more clusters to capture the correlation in the data. Interestingly, near the center of the simplex, the BMPK seemed to ignore correlation in contrast to the clusters near the edges of the simplex where more clusters were selected.

In the 2Pos1NegCorr scenario, similar results were observed except for MOTEF. MOTEF observed consistency in the models selected when two clusters were observed in the unsupervised scenario. When supervised, MOTEF selected the correct number of subclasses in most of the data sets (96.4%) and the corresponding centroids concentrated near the true. The supervised BTMPK method, even though it selected

too many clusters, concentrated most estimates near the true subclass centroids. The BMMG method performed similar to the unsupervised MOTEF method in the AllPosCorr and 2Pos1NegCorr scenarios (Figure 5.9).

In the TMPK 4 component simulation scenario, as expected, the proposed methods concentrated centroids near the true while the other methods captured different features of the data. Figures 5.10 - 5.13 display ternary plots of estimated cluster-specific convex combination centroids across all data sets by every three component sub-composition combination and method. The MOTEF and BTMPK methods concentrated centroids near the true across all sub-composition combinations, although the BTMPK selected more centroids than the true. MOTEF identified the true number of centroids in all of the data sets correctly concentrated centroid estimates but with some degree of bias in one of the true centroids. The BMMG seemed to capture different features of the data driven by the sub-composition (1,2,3) but the true centroids were mostly missed in the (1,3,4) and (2,3,4) sub-compositions. A similar finding was observed by the BMPK method with a larger number of clusters, however, there was consistent coverage of the true centroids.

In all, the proposed tensor mixture methodology performs best in the presence of essential zeros and may perform adequately under certain circumstances in the presence of rounded zeros. The tensor mixture methodology consistently achieved greater compression of group clustering, especially so for the MOTEF method across all scenarios. Additionally, the proposed methodology performed best in the presence of larger samples or a greater number of variables. The BMPK and BMMG methods performed well at revealing unique features of the data in the presence of essential zeros, with BMPK outperforming the latter in concentrating centroids near the true but at the cost of greatly over selecting clusters. A smaller number of clusters may be appealing to investigators when the goal is to select parsimonious distinguishable groupings in the presence of essential zeros, in which case the proposed methodology is recommended. However, if the goal is feature selection, then the BMMG or the BMPK methods may be preferable.

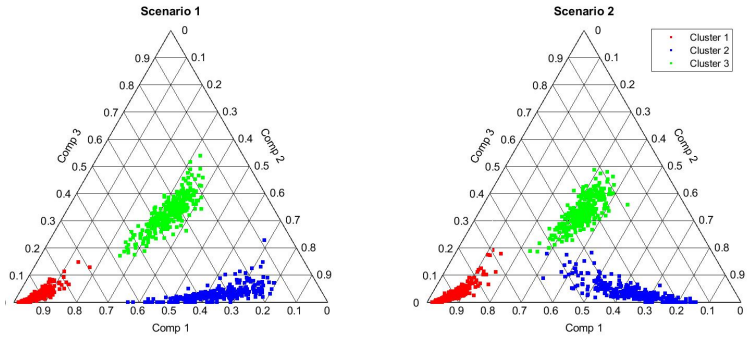

Figure 5.5: Ternary plot of a simulated data set for Scenario 1 (AllPosCorr) and 2 (2Pos1NegCorr).

Table 5.11: Simulation study clustering summary by method and scenario; mean (sd) for continuous variables, percentages for integer valued summaries (out of 500 data sets for each scenario)

		Method			
Scenario	**Summary**	**BTMPK**	**MOTEF**	**BMPK**	**BMMG**
AllPosCorr	Post. Mn. N	3.74 (0.5)	2.02 (0.2)	6.72 (0.5)	2.01 (0.1)
	No. Opt. Clus				
	2	0 (0.0%)	489 (97.8%)	0 (0.0%)	498 (99.6%)
	(*) 3	125 (25.0%)	11 (2.2%)	0 (0.0%)	2 (0.4%)
	4	372 (74.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	5	3 (0.6%)	0 (0.0%)	6 (1.2%)	0 (0.0%)
	6	0 (0.0%)	0 (0.0%)	138 (27.6%)	0 (0.0%)
	7	0 (0.0%)	0 (0.0%)	355 (71.0%)	0 (0.0%)
	8	0 (0.0%)	0 (0.0%)	1 (0.2%)	0 (0.0%)
2Pos1NegCorr	Post. Mn. N	3.79 (0.4)	2.05 (0.2)	6.55 (0.5)	2.03 (0.2)
	No. Opt. Clus				
	2	0 (0.0%)	489 (97.8%)	0 (0.0%)	494 (98.8%)
	(*) 3	123 (24.6%)	11 (2.2%)	0 (0.0%)	6 (1.2%)
	4	363 (72.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	5	14 (2.8%)	0 (0.0%)	17 (3.4%)	0 (0.0%)
	6	0 (0.0%)	0 (0.0%)	180 (36.0%)	0 (0.0%)
	7	0 (0.0%)	0 (0.0%)	302 (60.4%)	0 (0.0%)
	8	0 (0.0%)	0 (0.0%)	1 (0.2%)	0 (0.0%)
TMPK 4 comp.	Post. Mn. N	4.85 (0.5)	3.00 (0.0)	10.13 (0.5)	4.01 (0.2)
	No. Opt. Clus				
	(*) 3	49 (9.8%)	500 (100.0%)	0 (0.0%)	9 (1.8%)
	4	23 (4.6%)	0 (0.0%)	0 (0.0%)	483 (96.6%)
	5	280 (56.0%)	0 (0.0%)	0 (0.0%)	8 (1.6%)
	6	148 (29.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	9	0 (0.0%)	0 (0.0%)	33 (6.6%)	0 (0.0%)
	10	0 (0.0%)	0 (0.0%)	367 (73.4%)	0 (0.0%)
	11	0 (0.0%)	0 (0.0%)	98 (19.6%)	0 (0.0%)
	12	0 (0.0%)	0 (0.0%)	2 (0.4%)	0 (0.0%)

[1] Posterior N[1]: Posterior mean of number of clusters.
[2] (*): The true number of underlying subclasses.

Table 5.12: Simulation study clustering summary of supervised versions by method and scenario; mean (sd) for continuous variables, percentages for integer valued summaries (out of 500 data sets for each scenario)

		Method		
Scenario	**Summary**	**BTMPK**	**MOTEF**	**BMPK**
AllPosCorr	Post. Mn. N	4.59 (0.5)	2.51 (0.5)	6.65 (0.5)
	No. Opt. Clus			
	2	0 (0.0%)	213 (42.6%)	0 (0.0%)
	(*) 3	8 (1.6%)	287 (57.4%)	0 (0.0%)
	4	94 (18.8%)	0 (0.0%)	0 (0.0%)
	5	398 (79.6%)	0 (0.0%)	12 (2.4%)
	6	0 (0.0%)	0 (0.0%)	149 (29.8%)
	7	0 (0.0%)	0 (0.0%)	337 (67.4%)
	8	0 (0.0%)	0 (0.0%)	2 (0.4%)
2Pos1NegCorr	Post. Mn. N	4.86 (0.3)	3.00 (0.1)	6.54 (0.5)
	No. Opt. Clus			
	2	0 (0.0%)	17 (3.4%)	0 (0.0%)
	(*) 3	0 (0.0%)	482 (96.4%)	0 (0.0%)
	4	61 (12.2%)	1 (0.2%)	0 (0.0%)
	5	439 (87.8%)	0 (0.0%)	20 (4.0%)
	6	0 (0.0%)	0 (0.0%)	181 (36.2%)
	7	0 (0.0%)	0 (0.0%)	297 (59.4%)
	8	0 (0.0%)	0 (0.0%)	2 (0.4%)
TMPK 4 comp.	Post. Mn. N	4.59 (0.6)	3.00 (0.0)	10.20 (0.5)
	No. Opt. Clus			
	(*) 3	80 (16.0%)	500 (100.0%)	0 (0.0%)
	4	89 (17.8%)	0 (0.0%)	0 (0.0%)
	5	245 (49.0%)	0 (0.0%)	0 (0.0%)
	6	85 (17.0%)	0 (0.0%)	0 (0.0%)
	7	1 (0.2%)	0 (0.0%)	0 (0.0%)
	9	0 (0.0%)	0 (0.0%)	16 (3.2%)
	10	0 (0.0%)	0 (0.0%)	394 (78.8%)
	11	0 (0.0%)	0 (0.0%)	87 (17.4%)
	12	0 (0.0%)	0 (0.0%)	3 (0.6%)

[1] Posterior N[1]: Posterior mean of number of clusters.
[2] (*): The true number of underlying subclasses.

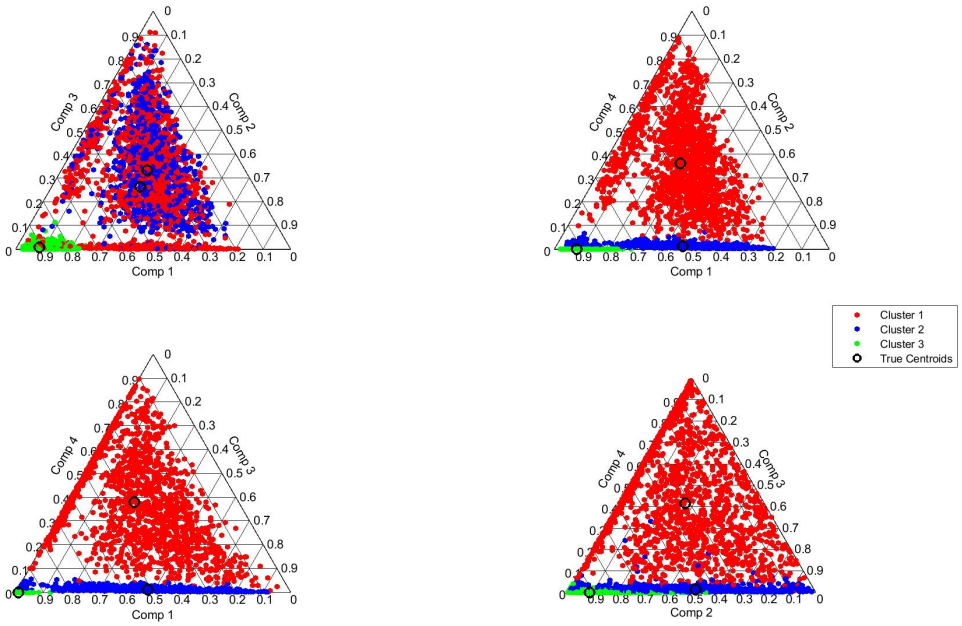

Figure 5.6: Ternary plots of a simulated data for the TMPK 4 component composition scenario displaying all four combinations of the three component sub-compositions.

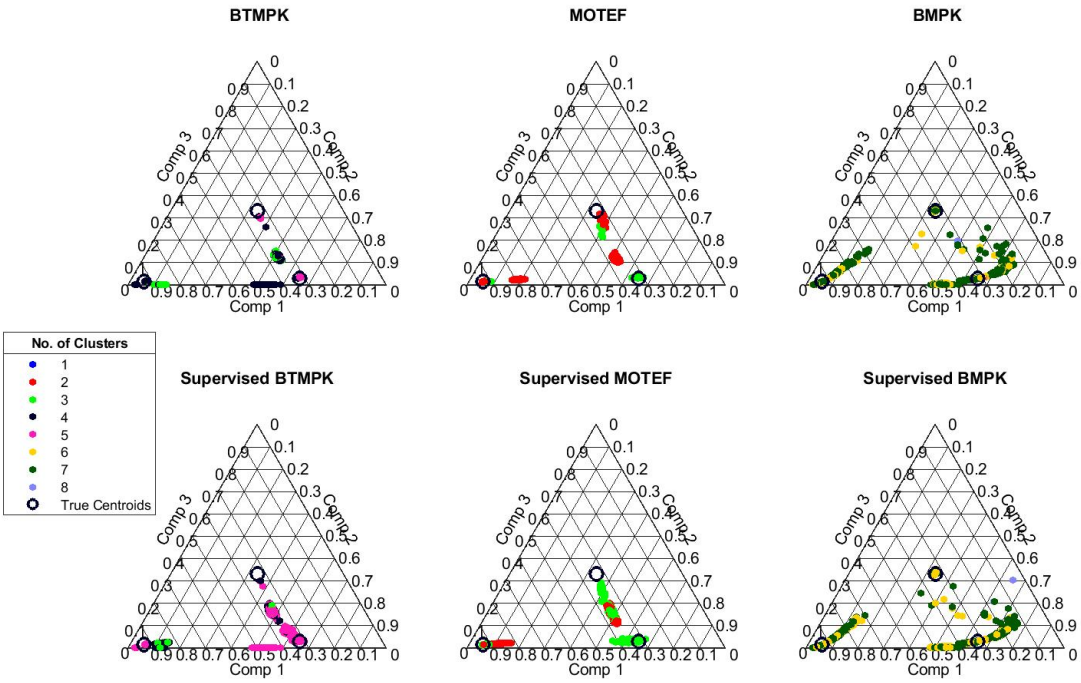

Figure 5.7: Results of unsupervised and supervised simulations for scenario 1 (AllPosCorr): Ternary plots of estimated cluster-specific convex combination centroids for each data set by applied method.

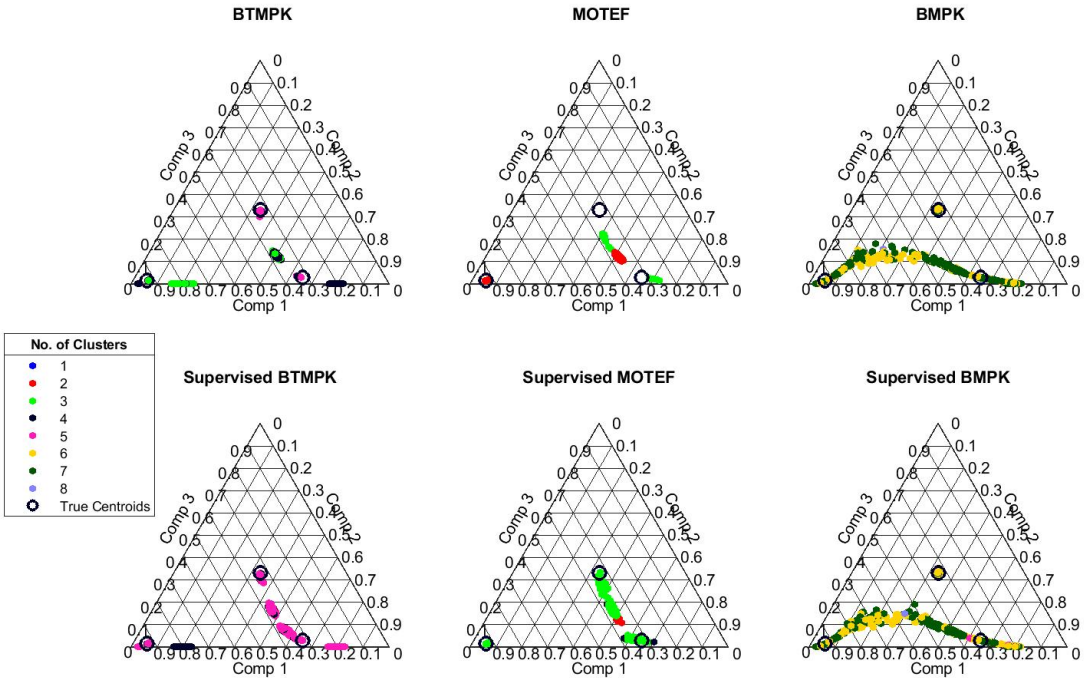

Figure 5.8: Results of unsupervised and supervised simulations for scenario 2 (2Pos1NegCorr): Ternary plots of estimated cluster-specific convex combination centroids for each data set by applied method.

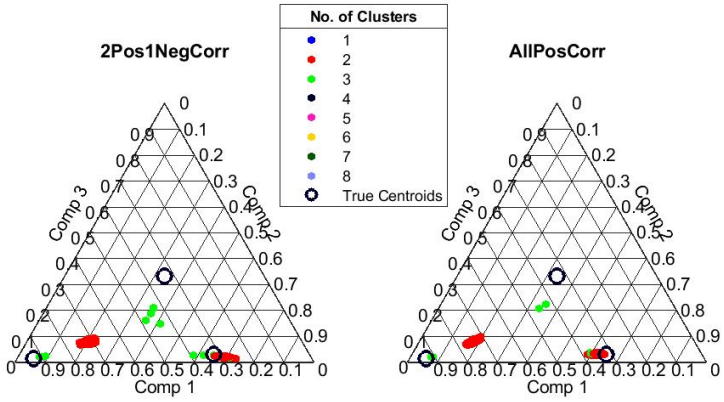

Figure 5.9: Results of unsupervised BMMG simulations for scenarios 1 (AllPosCorr) and 2 (2Pos1NegCorr): Ternary plots of estimated cluster-specific convex combination centroids for each data set by applied method.

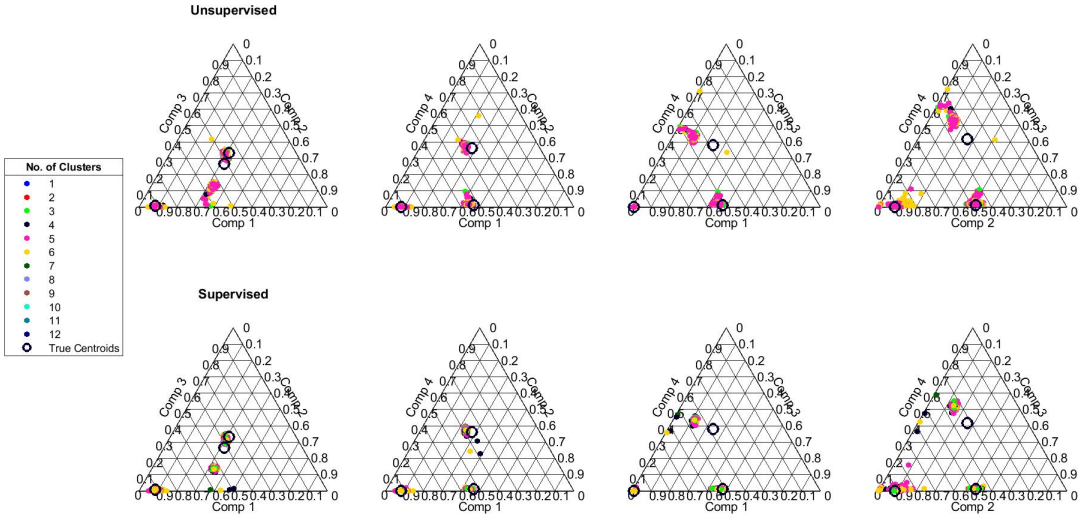

Figure 5.10: Results of unsupervised and supervised BTMPK simulations for scenario 3 (TMPK 4 component compositions): Ternary plots of estimated cluster-specific convex combination centroids for each data set by every 3 component sub-composition combination.

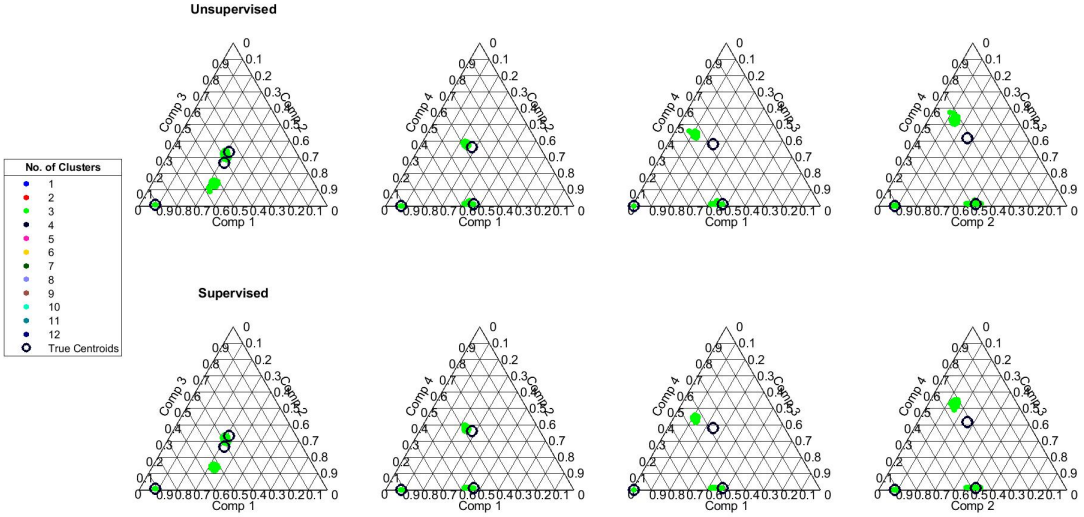

Figure 5.11: Results of unsupervised and supervised MOTEF simulations for scenario 3 (TMPK 4 component compositions): Ternary plots of estimated cluster-specific convex combination centroids for each data set by every 3 component sub-composition combination.

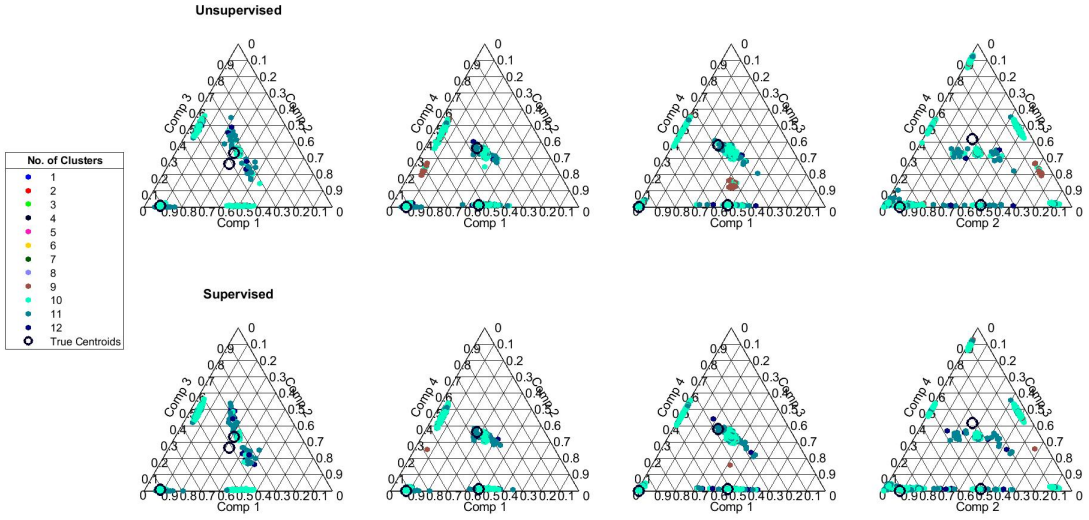

Figure 5.12: Results of unsupervised and supervised BMPK simulations for scenario 3 (TMPK 4 component compositions): Ternary plots of estimated cluster-specific convex combination centroids for each data set by every 3 component sub-composition combination.

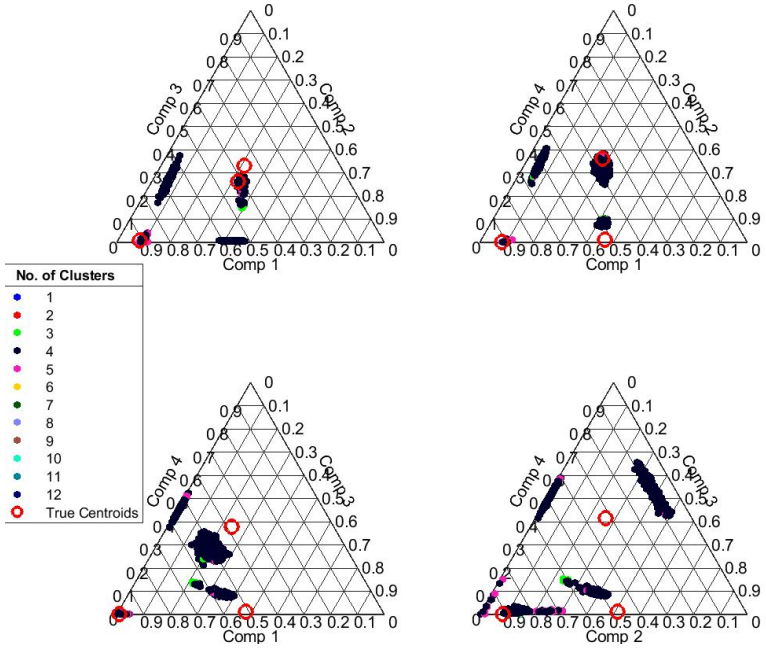

Figure 5.13: Results of BMMG simulations for scenario 3 (TMPK 4 component compositions): Ternary plots of estimated cluster-specific convex combination centroids for each data set by every 3 component sub-composition combination.

5.5 Hispanic Community Health Study/Study of Latinos

The aim of our study is to jointly profile weekday and weekend accelerometry-assessed sedentary behavior and physical activity in HCHS/SOL and assess associations with adiposity. Associations with adiposity were assessed using a two-stage approach. In the first stage, latent classes are derived using our proposed approach. In the second stage, associations between adiposity and the latent classes were assessed via logistic regression with pairwise comparisons of prevalence adiposity by latent classes. We consider the proposed (BTMPK) method with three other methods: Bayesian Mixture of Product Kernels (BMPK), Bayesian Mixture of Multivariate Gaussians (BMMG), and a modularized version of our proposed method (MOTEF). Additionally, we construct and compare abdominal adiposity supervised profiles using the BTMPK, MOTEF, and BMPK methods. Supervised profiling entailed jointly modeling the weekday and weekend accelerometry assessed variables together with abdominal adiposity.

5.5.1 Study population

The analysis was conducted on a subsample of approximately 25% HCHS/SOL participants with adherent accelerometer data from the baseline visit (at least three days with 10 or more hours/day of wear time) and at least one adherent weekend day (≈3,000 participants). Out of the 16,415 study participants, 12,750 wore the accelerometer for at least 10 hours per day for at least 3 days. Out of the accelerometer adherent participants, 11,328 wore the accelerometer for at least one weekend day. 2,954 of these participants were included in the stratified random sample provided for analysis.

5.5.2 Accelerometry data

The study accelerometer (Actical) records omnidirectional movement in g forces, and using a proprietary algorithm provides movement as a count per unit time (i.e. epoch). In HCHS/SOL, counts were recorded in one minute epochs, and non-wear was defined as at least 90 consecutive minutes of zero counts, with allowance of 1 or 2 minutes of nonzero counts if no counts were detected in a 30 minute window upstream and downstream of the 90 minute period (Choi et al., 2011). Days with less than 10 hours of wear time (i.e. non-adherent days) were excluded. For all other days, each minute of wear time was classified as one of four non-overlapping levels: sedentary, light, moderate, and vigorous using standard methods (Colley et al., 2011; Wong et al., 2011). We construct an aggregate total time in each intensity (sedentary, light, moderate, vigorous) separately for adherent weekday and weekend days, instead of taking the arithmetic mean of time within each intensity across adherent days. The aggregate total time in each intensity will be scaled by the

aggregate total wear time to construct compositional multivariate proportions (i.e. % sedentary, % light, % moderate, and % vigorous) for each participant separately for weekday and weekend measures. We hereby refer to these measures as the weekday and weekend time budget proportions (i.e. a total of eight variables). The time budget proportions give us a complete sense of determining how participants distribute their physical activity time by intensity during the weekday and weekend, respectively.

5.5.3 Adjustment for subsampling

The subsampling of the HCHS cohort resulted in the implementation of a two-phase design with stratified sampling in the second phase. All corresponding estimates were as such adjusted for the two phase implementation.

5.5.4 Latent class methods implementation

There were four latent class methods implemented which jointly model the distribution of weekday and weekend time budget proportions. Latent class methods included: 1) BTMPK, 2) MOTEF, 3) BMPK, and 4) BMMG. A feature of the latent class methods except for BMMG is that they can additionally jointly model mixed-scale variables. We leveraged this feature to derive abdominal adiposity (i.e. binary variable) supervised profiles using these methods by jointly modeling weekday and weekend sedentary behavior and physical activity with adiposity. The BMPK and BMMG cannot account for essential zeros and essential zeros were imputed with non-zero values. These methods are the same methods implemented in the simulation experiments, see Table 5.10.

The tensor mixture methodology (BTMPK) was previously. The modularized versions of the tensor mixture model on the simplex followed similar to those described in Chapter 3.

The zero imputed methods were implemented using a two-step approach. In the first step, essential zeros in time budget proportions were treated as rounded or zeros by limits of detection and were then imputed. In the second step, then the imputed data set was analyzed using BMPK and BMMG.

The mixture of multivariate Gaussians latent class method was implemented using the isometric log ratio method. That is, given a weekday and weekend time budget proportion, $(\boldsymbol{y}_{i,weekday}, \boldsymbol{y}_{I,weekday})$, and its corresponding latent class indicator, $z_i = h$ then the corresponding ILR coordinates $(h(\boldsymbol{y}_{i,weekday}), h(\boldsymbol{y}_{I,weekday}))$ are assumed multivariate Gaussian, $N_6(\boldsymbol{\mu}_h, \Sigma_h)$, from the h^{th} kernel. A description of this methodology is detailed elsewhere (Comas-Cufí et al., 2016).

The mixture of product kernels method was implemented similarly. The mixture of product kernels, or latent class analysis, modeling assumptions are such that variables are mutually independent given a latent

class membership, $z_i = h$. For implementation on the simplex, then it is further assumed, given $z_i = h$ then $y^*_{ijweekday}$ is assumed to be normal, $N(\mu_{jh}, \sigma^2_{jh})$ where $y^*_{ijweekday}$ is the jth ILR coordinate of the weekday time budget.

All methods were implemented within a Bayesian framework. Five chains were run on each method for 10,000 iterations discarding the first 5,000 as a burn-in and storing every fifth iterate for a total posterior sample size of 5,000. Each chain was initialized with random groups varying in size from 50 to 100. Standard weakly-informative and empirical priors were used for prior hyper parameters.

5.5.5 Cluster selection

All methods were implemented using a sparse finite mixture model specification with a symmetric Dirichlet prior on component weights and a precision parameter set to 1e-25 (**?**). This highly aggressive approach to reducing the redundant clusters effectively automated selection of modeling components. Uncertainty in the latent class membership was taken into account by selecting a single posterior sample to represent the identified cluster groups as recommended by Dahl (2006). The representative posterior latent class memberships were selected using the minimum least squares pairwise observation similarity approach (Dahl, 2006).

5.5.6 Associations with adiposity

The associations between abdominal adiposity and latent class sedentary behavior and physical activity membership were assessed via logistic regression. Tukey adjusted multiple comparisons were performed for pairwise comparisons of adiposity proportion by latent classes.

5.5.7 Statistical analysis software

Initial data management was conducted using SAS (Institute, 2015). The zero value imputation performed using the robCompositions package in R (Templ et al., 2018). The two phase and logistic regression analyses were implemented using the survey package (Lumley et al., 2004; Lumley and Lumley, 2019). The multiple comparisons were conducted using the multcomp package in R (Hothorn et al., 2017). All latent class analysis was performed using a Matlab Bayesian latent class analysis for compositional data toolbox developed for this paper.

5.5.8 Results

The stratified subsample reflects the characteristics of the HCHS/SOL cohort (Table 5.13). For instance, participants aged 45+ make up 40.2% of the HCHS/SOL cohort while these make up 40.8% of the analysis

subsample. The similar characteristics of the subsample to the full cohort gives us confidence that our findings can be generalizable to the HCHS/SOL target population.

Table 5.13: Summary of HCHS/SOL characteristics displaying n (weighted %) by full study sample, time budget adherent sample, and stratified subsample

			Analysis	
Characteristic	**Levels**	**Full**	**Adherent**	**Subsample**
All		16415	11328	2954
Age	18 - 44	6701 (59.8)	4166 (59.2)	1079 (59.2)
	45+	9714 (40.2)	7162 (40.8)	1875 (40.8)
Gender	Female	9835 (52.1)	6790 (52.4)	1773 (54.1)
	Male	6580 (47.9)	4538 (47.6)	1181 (45.9)
Background	Dominican	1473 (9.9)	1130 (10.7)	322 (10.2)
	Central American	1732 (7.4)	1096 (7.1)	279 (7.3)
	Cuban	2348 (20)	1393 (18.8)	313 (17.3)
	Mexican	6472 (37.4)	4655 (37.5)	1168 (37.6)
	Puerto Rican	2728 (16.1)	1961 (16.8)	579 (17.7)
	South American	1072 (5)	742 (4.9)	206 (4.9)
	More than one/Other heritage	503 (4.1)	327 (4.1)	80 (4.9)
BMI	Underweight	130 (1.2)	83 (1.1)	21 (1.4)
	Normal	3190 (22.1)	2211 (21.9)	574 (21.9)
	Overweight	6115 (37.2)	4312 (37.5)	1121 (35.7)
	Obese	6909 (39.6)	4701 (39.5)	1235 (40.9)

Table 5.14 displays a distributional breakdown of the different activity intensity zero configurations by weekday and weekend. In the HCHS/SOL cohort, approximately half of the participants do not achieve weekday vigorous physical activity, with 70% not achieving this level over the weekend. These characteristics of the data may imply that data imputation may not be a tenable approach when such a high percentage of the data is treated as missing.

Table 5.14: Summary of HCHS/SOL accelerometer-assessed sedentary behavior and physical activity configuration displaying unweighted n (weighted %) by full study sample, time budget adherent sample, and stratified subsample.

			Analysis	
Characteristic	**Levels**	**Full**	**Adherent**	**Subsample**
Weekday	No MV	510 (2.9)	366 (3)	95 (3)
Configuration	No Mod	7 (0)	5 (0)	1 (0)
	No Vig	8253 (48.6)	5585 (45.4)	1461 (47.4)
	No zeros	7645 (48.5)	5372 (51.6)	1397 (49.5)
Weekend	No LMV	5 (0)	3 (0)	2 (0.1)
Configuration	No MV	2008 (11.4)	1380 (10.9)	353 (11.2)
	No Mod	18 (0.2)	11 (0.2)	2 (0.1)
	No Vig	10170 (60.2)	6932 (58.7)	1807 (58.4)
	No zeros	4214 (28.2)	3002 (30.2)	790 (30.2)

Table 5.15 displays population adjusted estimates of overall weekday and weekend time budget proportions by type of estimators extrapolated to 16 hours (i.e. non-sleep hours) for simplicity in presentation. The table

displays centroids using the arithmetic mean, the arithmetic mean with non-zero time budgets (arithmetic mean no zeros), the geometric mean for non-zero data (centroid no zeros), and the proposed convex combination of geometric means by configuration (centroid convex combination). Overall, we find that both the arithmetic mean and the convex combination estimators indicate that participants slightly increase their weekend sedentary time relative to the weekday by displacing weekday time across all physical activity behaviors. The centroid restricted to non-zero data indicates the opposite and displace some sedentary time for light physical activity. The arithmetic mean and the non-zero centroid have similar weekday behavior times which indicates the arithmetic mean likely overestimates time in physical activity behaviors. As such, hereafter the convex combination estimator is used to convey summaries of time budgets due to the large number of zeros in the data for better representation of the true centroids. With this measure, typical participants were found to spend over 12 hours in sedentary, three and a half hours in light, approximately 16 minutes in moderate, and one minute in vigorous activities during the weekday (Table 5.15). Approximately twelve minutes of weekday physical activity are displaced by sedentary time in the weekend.

Table 5.15: Population adjusted estimates of overall weekday and weekend time budget proportions by type of estimators displayed in digital time format (HH:MM:SS) out of 16 hours

		Activity Intensity			
Estimator	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
Arithmetic mean	WEEKDAY	11:52:43	03:41:53	00:21:58	00:03:24
	WEEKEND	12:02:46	03:37:27	00:17:19	00:02:26
Arithmetic mean (no zeros)	WEEKDAY	11:22:18	03:58:21	00:32:27	00:06:52
	WEEKEND	11:14:09	04:03:44	00:33:59	00:08:06
Centroid (conv. comb.)	WEEKDAY	12:12:19	03:30:49	00:15:46	00:01:04
	WEEKEND	12:23:52	03:23:32	00:11:41	00:00:53
Centroid (no zeros)	WEEKDAY	11:42:11	03:50:32	00:25:05	00:02:10
	WEEKEND	11:36:45	03:54:24	00:25:54	00:02:56

5.5.9 Latent classes

The number of identified latent classes, or clusters, varied by method (Table 5.16). The BMPK selected the most clusters at 18 while MOTEF selected the least at three. BMMG, BTMPK, and the supervised BTMPK methods selected nine, 10, and seven clusters, respectively. The modularized methods selected the most robust cluster groupings with each finding two large clusters of size approximately 47% and 42% and one small cluster of size 11%. The BMMG and BTMPK picked up a few clusters composed of outlying observations. However, the grouping were largely different. The two largest BMMG clusters accounted for approximately 88% with all other clusters ranging from 0.4% to 4.6%. Three BTMPK latent classes were approximately balanced in size at 30%, 24.2%, and 21.5% with the rest ranging from 0% to 9.8%. The BMPK

latent classes ranged in distributional breakdown from 0.5% to 13.1%. The supervised clustering seemed useful for re-allocating singleton clusters in the BTMPK.

Table 5.16: Summary of number of latent classes selected by method.

Method	Unsupervised	Supervised
BTMPK	10	7
MOTEF	3	3
BMPK	18	18
BMMG	9	

5.5.10 Time-budget centroids

The derived accelerometry-accessed weekday and weekend sedentary behaviors and physical activity profiles had some commonalities. In the descriptions that follow, all time budget centroids are described out of 16 hours of wear time. Most latent classes identified had weekend sedentary time remain the same or increase from week day sedentary time across all methods (Tables 5.25 - 5.31). The few clusters whose sedentary time notably decreased from weekday to weekend, decreased by approximately 20 and 40 minutes in BTMPK cluster 6 (n=141) and BMPK cluster 3 (n=22), respectively. Additionally, there were some cluster profiles seemingly present across all methods except for BMPK. For instance, the MOTEF (clusters 3, n=1150;1127 supervised), BTMPK (cluster 8 n=650, supervised cluster 7 n=649), and the BMMG (cluster 6, n=714) methods each derived a cluster with approximately 25+ moderate to vigorous physical activity and 11.5 hours of sedentary time in weekday and weekend days. However, the size of this cluster varied greatly.

The MOTEF methods observed the most interpretable and robust latent classes with clear incremental physical activity pattern distinctions. Tables 5.25 and 5.26 display summaries of the weekday and weekend time budget centroids for the MOTEF methods. Similar trends are observed in both the unsupervised and supervised derived latent classes. Cluster 1 (n=318;319 supervised) is the most sedentary group with only two minutes of moderate weekday physical activity, little to no vigorous activity, and approximately two hours of light physical activity time. The second clusters (n=1486;1505 supervised) have approximately 10 and 5 minutes of weekday and weekend moderate physical activity, respectively, with very few seconds of vigorous physical activity. The third clusters (n=1150;1127 supervised) are the most active groups with close to 4 hours, more than 20 minutes, and 2 minutes of light, moderate, and vigorous physical activity, respectively, during both the weekday and weekend.

The BTMPK methods, discovered more which depict greater variation in the weekday and weekend moderate to vigorous physical activity. For instance, clusters 5 (n=47) and 3 (n=47) in the unsupervised and adiposity supervised, respectively, are clusters with the same profile characterized by displacement of

all weekday MVPA and 40+ minutes of LPA into sedentary weekend time. Tables 5.27 and 5.28 display centroid summaries for the unsupervised and supervised derived latent classes. BTMPK methods additionally discovered another common cluster (cluster 6, n=141; cluster 5, n=135 supervised) characterized by a displacement of approximately 18 minutes of weekday sedentary time into 10 weekend LPA and 8 weekend MVPA minutes.

The supervised BTMPK profiles seemed more robust to outliers relative to the unsupervised profiles. The unsupervised profiles identified clusters 1 and 2 as singleton clusters with unusual physical activity patterns. Cluster 1 observed no moderate weekend physical activity but some VPA while cluster 2 observed the same pattern during the weekday. Additionally, the unsupervised method identified two clusters, clusters 9 (n=130) and 10 (n=83), who had much higher weekday LPA at more than 6.5 hours while weekday LPA was at most just under 4 hours in the other clusters. The supervision was successful in collapsing reallocating the outlying centroid latent classes and thereby reducing the number of latent classes from 10 to 7.

The BMMG method selected clusters with few zero pattern configurations in difference to the BTMPK but seemed to pick up multiple outliers (Table 5.29). For instance, cluster 1 (n=3) is composed of highly sedentary individuals with only about 20 minutes of light activity thought the week. On the other end, cluster 9 (n=11) was estimated to have more time in light PA at almost 10 hours with just shy of six hours of sedentary time. Cluster 8 (n=37) had the largest increase in with 4 more hours of sedentary time in the weekend versus the weekend, which was largely displaced from weekday light activity. Cluster 7 (n=3) was an outlier with individuals estimated to have close to an hour of moderate PA during the week and 25 minutes in the weekend. The largest cluster (n=1867), cluster 4, was estimated to have sedentary behavior and physical activity similar to the overall time budget with just over 12 hours of sedentary time, 3.5 hours of light activity, and 10 minutes of MVPA.

The BMPK methods derived many clusters with little discernible differences but was successful in picking out highly physically active groups. For instance, in the unsupervised approach, clusters 15 – 18 were high active with 36, 43, 85, and 52 weekday MVPA minutes. Cluster 15, however, had a noticeable decrease in weekend MVPA to 14 minutes. The rest of the clusters had somewhat sustained MVPA time. Clusters 17 (n=88;87 supervised) had remarkable time in vigorous physical activity at 27+ and 24+ minutes in the weekday and weekend, respectively. See Tables 5.30 and 5.31.

5.5.11 Associations with adiposity

Adiposity generally decreased with higher levels of physical activity as expected. Figures 5.34 and 5.32 display confidence interval plots of estimated group adiposity prevalence by method and latent class. Note, the

cluster labels correspond to ranks (order statistics) of the corresponding group centroid weekday sedentary time. All methods had at least two non-overlapping 95% confidence intervals where the profiles of the different latent classes were characterized by different levels of physical activity. For instance, in the MOTEF analysis, cluster 3 observed 25+ MVPA weekday and weekend minutes and estimated adiposity estimate of 0.42 (95% ci, 0.38 – 0.45) while clusters 1 and 2 observed less than 11 MVPA weekday and weekend minutes and estimated adiposity prevalences of 0.76 (95% ci, 0.67 – 0.84) and 0.65 (95% ci, 0.60 – 0.69), respectively. On the other end, for the supervised BMPK analysis, as an example, one of its most physically active latent class (cluster 16) had 25+ weekday and weekend MVPA minutes with an observed adiposity prevalence of 0.32 (95% ci, 0.25 – 0.39) compared to 0.91 and 0.99 (95% ci, 0.88 – 1.00) and 0.85 (95% 0.69 – 0.92) in the least physically active clusters.

Figures 5.33 and 5.31 display summaries of Tukey adjusted all pairwise comparisons of adiposity by method and latent class membership. Note, pairwise significance was influenced by cluster size and estimated adiposity in the unsupervised BTMPK and BMMG methods. A lot of the significance in the pairwise comparisons for these methods were identified because a cluster had an extreme value estimate in adiposity. For instance, cluster 2 in unsupervised BTMPK and cluster 7 in BMMG had estimated adiposity of 1. Additionally, cluster 1 in BTMPK was a singleton cluster with an estimated adiposity of 0.00. The BMPK method had the smallest rate of significant findings out of all pairwise comparisons at 34/153 while the MOTEF supervised method determined adiposity significantly differed among all the groups at 3/3 pairwise comparisons. The BTMPK, BMMG, and unsupervised MOTEF methods each had 22/45, 12/36, and 2/3 pairwise comparisons flagged as significant, respectively. The supervised BTMPK observed 9/21 significant pairwise comparisons.

The supervised MOTEF method produced the clearest associations between adiposity and latent classes (Table 5.26). The inverse relationship between physical activity and adiposity is clear with estimated adiposity of 0.83 (0.76 – 0.88), 0.69 (0.64 – 0.73), and 0.36 (0.32 – 0.39) for clusters 1 - 3 which had approximately 1 hour differences in weekday and weekend sedentary time. Another way to look at the characterization of the physical activity is, clusters 1 - 3 had estimated 12, 65, and 211 minutes of moderate physical activity per week, respectively. Note, by construction, time budgets centroid summaries by weekday and weekend depict a typical day. In the summary tables used for characterizing the profiles, these are extrapolated to 16 hours which represent wake hours. In a similar manner, the centroids can be extrapolated to depict approximate week long estimates of duration in each activity intensity. This depiction is most useful for comparison with the US recommendation guidelines of physical activity as 150 minutes of MPA or 75 minutes of VPA per week or a combination of the two. Interestingly, only cluster 3 appeared to achieve the American Heart Association physical activity recommendation of 150 minutes of MPA, and it had the lowest prevalence of adiposity.

The proposed supervised BTMPK method provides another clear relationship between physical activity and adiposity. Clusters 2, 1, and 4 had the largest prevalences of adiposity at 0.93, 0.77, and 0.69 and the least estimated amount of estimate MPA per week at only 8, 12, and 50 minutes, respectively. Clusters 5 and 3 had moderate amounts of MPA each with 98 minutes of MPA per week and estimate adiposity prevalences of 0.59 and 0.52, respectively. A feature of jointly modeling the weekday and weekend time budgets is that it allows one to see how participants distribute their time activity intensities. Clusters 5 and 3 with seemingly similar week long MPA, have different routes to this accumulation where cluster 3 accumulates all MVPA during the weekdays while cluster 5 accumulate MVPA through out the week (Table 5.28). Clusters 6 and 7 had the greatest amount of MPA minutes per week at 132 and 268 with adiposity prevalences of 0.47 and 0.38, respectively. The associations observed with the latent classes support recent findings where similar associations between accelerometry-assessed MVPA were observed in US, Greenland, and UK cohorts (Murabito et al., 2015; Dahl-Petersen et al., 2017; McGrath et al., 2017; Myers et al., 2018; Moon et al., 2017).

5.6 Discussion

In this chapter, we have proposed a Bayesian joint mixture model for compositional data with essential zeros for profiling of physical activity and health risk via the tensor mixture of product kernels. The tensor mixture methodology together with the principal of working in the coordinates allowed the building of a mixture modeling framework that is the first of a kind to be able to account for essential zeros in compositional data for subclass identification. The methodology leveraged a sparse mixture model specification for automatic deletion of redundant latent classes thereby automating cluster selection. We found the proposed methodology outperforms competitors in mixtures of product kernels and multivariate Gaussian adapted to compositional data with imputation of zeros. We additionally developed a Matlab package that can implement the joint mixture models presented in this chapter, and was combined with capability to for including mixed-scale variables (i.e. continuous, count, categorical).

Although the proposed methodology performed well in certain scenarios, there are a few opportunities for improvement. The first, is to conduct further investigation to understand scenarios where the proposed methodology has a loss of information. The second, is to adequately extend the mixture methodology to account for repeated and correlated observations. Lastly, it may be useful to build in a model based adjustment for complex survey designs.

APPENDIX A: SIMULATION DETAILS FOR CHAPTER 4

The purpose of this section is to elucidate details of the details of the simulation experiments for Bayesian semi-parametric modeling with variable selection.

The predictors and confounding variables are assumed to have a simple mixed-scale distribution:

$$f(\boldsymbol{x}_i, \boldsymbol{z}_i) = \int_{R(\boldsymbol{z}_i)} f(\boldsymbol{x}_i, \boldsymbol{z}_i^*) d\boldsymbol{z}_i^*, \tag{5.77}$$

where $R(\boldsymbol{z}_i) = \{\boldsymbol{z}_i^* \in \mathbb{R}^{p_0+1} : z_{ij} = g(z_{ij}^*), j = 1, \ldots, p_0 + 1\}$ and $(\boldsymbol{x}_i, \boldsymbol{z}_i^*) \sim \mathrm{N}_{p+p_0+1}(\mathbf{0}, \boldsymbol{\Sigma})$. We impose the following functions to induce the different scales on the confounding variables:

- indicator variables coded for three level nominal variable: $z_{ij} = I(z_{i,j+1}^* = \max_{1 \le j \le 3} z_{ij}^*)$, for $j = 1, 2$.
- binary variables: $z_{ij} = I(z_{i,j+1} > a_j)$ for $j = 3, \ldots, 6, 8$ where $(a_3, \ldots, a_6, a_8)$ = (1.2816, 0.2533, -0.5244, -0.8416, 1/3).
- bounded ordinal: $z_{ij} = (\sum_{r=0}^{31} (r + 14) I(a_{jr} < 10 * z_{i,j+1}^* + 25 \le a_{j,r+1}) - 25)/10$ where $j = 7$, $a_{jr} = 14 + r$ for $r = 1, \ldots, 31$, and set $a_{j0}, a_j = -\infty, \infty$.
- continuous: $z_{ij} = z_{i,j+1}^*$ for $j = 9, 10$.

Note, continuous and bounded ordinal variables are scaled by 2 to facilitate comparison with the indicator variables to used in modeling with mixed-scale confounding variables (as done and suggested in Buckley et al. (2015) and Gelman et al. (2013), respectively). Further, we let

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_X} & \boldsymbol{\Sigma}'_{\boldsymbol{XZ}} \\ \boldsymbol{\Sigma_{XZ}} & \boldsymbol{\Sigma_Z} \end{bmatrix} \tag{5.78}$$

the covariance matrix is partitioned by sub-components corresponding to predictor, confounder, and predictor-confounder covariance matrices. Setting $\boldsymbol{\Sigma_{XZ}}$ to a non-zero matrix ensures association between the predictor and confounding variables.

Below we elucidate correlation structures used for generating data in all scenarios unless otherwise detailed in Table 4.4.

$$\boldsymbol{\Sigma_X} = \begin{bmatrix} 1.00 & 0.52 & 0.63 & 0.68 & 0.84 \\ 0.52 & 1.00 & 0.92 & 0.77 & 0.51 \\ 0.63 & 0.92 & 1.00 & 0.85 & 0.54 \\ 0.68 & 0.77 & 0.85 & 1.00 & 0.59 \\ 0.84 & 0.51 & 0.54 & 0.59 & 1.00 \end{bmatrix} \tag{5.79}$$

$$\boldsymbol{\Sigma_Z} = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.1 & 0.2 & 0.0 & 0.1 & -0.1 & 0.0 & -0.1 & 0.1 \\ 0.0 & 1.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & -0.1 & 0.0 & 0.1 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.2 & 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & -0.1 \\ 0.1 & 0.1 & 0.0 & 1.0 & -0.1 & -0.1 & 0.0 & -0.1 & -0.1 & 0.0 & 0.1 \\ 0.2 & 0.1 & 0.2 & -0.1 & 1.0 & 0.1 & -0.1 & 0.0 & 0.1 & 0.0 & -0.1 \\ 0.0 & 0.1 & 0.0 & -0.1 & 0.1 & 1.0 & -0.1 & -0.1 & 0.1 & 0.0 & -0.1 \\ 0.1 & 0.1 & 0.1 & 0.0 & -0.1 & -0.1 & 1.0 & 0.1 & 0.1 & 0.0 & 0.0 \\ -0.1 & -0.1 & 0.1 & -0.1 & 0.0 & -0.1 & 0.1 & 1.0 & 0.1 & 0.0 & 0.1 \\ 0.0 & 0.0 & 0.1 & -0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1.0 & 0.0 & -0.1 \\ -0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.2 \\ 0.1 & 0.0 & -0.1 & 0.1 & -0.1 & -0.1 & 0.0 & 0.1 & -0.1 & 0.2 & 1.0 \end{bmatrix} \tag{5.80}$$

$$\boldsymbol{\Sigma_{XZ}} = \begin{bmatrix} -0.1 & -0.1 & -0.1 & -0.1 & -0.1 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.3 & -0.3 & -0.3 & -0.3 & -0.3 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix} \tag{5.81}$$

Table 5.17: Summary of data set characteristics by simulation scenario, type of confounding variables used, and true outcome standard deviations. Each simulation scenario configuration (row) is composed of 500 data sets each with 500 observations. Summary measures display mean and ($2.5^{th} - -97.5^{th}$) percentiles across all data sets within the row configuration. R^2 and R^2_{adj} correspond to mean of estimated coefficient of determination values for each data set computed from a model with only the model terms used in the generation. Sensitivity displays the mean of the proportions of true terms detected (excluding confounding variables where applicable) by the fitted model for each data set.

Scenario	Confounders	σ	R^2	R^2_{adj}	Sensitivity
1	none	13.5	0.10 (0.05 – 0.15)	0.09 (0.04 – 0.15)	0.11 (0.00 – 0.40)
1	none	8.9	0.20 (0.14 – 0.26)	0.19 (0.13 – 0.25)	0.19 (0.00 – 0.40)
1	none	6.8	0.29 (0.23 – 0.36)	0.29 (0.23 – 0.35)	0.29 (0.00 – 0.60)
1	continuous	13.5	0.11 (0.06 – 0.16)	0.08 (0.03 – 0.14)	0.10 (0.00 – 0.40)
1	continuous	8.9	0.19 (0.14 – 0.26)	0.17 (0.11 – 0.24)	0.18 (0.00 – 0.40)
1	continuous	6.8	0.28 (0.22 – 0.34)	0.26 (0.19 – 0.32)	0.28 (0.00 – 0.60)
1	mixed-scale	8	0.10 (0.06 – 0.15)	0.07 (0.03 – 0.12)	0.10 (0.00 – 0.60)
1	mixed-scale	4.7	0.20 (0.14 – 0.26)	0.17 (0.11 – 0.24)	0.17 (0.00 – 0.40)
1	mixed-scale	3.5	0.30 (0.23 – 0.36)	0.27 (0.20 – 0.34)	0.28 (0.00 – 0.60)
2	none	17.9	0.10 (0.05 – 0.16)	0.09 (0.04 – 0.15)	0.10 (0.00 – 0.38)
2	none	11.9	0.20 (0.13 – 0.27)	0.18 (0.11 – 0.25)	0.19 (0.00 – 0.50)
2	none	8.7	0.31 (0.23 – 0.39)	0.30 (0.22 – 0.38)	0.31 (0.00 – 0.50)
2	continuous	21	0.10 (0.06 – 0.16)	0.07 (0.02 – 0.12)	0.09 (0.00 – 0.38)
2	continuous	12.5	0.20 (0.13 – 0.26)	0.17 (0.10 – 0.24)	0.17 (0.00 – 0.38)
2	continuous	9	0.31 (0.24 – 0.39)	0.28 (0.21 – 0.37)	0.29 (0.00 – 0.50)
2	mixed-scale	8	0.10 (0.06 – 0.15)	0.07 (0.02 – 0.12)	0.08 (0.00 – 0.25)
2	mixed-scale	4.7	0.21 (0.14 – 0.27)	0.18 (0.11 – 0.24)	0.18 (0.00 – 0.38)
2	mixed-scale	3.5	0.30 (0.23 – 0.37)	0.27 (0.20 – 0.35)	0.26 (0.00 – 0.50)
8	none	23.3	0.10 (0.05 – 0.17)	0.09 (0.03 – 0.16)	0.11 (0.00 – 0.33)
8	none	15.2	0.20 (0.12 – 0.30)	0.19 (0.10 – 0.29)	0.20 (0.00 – 0.44)
8	none	11.4	0.30 (0.21 – 0.41)	0.29 (0.19 – 0.40)	0.27 (0.11 – 0.56)
8	continuous	25	0.10 (0.06 – 0.17)	0.07 (0.02 – 0.14)	0.10 (0.00 – 0.33)
8	continuous	15	0.20 (0.13 – 0.29)	0.17 (0.09 – 0.26)	0.19 (0.00 – 0.44)
8	continuous	11	0.30 (0.21 – 0.41)	0.28 (0.18 – 0.39)	0.29 (0.11 – 0.56)
8	mixed-scale	8.3	0.11 (0.06 – 0.16)	0.07 (0.03 – 0.13)	0.08 (0.00 – 0.33)
8	mixed-scale	5.1	0.20 (0.14 – 0.27)	0.17 (0.11 – 0.24)	0.15 (0.00 – 0.33)
8	mixed-scale	3.9	0.29 (0.22 – 0.37)	0.26 (0.18 – 0.34)	0.22 (0.00 – 0.44)

Notable findings from the computation of data characteristics:

- including more predictor terms generally increases model precision thereby requiring a larger outcome standard deviation to achieve a desired R^2 level.
- there is an inverse relationship between outcome standard deviation and R^2.
- including a set of 10 continuous confounding variables does not seem to have a material effect.
- including mixed-scale variables decreases model precision thereby requiring a smaller outcome standard deviation, and has a reduction in sensitivity.

APPENDIX B: SIMULATION RESULTS FIGURES FOR CHAPTER 4

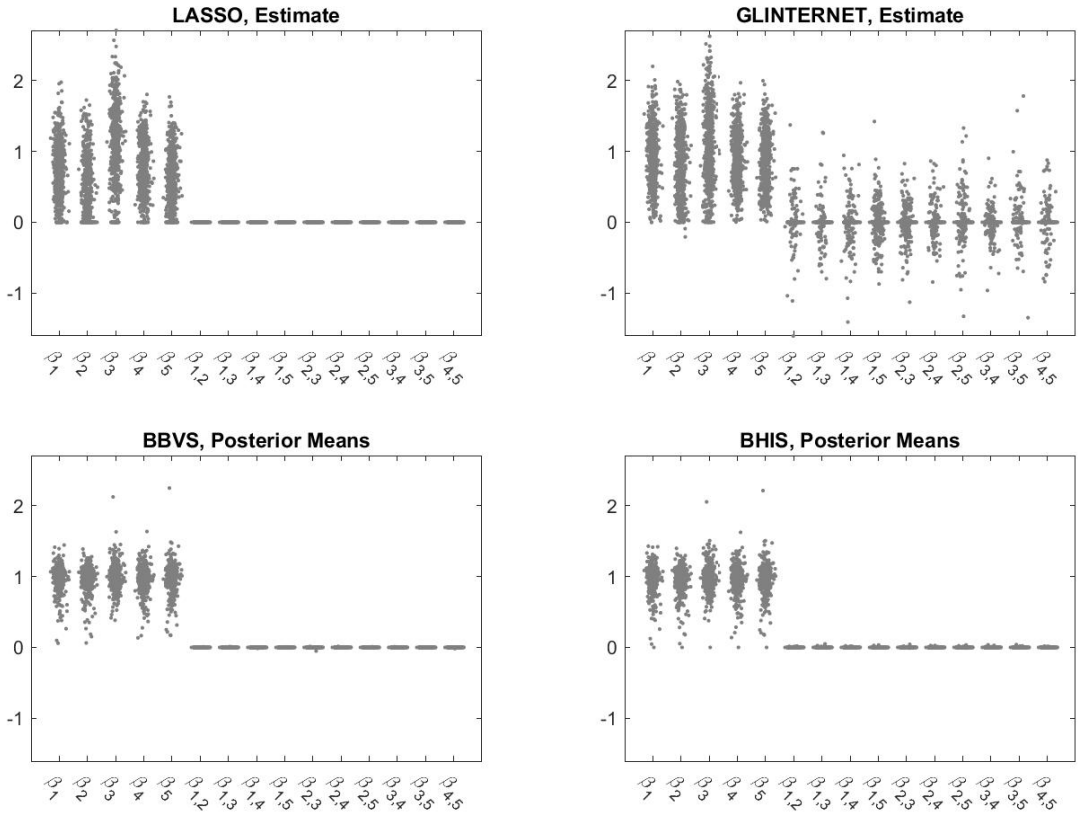

Figure 5.14: Results of simulations for scenario 1 for all 500 data sets each of size 500: Coefficient estimates.

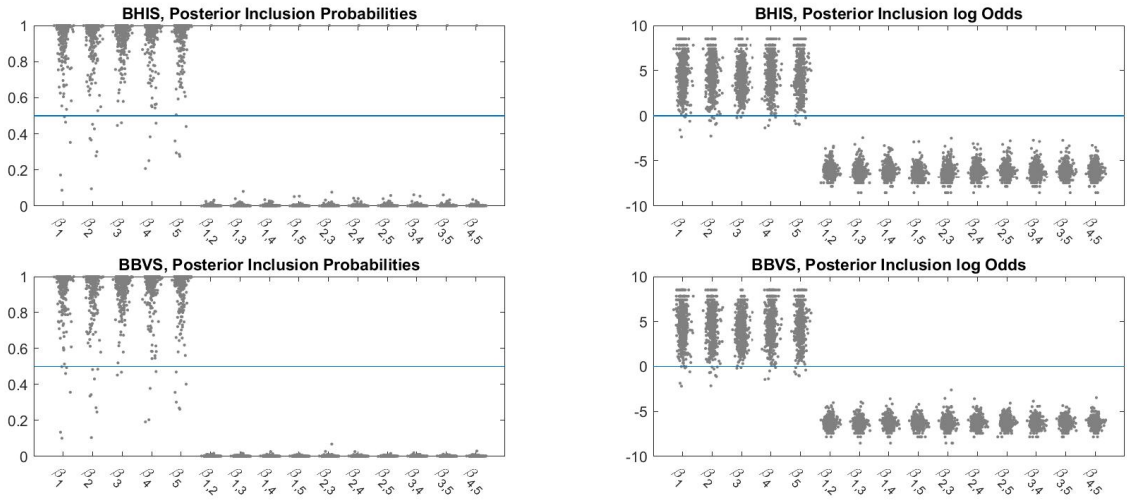

Figure 5.15: Results of simulations for scenario 1 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

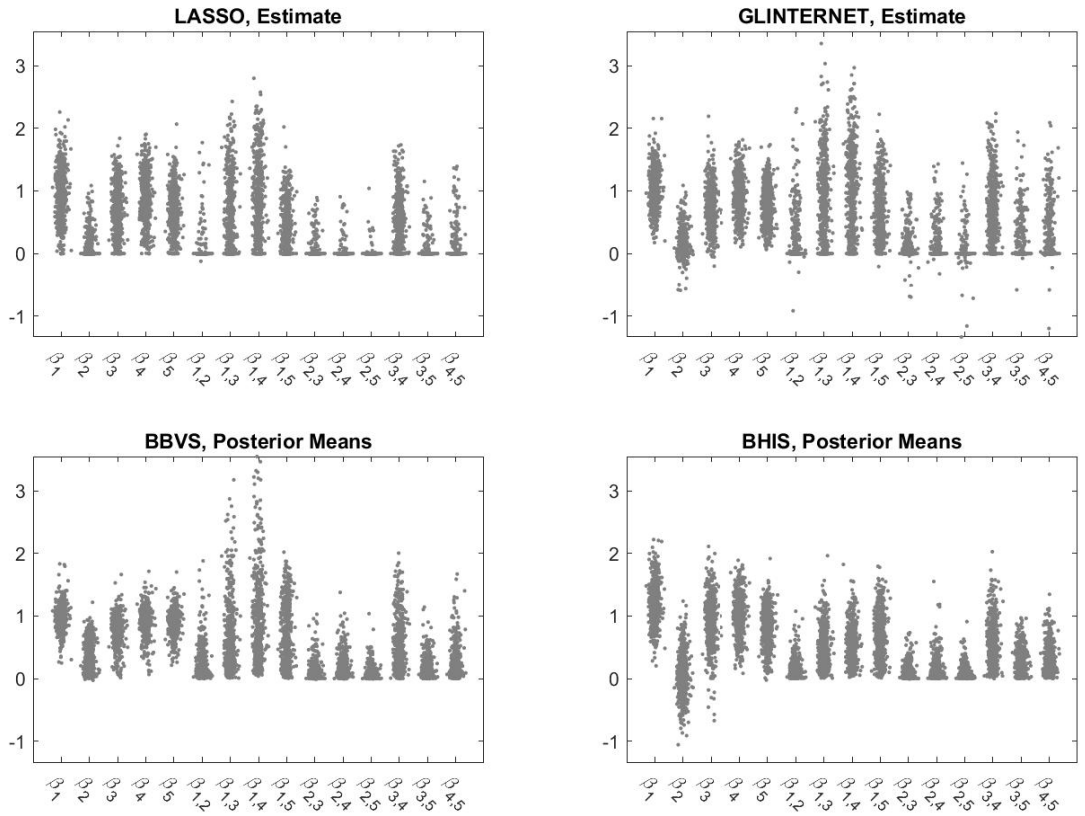

Figure 5.16: Results of simulations for scenario 2 for all 500 data sets each of size 500: Coefficient estimates.

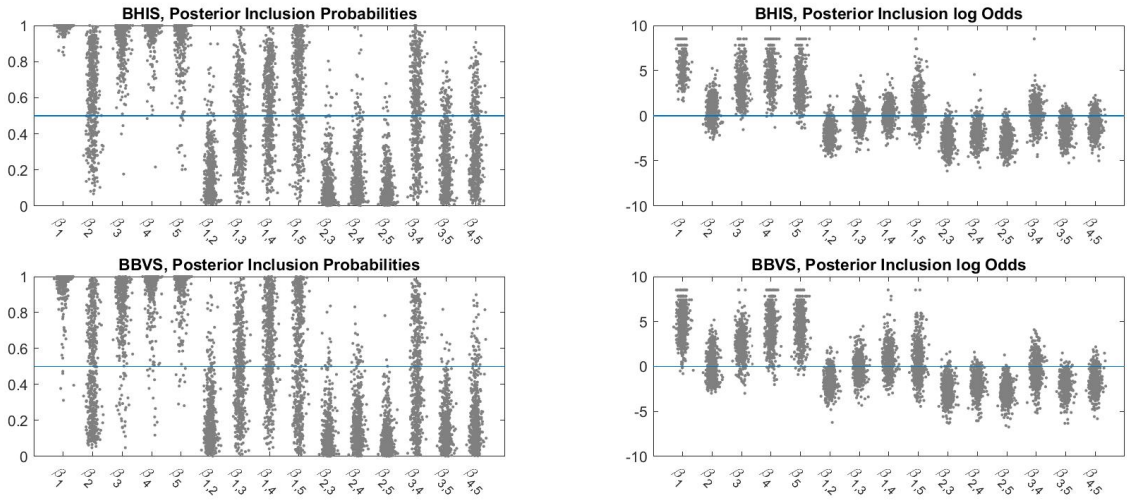

Figure 5.17: Results of simulations for scenario 2 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

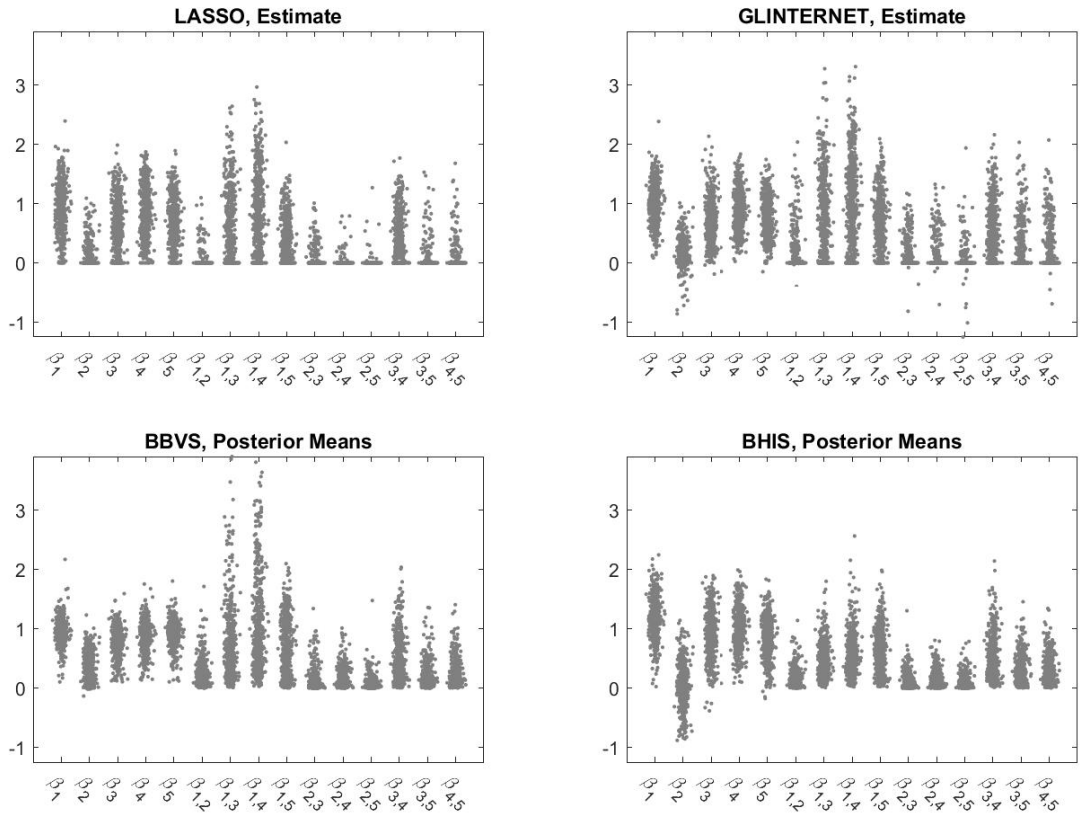

Figure 5.18: Results of simulations for scenario 3 for all 500 data sets each of size 500: Coefficient estimates.

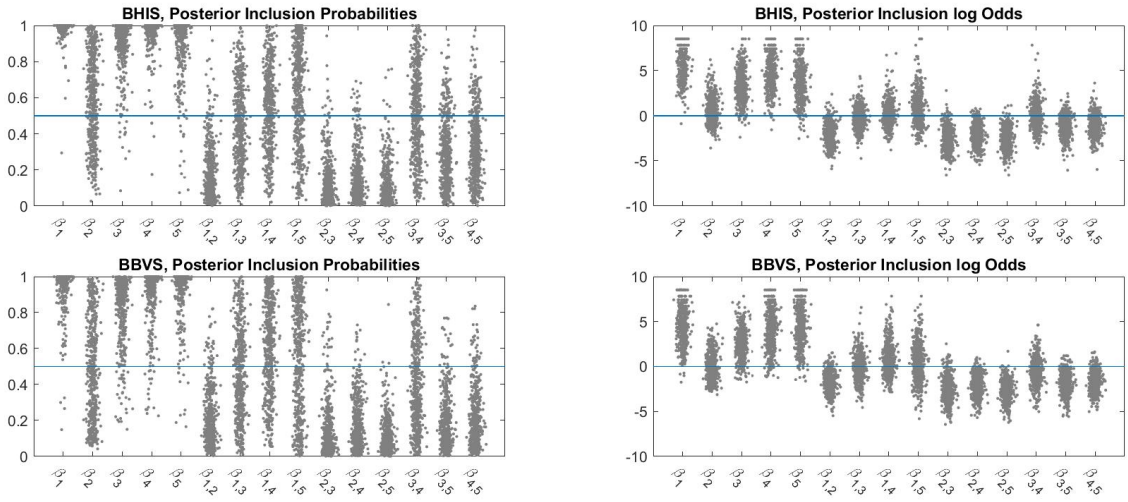

Figure 5.19: Results of simulations for scenario 3 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

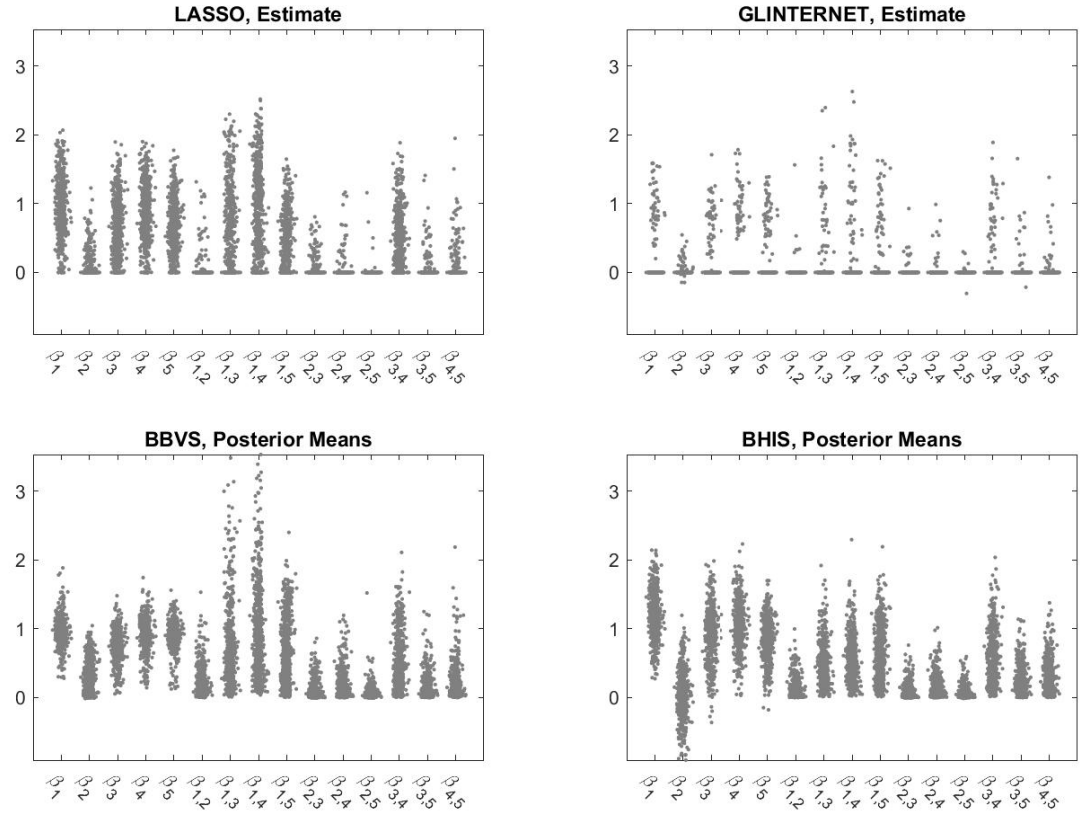

Figure 5.20: Results of simulations for scenario 4 for all 500 data sets each of size 500: Coefficient estimates.

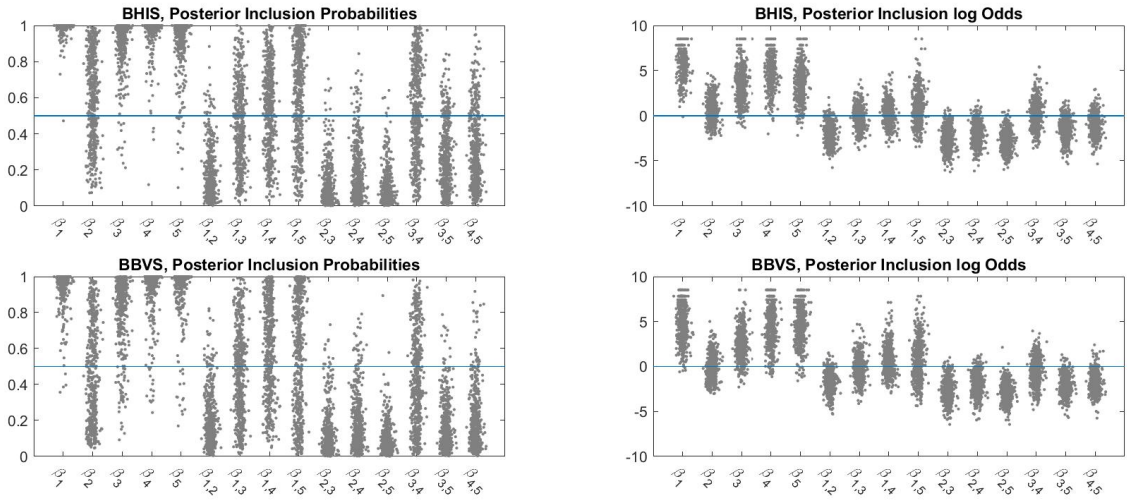

Figure 5.21: Results of simulations for scenario 4 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

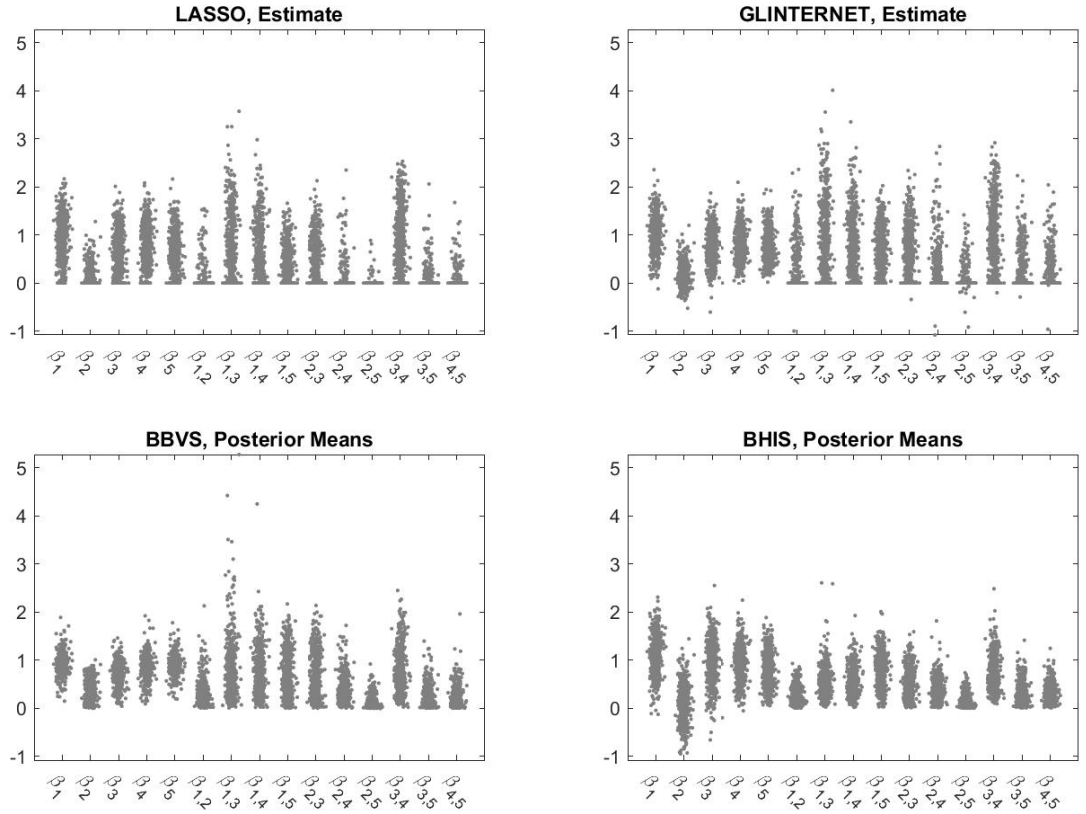

Figure 5.22: Results of simulations for scenario 5 for all 500 data sets each of size 500: Coefficient estimates.

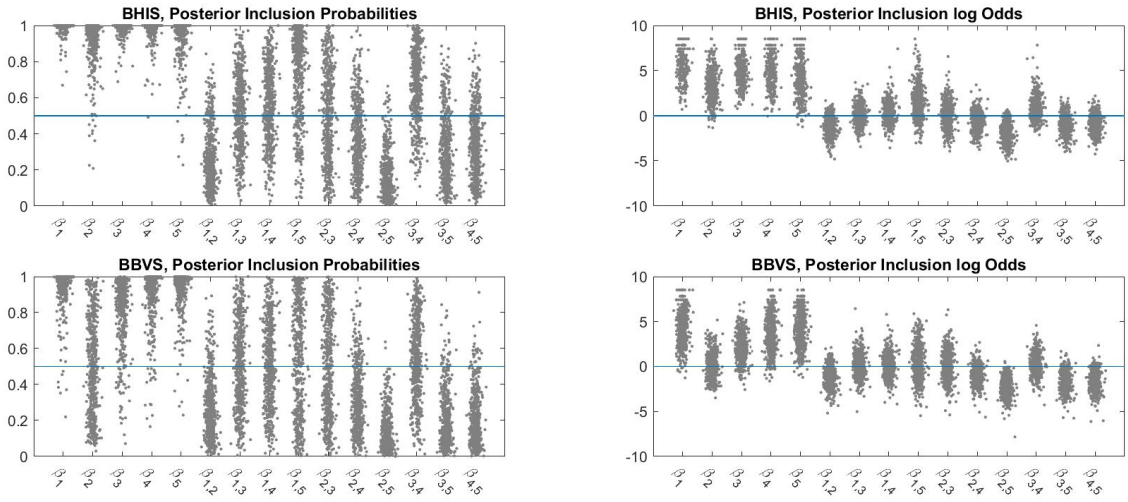

Figure 5.23: Results of simulations for scenario 5 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

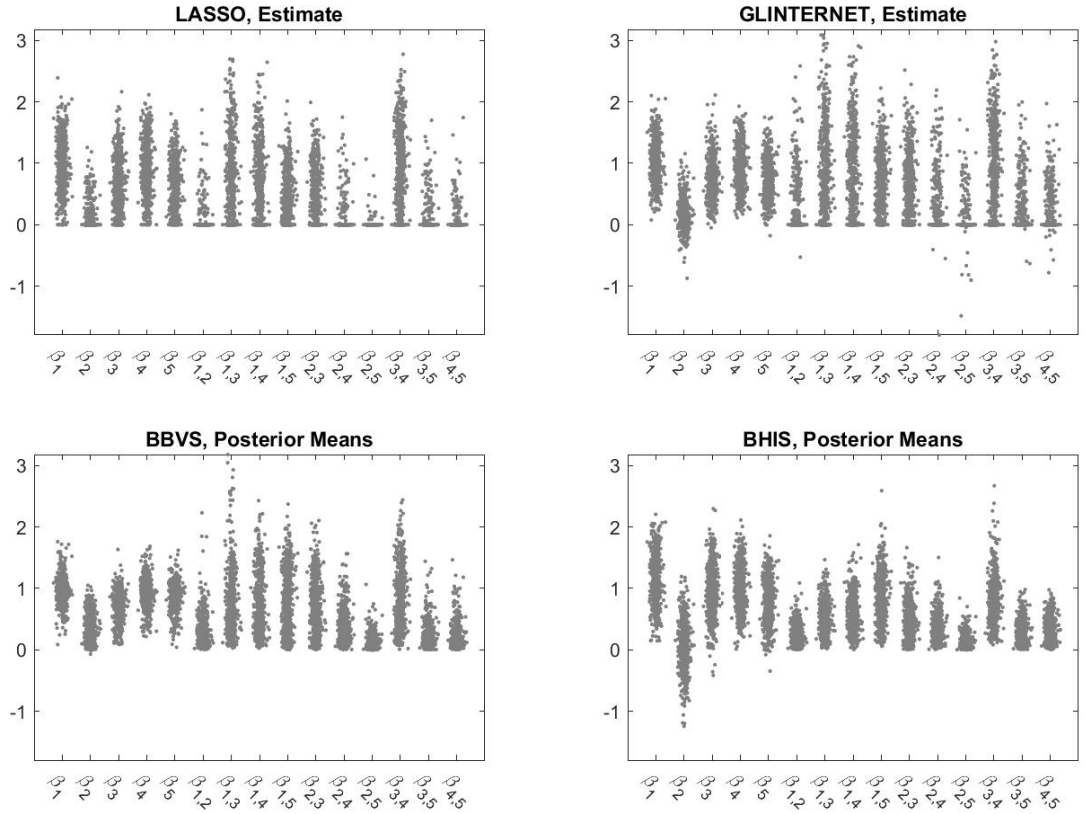

Figure 5.24: Results of simulations for scenario 6 for all 500 data sets each of size 500: Coefficient estimates.

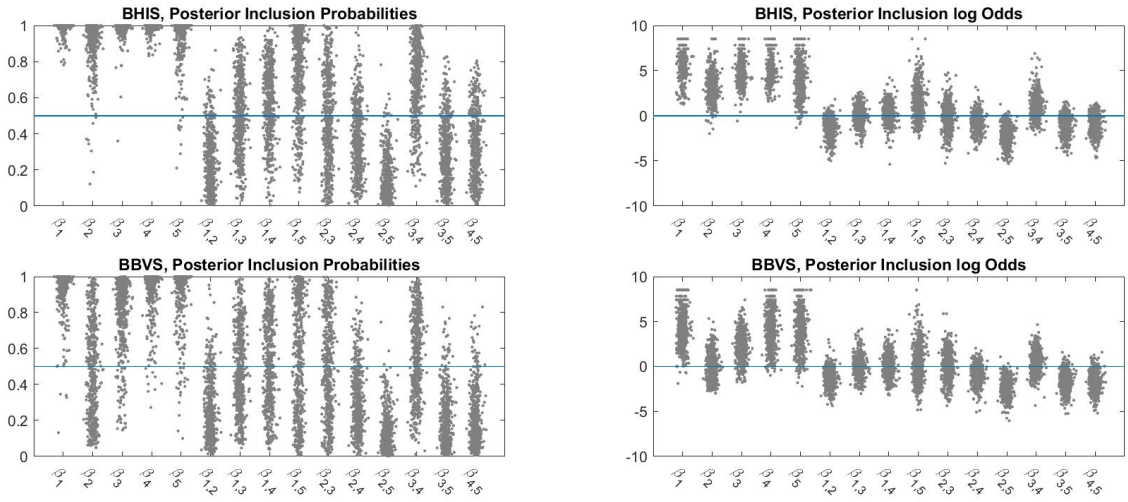

Figure 5.25: Results of simulations for scenario 6 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

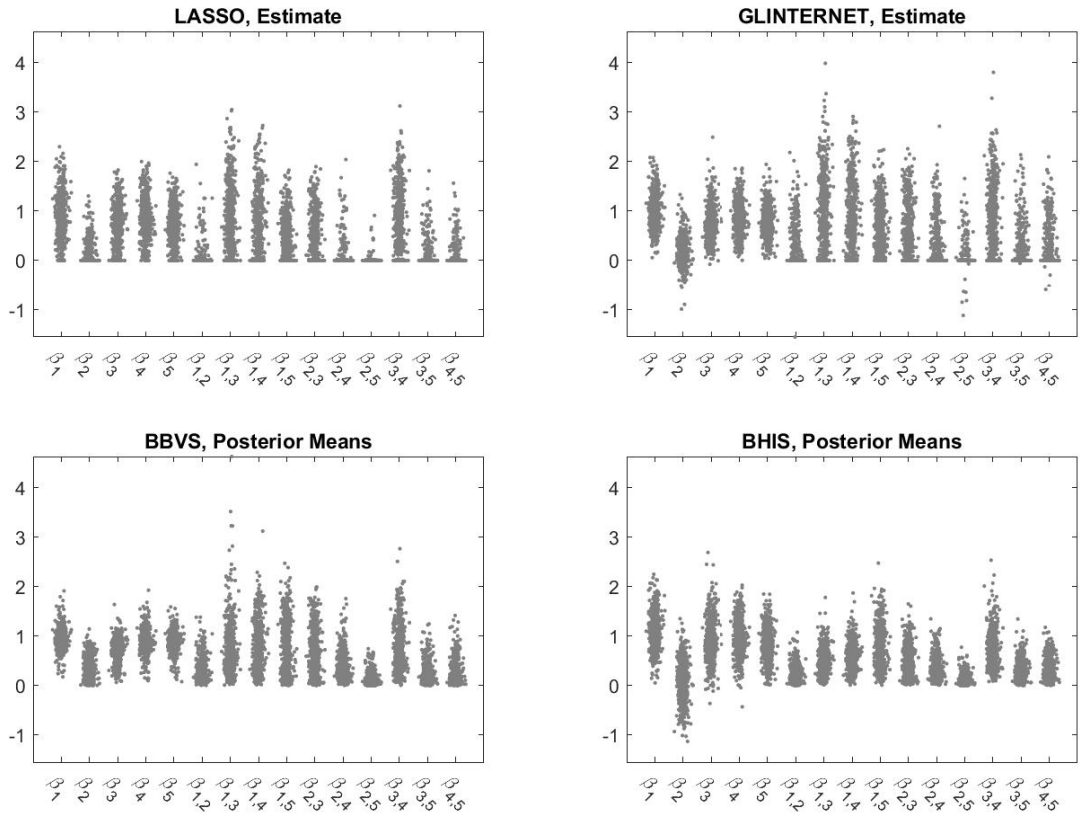

Figure 5.26: Results of simulations for scenario 7 for all 500 data sets each of size 500: Coefficient estimates.

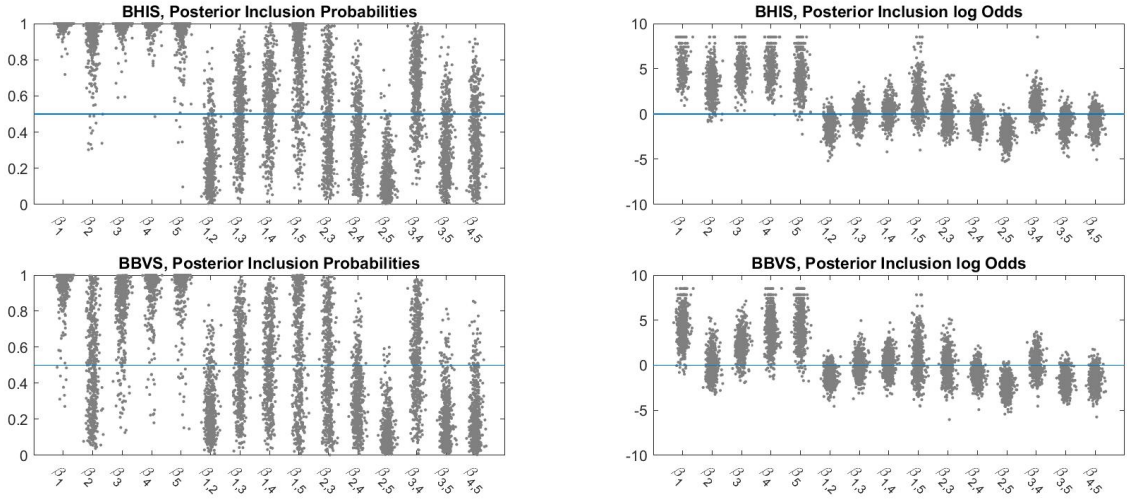

Figure 5.27: Results of simulations for scenario 7 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

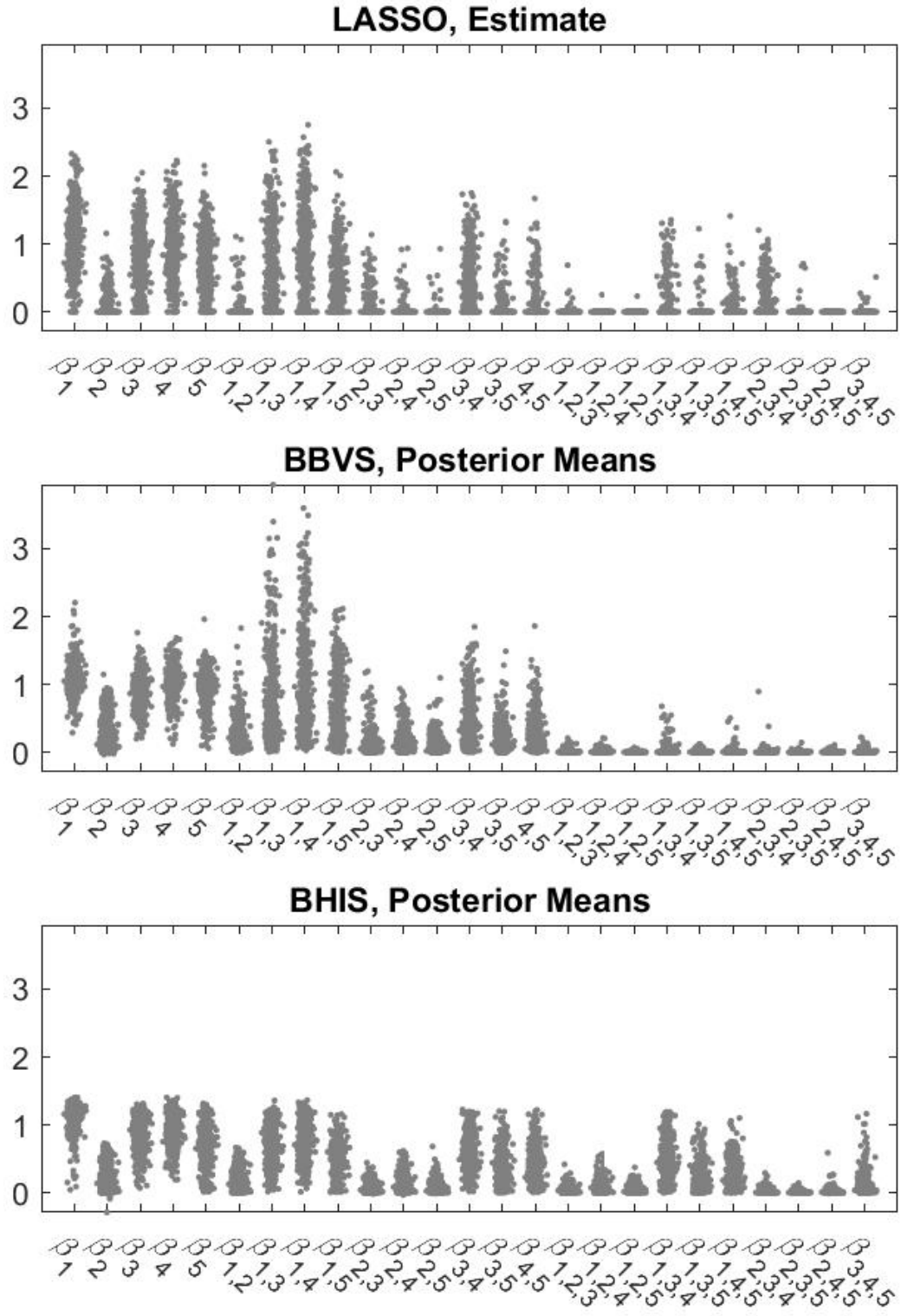

Figure 5.28: Results of simulations for scenario 8 for all 500 data sets each of size 500: Coefficient estimates.

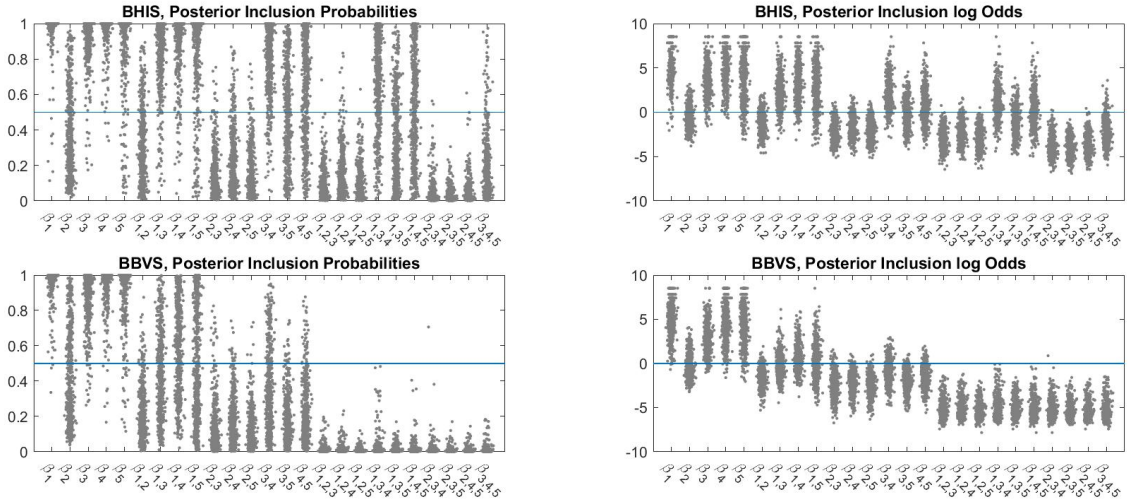

Figure 5.29: Results of simulations for scenario 8 displaying coefficient specific posterior probability and log odds for inclusion for each data set.

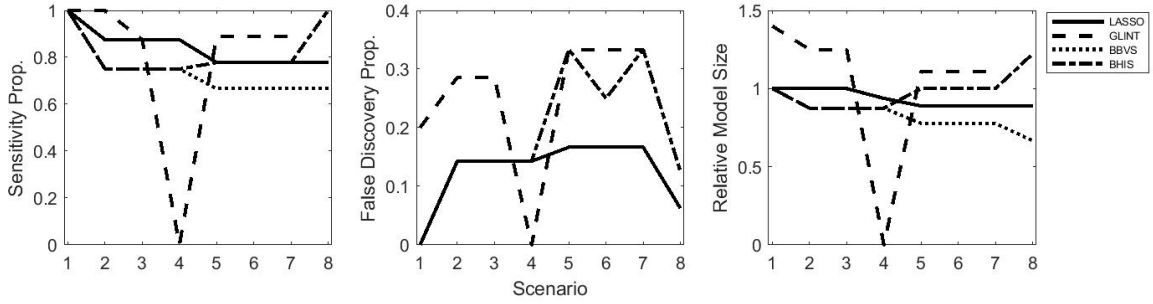

Figure 5.30: Results of all simulation scenarios displaying median of summary measures across all 500 data sets.

APPENDIX C: SIMULATION RESULTS TABLES FOR CHAPTER 4

Table 5.18: Simulation scenario 1 results summary comparing LASSO, GLINTERNET, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		GLINTERNET		BBVS		BHIS	
	true	Median	Incl Prop	Median	Incl Prop	Median		Median	
Coef.	value	Estimate	Sel flag	Estimate	Sel flag	Estimate	Incl Prop	Estimate	Incl Prop
β_1	1	0.72 (0.00 – 1.53)	0.96	0.88 (0.13 – 1.70)	1.00	1.00 (0.58 – 1.23)	0.99	1.00 (0.53 – 1.25)	0.99
β_2	1	0.38 (0.00 – 1.43)	0.75	0.79 (0.01 – 1.76)	0.98	0.99 (0.55 – 1.21)	0.98	0.99 (0.54 – 1.21)	0.99
β_3	1	1.17 (0.00 – 2.11)	0.96	1.04 (0.01 – 2.21)	0.98	0.99 (0.68 – 1.33)	1.00	0.99 (0.65 – 1.33)	1.00
β_4	1	0.78 (0.00 – 1.51)	0.97	0.88 (0.17 – 1.65)	1.00	0.99 (0.55 – 1.26)	0.99	0.99 (0.55 – 1.26)	0.99
β_5	1	0.60 (0.00 – 1.34)	0.94	0.84 (0.19 – 1.59)	1.00	0.98 (0.54 – 1.28)	0.99	0.98 (0.53 – 1.28)	0.99
$\beta_{1,2}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.32 – 0.41)	0.16	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{1,3}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.37 – 0.24)	0.15	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{1,4}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.43 – 0.38)	0.22	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{1,5}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.48 – 0.42)	0.39	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{2,3}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.43 – 0.34)	0.28	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{2,4}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.34 – 0.39)	0.21	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{2,5}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.33 – 0.53)	0.24	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{3,4}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.31 – 0.26)	0.20	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{3,5}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.29 – 0.47)	0.17	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00
$\beta_{4,5}$	0	0.00 (0.00 – 0.00)	0.00	0.00 (-0.39 – 0.37)	0.17	0.00 (-0.00 – 0.00)	0.00	0.00 (-0.00 – 0.00)	0.00

Table 5.19: Simulation scenario 2 results summary comparing LASSO, GLINTERNET, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		GLINTERNET		BBVS		BHIS	
	true	Median	Incl Prop	Median	Incl Prop	Median		Median	
Coef.	value	Estimate	Sel flag	Estimate	Sel flag	Estimate	Incl Prop	Estimate	Incl Prop
β_1	1	0.94 (0.07 – 1.75)	0.98	1.04 (0.41 – 1.64)	1.00	0.96 (0.56 – 1.37)	0.99	1.17 (0.58 – 1.86)	1.00
β_2	0	0.00 (0.00 – 0.68)	0.27	0.08 (-0.16 – 0.75)	0.87	0.27 (0.02 – 0.85)	0.44	0.06 (-0.58 – 0.79)	0.69
β_3	1	0.70 (0.00 – 1.44)	0.93	0.71 (0.09 – 1.40)	1.00	0.81 (0.17 – 1.12)	0.94	0.92 (0.09 – 1.65)	0.99
β_4	1	0.89 (0.11 – 1.64)	0.99	0.90 (0.31 – 1.63)	1.00	0.92 (0.39 – 1.31)	0.98	1.05 (0.35 – 1.65)	1.00
β_5	1	0.70 (0.00 – 1.40)	0.96	0.75 (0.20 – 1.36)	1.00	0.93 (0.43 – 1.28)	0.98	0.90 (0.16 – 1.35)	0.98
$\beta_{1,2}$	0	0.00 (0.00 – 0.71)	0.09	0.00 (0.00 – 1.17)	0.27	0.13 (0.01 – 1.02)	0.06	0.09 (0.01 – 0.56)	0.04
$\beta_{1,3}$	1	0.39 (0.00 – 1.96)	0.66	0.69 (0.00 – 2.20)	0.78	0.55 (0.07 – 2.05)	0.44	0.39 (0.05 – 1.20)	0.41
$\beta_{1,4}$	1	0.81 (0.00 – 2.15)	0.86	0.95 (0.00 – 2.45)	0.87	0.77 (0.11 – 2.79)	0.63	0.52 (0.08 – 1.22)	0.59
$\beta_{1,5}$	1	0.33 (0.00 – 1.22)	0.72	0.70 (0.00 – 1.64)	0.92	0.64 (0.03 – 1.68)	0.56	0.62 (0.06 – 1.44)	0.66
$\beta_{2,3}$	0	0.00 (0.00 – 0.51)	0.12	0.00 (0.00 – 0.82)	0.28	0.04 (0.00 – 0.54)	0.03	0.04 (0.00 – 0.41)	0.02
$\beta_{2,4}$	0	0.00 (0.00 – 0.31)	0.06	0.00 (0.00 – 0.85)	0.21	0.10 (0.00 – 0.70)	0.05	0.09 (0.01 – 0.60)	0.05
$\beta_{2,5}$	0	0.00 (0.00 – 0.01)	0.03	0.00 (-0.11 – 0.62)	0.14	0.04 (0.00 – 0.40)	0.01	0.04 (0.00 – 0.35)	0.01
$\beta_{3,4}$	1	0.36 (0.00 – 1.44)	0.71	0.56 (0.00 – 1.76)	0.81	0.41 (0.02 – 1.55)	0.38	0.49 (0.04 – 1.42)	0.54
$\beta_{3,5}$	0	0.00 (0.00 – 0.55)	0.11	0.00 (0.00 – 1.01)	0.29	0.10 (0.00 – 0.61)	0.02	0.19 (0.01 – 0.74)	0.12
$\beta_{4,5}$	0	0.00 (0.00 – 0.89)	0.16	0.00 (0.00 – 1.35)	0.37	0.14 (0.01 – 0.98)	0.10	0.23 (0.02 – 0.88)	0.20

Table 5.20: Simulation scenario 3 results summary comparing LASSO, GLINTERNET, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		GLINTERNET		BBVS		BHIS	
	true	Median	Incl Prop	Median	Incl Prop	Median		Median	
Coef.	value	Estimate	Sel flag	Estimate	Sel flag	Estimate	Incl Prop	Estimate	Incl Prop
β_1	1	0.93 (0.07 – 1.69)	0.98	1.01 (0.37 – 1.63)	1.00	0.95 (0.52 – 1.35)	0.99	1.17 (0.50 – 1.85)	1.00
β_2	0	0.00 (0.00 – 0.75)	0.29	0.07 (-0.23 – 0.70)	0.85	0.30 (0.01 – 0.85)	0.43	0.06 (-0.71 – 0.84)	0.63
β_3	1	0.67 (0.00 – 1.48)	0.92	0.71 (0.10 – 1.54)	1.00	0.81 (0.22 – 1.18)	0.94	0.95 (0.12 – 1.64)	0.97
β_4	1	0.86 (0.06 – 1.62)	0.98	0.89 (0.29 – 1.64)	1.00	0.91 (0.34 – 1.29)	0.97	1.03 (0.36 – 1.71)	0.99
β_5	1	0.71 (0.00 – 1.45)	0.96	0.75 (0.19 – 1.39)	1.00	0.92 (0.38 – 1.27)	0.98	0.91 (0.12 – 1.50)	0.97
$\beta_{1,2}$	0	0.00 (0.00 – 0.48)	0.08	0.00 (0.00 – 1.05)	0.27	0.12 (0.01 – 0.79)	0.07	0.08 (0.01 – 0.56)	0.05
$\beta_{1,3}$	1	0.33 (0.00 – 1.95)	0.66	0.65 (0.00 – 2.18)	0.74	0.51 (0.06 – 2.57)	0.43	0.39 (0.05 – 1.18)	0.41
$\beta_{1,4}$	1	0.80 (0.00 – 2.25)	0.82	0.95 (0.00 – 2.49)	0.90	0.82 (0.10 – 2.97)	0.65	0.50 (0.08 – 1.40)	0.59
$\beta_{1,5}$	1	0.33 (0.00 – 1.25)	0.73	0.73 (0.00 – 1.65)	0.91	0.67 (0.02 – 1.56)	0.57	0.58 (0.05 – 1.44)	0.65
$\beta_{2,3}$	0	0.00 (0.00 – 0.53)	0.15	0.00 (0.00 – 0.76)	0.27	0.05 (0.00 – 0.71)	0.04	0.04 (0.00 – 0.44)	0.03
$\beta_{2,4}$	0	0.00 (0.00 – 0.19)	0.05	0.00 (0.00 – 0.83)	0.20	0.09 (0.01 – 0.60)	0.03	0.07 (0.00 – 0.51)	0.04
$\beta_{2,5}$	0	0.00 (0.00 – 0.00)	0.02	0.00 (0.00 – 0.52)	0.14	0.04 (0.00 – 0.37)	0.01	0.04 (0.00 – 0.38)	0.01
$\beta_{3,4}$	1	0.29 (0.00 – 1.33)	0.66	0.53 (0.00 – 1.58)	0.80	0.35 (0.02 – 1.46)	0.33	0.46 (0.05 – 1.34)	0.49
$\beta_{3,5}$	0	0.00 (0.00 – 0.74)	0.12	0.00 (0.00 – 1.27)	0.32	0.10 (0.01 – 0.72)	0.04	0.21 (0.01 – 0.85)	0.16
$\beta_{4,5}$	0	0.00 (0.00 – 0.78)	0.14	0.00 (0.00 – 1.16)	0.35	0.13 (0.01 – 0.83)	0.08	0.25 (0.02 – 0.84)	0.18

Table 5.21: Simulation scenario 4 results summary comparing LASSO, GLINTERNET, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		GLINTERNET		BBVS		BHIS	
	true	Median	Incl Prop	Median	Incl Prop	Median		Median	
Coef.	value	Estimate	Sel flag	Estimate	Sel flag	Estimate	Incl Prop	Estimate	Incl Prop
β_1	1	0.97 (0.05 – 1.78)	0.98	0.00 (0.00 – 1.18)	0.10	0.96 (0.54 – 1.39)	0.99	1.19 (0.53 – 1.89)	1.00
β_2	0	0.00 (0.00 – 0.61)	0.27	0.00 (0.00 – 0.16)	0.07	0.27 (0.01 – 0.84)	0.43	0.04 (-0.65 – 0.76)	0.64
β_3	1	0.72 (0.00 – 1.55)	0.93	0.00 (0.00 – 0.97)	0.10	0.81 (0.21 – 1.13)	0.92	0.94 (0.11 – 1.68)	0.98
β_4	1	0.90 (0.00 – 1.67)	0.97	0.00 (0.00 – 1.20)	0.10	0.92 (0.37 – 1.33)	0.97	1.06 (0.34 – 1.71)	0.99
β_5	1	0.67 (0.00 – 1.46)	0.96	0.00 (0.00 – 1.04)	0.10	0.92 (0.36 – 1.29)	0.97	0.87 (0.21 – 1.40)	0.97
$\beta_{1,2}$	0	0.00 (0.00 – 0.55)	0.07	0.00 (0.00 – 0.00)	0.01	0.12 (0.01 – 0.84)	0.06	0.08 (0.01 – 0.55)	0.04
$\beta_{1,3}$	1	0.42 (0.00 – 2.01)	0.65	0.00 (0.00 – 1.20)	0.07	0.57 (0.05 – 2.41)	0.46	0.39 (0.05 – 1.23)	0.42
$\beta_{1,4}$	1	0.84 (0.00 – 2.15)	0.84	0.00 (0.00 – 1.31)	0.09	0.80 (0.10 – 2.72)	0.64	0.50 (0.07 – 1.33)	0.58
$\beta_{1,5}$	1	0.31 (0.00 – 1.34)	0.71	0.00 (0.00 – 1.15)	0.09	0.65 (0.03 – 1.65)	0.58	0.61 (0.05 – 1.45)	0.66
$\beta_{2,3}$	0	0.00 (0.00 – 0.42)	0.12	0.00 (0.00 – 0.00)	0.02	0.04 (0.00 – 0.50)	0.02	0.04 (0.00 – 0.39)	0.01
$\beta_{2,4}$	0	0.00 (0.00 – 0.41)	0.06	0.00 (0.00 – 0.00)	0.02	0.11 (0.01 – 0.77)	0.06	0.08 (0.00 – 0.60)	0.05
$\beta_{2,5}$	0	0.00 (0.00 – 0.00)	0.02	0.00 (0.00 – 0.00)	0.01	0.03 (0.00 – 0.27)	0.00	0.03 (0.00 – 0.32)	0.01
$\beta_{3,4}$	1	0.31 (0.00 – 1.45)	0.72	0.00 (0.00 – 1.08)	0.09	0.44 (0.02 – 1.43)	0.39	0.49 (0.04 – 1.43)	0.53
$\beta_{3,5}$	0	0.00 (0.00 – 0.50)	0.08	0.00 (0.00 – 0.13)	0.03	0.09 (0.01 – 0.61)	0.03	0.19 (0.01 – 0.78)	0.14
$\beta_{4,5}$	0	0.00 (0.00 – 0.65)	0.14	0.00 (0.00 – 0.18)	0.04	0.12 (0.01 – 0.97)	0.07	0.22 (0.02 – 0.95)	0.19

Table 5.22: Simulation scenario 5 results summary comparing LASSO, GLINTERNET, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		GLINTERNET		BBVS		BHIS	
	true	Median	Incl Prop	Median	Incl Prop	Median		Median	
Coef.	value	Estimate	Sel flag	Estimate	Sel flag	Estimate	Incl Prop	Estimate	Incl Prop
β_1	1	0.95 (0.00 – 1.80)	0.98	0.99 (0.30 – 1.72)	1.00	0.94 (0.50 – 1.37)	0.99	1.12 (0.32 – 1.89)	1.00
β_2	0	0.00 (0.00 – 0.68)	0.32	0.13 (-0.23 – 0.70)	0.97	0.33 (0.02 – 0.81)	0.49	0.12 (-0.65 – 0.88)	0.98
β_3	1	0.65 (0.00 – 1.49)	0.90	0.74 (0.11 – 1.42)	1.00	0.78 (0.24 – 1.17)	0.92	0.94 (0.03 – 1.80)	1.00
β_4	1	0.84 (0.00 – 1.63)	0.97	0.84 (0.21 – 1.51)	1.00	0.89 (0.35 – 1.29)	0.97	0.94 (0.23 – 1.62)	1.00
β_5	1	0.69 (0.00 – 1.51)	0.96	0.74 (0.22 – 1.47)	1.00	0.92 (0.49 – 1.30)	0.99	0.85 (0.18 – 1.51)	0.99
$\beta_{1,2}$	0	0.00 (0.00 – 1.01)	0.13	0.00 (0.00 – 1.42)	0.32	0.22 (0.03 – 1.03)	0.10	0.20 (0.02 – 0.69)	0.12
$\beta_{1,3}$	1	0.57 (0.00 – 2.25)	0.72	0.73 (0.00 – 2.66)	0.75	0.62 (0.08 – 2.41)	0.50	0.48 (0.10 – 1.17)	0.50
$\beta_{1,4}$	1	0.39 (0.00 – 2.05)	0.65	0.68 (0.00 – 2.32)	0.78	0.61 (0.09 – 1.83)	0.49	0.56 (0.13 – 1.19)	0.61
$\beta_{1,5}$	1	0.29 (0.00 – 1.25)	0.69	0.70 (0.00 – 1.61)	0.90	0.65 (0.03 – 1.68)	0.56	0.84 (0.10 – 1.49)	0.81
$\beta_{2,3}$	1	0.22 (0.00 – 1.45)	0.60	0.47 (0.00 – 1.75)	0.70	0.54 (0.04 – 1.65)	0.48	0.44 (0.05 – 1.20)	0.46
$\beta_{2,4}$	0	0.00 (0.00 – 0.98)	0.14	0.00 (0.00 – 1.64)	0.35	0.33 (0.03 – 1.04)	0.16	0.31 (0.04 – 0.87)	0.24
$\beta_{2,5}$	0	0.00 (0.00 – 0.15)	0.03	0.00 (0.00 – 0.76)	0.17	0.06 (0.00 – 0.45)	0.00	0.09 (0.01 – 0.45)	0.02
$\beta_{3,4}$	1	0.88 (0.00 – 2.26)	0.84	0.87 (0.00 – 2.46)	0.82	0.70 (0.10 – 1.79)	0.58	0.71 (0.20 – 1.53)	0.80
$\beta_{3,5}$	0	0.00 (0.00 – 0.72)	0.17	0.00 (0.00 – 1.24)	0.35	0.13 (0.01 – 0.80)	0.06	0.23 (0.03 – 0.82)	0.16
$\beta_{4,5}$	0	0.00 (0.00 – 0.56)	0.12	0.00 (0.00 – 1.22)	0.31	0.14 (0.01 – 0.75)	0.05	0.27 (0.04 – 0.74)	0.20

Table 5.23: Simulation scenario 6 results summary comparing LASSO, GLINTERNET, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		GLINTERNET		BBVS		BHIS	
	true	Median	Incl Prop	Median	Incl Prop	Median		Median	
Coef.	value	Estimate	Sel flag	Estimate	Sel flag	Estimate	Incl Prop	Estimate	Incl Prop
β_1	1	0.93 (0.05 – 1.83)	0.98	1.01 (0.36 – 1.74)	1.00	0.95 (0.51 – 1.47)	0.99	1.11 (0.40 – 1.93)	1.00
β_2	0	0.00 (0.00 – 0.72)	0.28	0.11 (-0.28 – 0.76)	0.95	0.30 (0.01 – 0.83)	0.46	0.07 (-0.75 – 0.85)	0.98
β_3	1	0.58 (0.00 – 1.54)	0.90	0.72 (0.14 – 1.46)	1.00	0.76 (0.17 – 1.10)	0.91	0.92 (0.11 – 1.75)	1.00
β_4	1	0.86 (0.03 – 1.77)	0.98	0.87 (0.28 – 1.65)	1.00	0.92 (0.39 – 1.43)	0.98	1.00 (0.33 – 1.76)	1.00
β_5	1	0.65 (0.00 – 1.44)	0.94	0.73 (0.13 – 1.39)	1.00	0.90 (0.34 – 1.30)	0.96	0.82 (0.08 – 1.46)	0.98
$\beta_{1,2}$	0	0.00 (0.00 – 0.76)	0.12	0.00 (0.00 – 1.38)	0.30	0.20 (0.02 – 0.92)	0.10	0.18 (0.02 – 0.62)	0.09
$\beta_{1,3}$	1	0.53 (0.00 – 2.22)	0.70	0.66 (0.00 – 2.49)	0.74	0.56 (0.08 – 2.35)	0.44	0.48 (0.07 – 1.09)	0.50
$\beta_{1,4}$	1	0.40 (0.00 – 1.93)	0.68	0.71 (0.00 – 2.40)	0.76	0.59 (0.08 – 1.82)	0.49	0.56 (0.13 – 1.18)	0.64
$\beta_{1,5}$	1	0.27 (0.00 – 1.28)	0.67	0.71 (0.00 – 1.70)	0.91	0.64 (0.03 – 1.78)	0.56	0.83 (0.13 – 1.66)	0.81
$\beta_{2,3}$	1	0.24 (0.00 – 1.46)	0.61	0.49 (0.00 – 1.76)	0.74	0.54 (0.03 – 1.61)	0.46	0.41 (0.02 – 1.19)	0.44
$\beta_{2,4}$	0	0.00 (0.00 – 0.87)	0.13	0.00 (0.00 – 1.59)	0.34	0.30 (0.03 – 1.06)	0.15	0.30 (0.05 – 0.87)	0.21
$\beta_{2,5}$	0	0.00 (0.00 – 0.09)	0.03	0.00 (0.00 – 0.84)	0.14	0.06 (0.00 – 0.44)	0.01	0.08 (0.00 – 0.44)	0.02
$\beta_{3,4}$	1	0.95 (0.00 – 2.20)	0.84	0.91 (0.00 – 2.45)	0.83	0.74 (0.12 – 1.93)	0.59	0.75 (0.18 – 1.79)	0.83
$\beta_{3,5}$	0	0.00 (0.00 – 0.86)	0.16	0.00 (0.00 – 1.28)	0.34	0.14 (0.01 – 0.75)	0.06	0.24 (0.02 – 0.73)	0.15
$\beta_{4,5}$	0	0.00 (0.00 – 0.61)	0.13	0.00 (0.00 – 1.09)	0.30	0.12 (0.01 – 0.80)	0.06	0.27 (0.03 – 0.77)	0.23

Table 5.24: Simulation scenario 7 results summary comparing LASSO, GLINTERNET, BBVS versus BHIS. Median estimate column displays the median ($2.5^{th} - -97.5^{th}$) percentiles of the posterior mean across all 500 data sets. Inclusion probability column displays the proportion of data sets with estimated log odds greater than 0 flagging for inclusion.

		LASSO		GLINTERNET		BBVS		BHIS	
	true	Median	Incl Prop	Median	Incl Prop	Median		Median	
Coef.	value	Estimate	Sel flag	Estimate	Sel flag	Estimate	Incl Prop	Estimate	Incl Prop
β_1	1	0.91 (0.07 – 1.85)	0.98	0.99 (0.40 – 1.75)	1.00	0.94 (0.55 – 1.39)	0.98	1.10 (0.46 – 1.88)	1.00
β_2	0	0.00 (0.00 – 0.86)	0.28	0.14 (-0.31 – 0.87)	0.96	0.32 (0.02 – 0.85)	0.48	0.10 (-0.74 – 0.88)	0.98
β_3	1	0.67 (0.00 – 1.53)	0.91	0.74 (0.14 – 1.50)	1.00	0.79 (0.21 – 1.12)	0.92	0.93 (0.12 – 1.84)	1.00
β_4	1	0.87 (0.00 – 1.69)	0.97	0.88 (0.21 – 1.57)	1.00	0.91 (0.38 – 1.39)	0.98	1.00 (0.23 – 1.67)	1.00
β_5	1	0.70 (0.00 – 1.47)	0.95	0.76 (0.19 – 1.40)	1.00	0.92 (0.38 – 1.29)	0.97	0.87 (0.18 – 1.48)	0.99
$\beta_{1,2}$	0	0.00 (0.00 – 0.80)	0.11	0.00 (0.00 – 1.29)	0.31	0.22 (0.03 – 0.95)	0.10	0.19 (0.02 – 0.67)	0.09
$\beta_{1,3}$	1	0.62 (0.00 – 2.31)	0.72	0.76 (0.00 – 2.48)	0.77	0.58 (0.08 – 2.18)	0.47	0.49 (0.07 – 1.08)	0.54
$\beta_{1,4}$	1	0.43 (0.00 – 2.26)	0.66	0.69 (0.00 – 2.39)	0.76	0.63 (0.10 – 1.81)	0.53	0.54 (0.13 – 1.19)	0.63
$\beta_{1,5}$	1	0.31 (0.00 – 1.38)	0.70	0.71 (0.00 – 1.66)	0.90	0.70 (0.03 – 1.86)	0.57	0.82 (0.13 – 1.69)	0.81
$\beta_{2,3}$	1	0.30 (0.00 – 1.53)	0.64	0.45 (0.00 – 1.79)	0.74	0.54 (0.04 – 1.66)	0.46	0.42 (0.04 – 1.20)	0.44
$\beta_{2,4}$	0	0.00 (0.00 – 1.00)	0.13	0.00 (0.00 – 1.54)	0.35	0.31 (0.04 – 1.06)	0.15	0.30 (0.04 – 0.95)	0.24
$\beta_{2,5}$	0	0.00 (0.00 – 0.05)	0.04	0.00 (0.00 – 0.77)	0.16	0.06 (0.00 – 0.47)	0.02	0.09 (0.01 – 0.43)	0.01
$\beta_{3,4}$	1	0.77 (0.00 – 2.14)	0.80	0.73 (0.00 – 2.33)	0.80	0.65 (0.10 – 1.86)	0.54	0.71 (0.20 – 1.60)	0.78
$\beta_{3,5}$	0	0.00 (0.00 – 0.90)	0.16	0.00 (0.00 – 1.34)	0.31	0.13 (0.01 – 0.83)	0.06	0.25 (0.02 – 0.77)	0.17
$\beta_{4,5}$	0	0.00 (0.00 – 0.73)	0.16	0.00 (0.00 – 1.27)	0.38	0.14 (0.01 – 0.84)	0.09	0.30 (0.03 – 0.83)	0.27

APPENDIX D: HCHS/SOL RESULTS TABLES FOR CHAPTER 5

Table 5.25: Convex combination centroid summary for the modularized tensor mixture model by latent class identified displayed as time (HH:MM:SS) in activity intensity out of 16 hours.

	Sample	**Adiposity**		**Time in Activity Intensity (out of 16 hrs)**			
Cluster	**n (%)**	**% (95% CI)**	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
1	318	0.76	WEEKDAY	14:03:50	01:53:49	00:02:19	00:00:00
	(11.6)	(0.67 – 0.84)	WEEKEND	14:06:39	01:50:39	00:02:36	00:00:04
2	1486	0.65	WEEKDAY	12:19:27	03:30:04	00:10:14	00:00:12
	(46.7)	(0.60 – 0.69)	WEEKEND	12:35:05	03:19:13	00:05:39	00:00:02
3	1150	0.42	WEEKDAY	11:26:47	04:03:38	00:26:55	00:02:38
	(41.7)	(0.38 – 0.45)	WEEKEND	11:37:28	03:58:16	00:22:02	00:02:12

Table 5.26: Convex combination centroid summary for the supervised modularized tensor mixture model by latent class identified displayed as time (HH:MM:SS) in activity intensity out of 16 hours.

	Sample	**Adiposity**		**Time in Activity Intensity (out of 16 hrs)**			
Cluster	**n (%)**	**% (95% CI)**	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
1	319	0.83	WEEKDAY	13:46:42	02:11:10	00:02:06	00:00:00
	(10.3)	(0.76 – 0.88)	WEEKEND	13:56:28	02:02:36	00:00:51	00:00:03
2	1505	0.69	WEEKDAY	12:25:55	03:23:38	00:10:13	00:00:12
	(48.2)	(0.64 – 0.73)	WEEKEND	12:39:02	03:14:44	00:06:09	00:00:03
3	1127	0.36	WEEKDAY	11:29:48	04:00:54	00:26:37	00:02:39
	(41.5)	(0.32 – 0.39)	WEEKEND	11:40:52	03:55:11	00:21:43	00:02:12

Table 5.27: Convex combination centroid summary for the proposed Bayesian tensor mixture of product kernels (BTMPK) by latent class identified displayed as time (HH:MM:SS) in activity intensity out of 16 hours.

	Sample	Adiposity		Time in Activity Intensity (out of 16 hrs)			
Cluster	**n (%)**	**% (95% CI)**	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
1	1	0.00	WEEKDAY	14:37:53	01:18:25	00:02:13	00:01:28
	(0.1)	(0.00 – 0.00)	WEEKEND	14:52:56	01:06:30	00:00:00	00:00:32
2	1	1.00	WEEKDAY	14:20:25	01:39:17	00:00:00	00:00:16
	(0.0)	(1.00 – 1.00)	WEEKEND	14:20:21	01:39:07	00:00:30	00:00:00
3	308	0.78	WEEKDAY	13:30:40	02:26:45	00:02:34	00:00:00
	(9.8)	(0.69 – 0.84)	WEEKEND	13:52:25	02:07:34	00:00:00	00:00:00
4	977	0.70	WEEKDAY	12:51:30	03:01:26	00:07:03	00:00:00
	(30.2)	(0.66 – 0.75)	WEEKEND	12:57:09	02:57:58	00:04:52	00:00:00
5	47	0.54	WEEKDAY	12:29:38	03:11:52	00:17:16	00:01:12
	(1.5)	(0.32 – 0.75)	WEEKEND	13:35:57	02:24:02	00:00:00	00:00:00
6	141	0.59	WEEKDAY	12:24:23	03:24:25	00:11:10	00:00:00
	(6.0)	(0.43 – 0.74)	WEEKEND	12:05:04	03:35:32	00:18:08	00:01:14
7	616	0.46	WEEKDAY	12:07:10	03:32:14	00:19:21	00:01:13
	(21.5)	(0.41 – 0.52)	WEEKEND	12:29:39	03:21:20	00:09:00	00:00:00
8	650	0.38	WEEKDAY	11:28:54	03:55:22	00:31:53	00:03:49
	(24.2)	(0.33 – 0.42)	WEEKEND	11:29:00	03:59:07	00:28:14	00:03:37
9	130	0.65	WEEKDAY	09:14:07	06:34:42	00:11:09	00:00:00
	(4.4)	(0.52 – 0.76)	WEEKEND	10:01:55	05:49:19	00:08:44	00:00:00
10	83	0.55	WEEKDAY	08:41:41	06:49:01	00:27:34	00:01:41
	(2.3)	(0.39 – 0.69)	WEEKEND	10:23:53	05:24:15	00:11:51	00:00:00

Table 5.28: Convex combination centroid summary for the proposed supervised Bayesian tensor mixture of product kernels (BTMPK) by latent class identified displayed as time (HH:MM:SS) in activity intensity out of 16 hours.

	Sample	Adiposity		Time in Activity Intensity (out of 16 hrs)			
Cluster	**n (%)**	**% (95% CI)**	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
1	309	0.77	WEEKDAY	13:30:19	02:27:08	00:02:31	00:00:00
	(9.8)	(0.68 – 0.84)	WEEKEND	13:52:14	02:07:45	00:00:00	00:00:00
2	40	0.93	WEEKDAY	13:30:04	02:29:55	00:00:00	00:00:00
	(1.3)	(0.82 – 0.98)	WEEKEND	13:16:29	02:39:35	00:03:32	00:00:22
3	47	0.54	WEEKDAY	12:29:38	03:11:52	00:17:16	00:01:12
	(1.5)	(0.32 – 0.75)	WEEKEND	13:35:57	02:24:02	00:00:00	00:00:00
4	1073	0.69	WEEKDAY	12:27:15	03:24:53	00:07:51	00:00:00
	(33.5)	(0.64 – 0.73)	WEEKEND	12:37:22	03:17:08	00:05:28	00:00:00
5	135	0.59	WEEKDAY	12:23:32	03:24:54	00:11:32	00:00:00
	(5.9)	(0.42 – 0.74)	WEEKEND	12:05:39	03:34:56	00:18:10	00:01:13
6	698	0.47	WEEKDAY	11:50:08	03:48:20	00:20:14	00:01:16
	(23.8)	(0.42 – 0.52)	WEEKEND	12:18:59	03:31:42	00:09:17	00:00:00
7	649	0.38	WEEKDAY	11:29:16	03:55:03	00:31:51	00:03:48
	(24.1)	(0.33 – 0.42)	WEEKEND	11:29:32	03:58:34	00:28:14	00:03:38

Table 5.29: Convex combination centroid summary for the Bayesian mixture of multivariate Gaussians (BMMG) by latent class identified displayed as time (HH:MM:SS) in activity intensity out of 16 hours.

	Sample	**Adiposity**		**Time in Activity Intensity (out of 16 hrs)**			
Cluster	**n (%)**	**% (95% CI)**	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
1	13	0.90	WEEKDAY	15:38:59	00:20:48	00:00:11	00:00:00
	(0.4)	(0.67 – 0.98)	WEEKEND	15:37:22	00:22:37	00:00:00	00:00:00
2	126	0.82	WEEKDAY	13:53:43	02:05:57	00:00:17	00:00:01
	(4.6)	(0.67 – 0.91)	WEEKEND	14:04:32	01:53:02	00:01:53	00:00:31
3	154	0.68	WEEKDAY	12:48:08	03:00:48	00:10:44	00:00:18
	(5.0)	(0.56 – 0.78)	WEEKEND	12:58:45	03:00:22	00:00:48	00:00:03
4	1867	0.60	WEEKDAY	12:15:43	03:30:46	00:12:58	00:00:32
	(59.7)	(0.56 – 0.63)	WEEKEND	12:29:35	03:22:55	00:07:25	00:00:03
5	29	0.51	WEEKDAY	12:11:35	03:39:27	00:05:46	00:03:09
	(1.1)	(0.28 – 0.73)	WEEKEND	12:41:13	03:13:44	00:04:57	00:00:04
6	714	0.42	WEEKDAY	11:41:44	03:46:25	00:28:42	00:03:08
	(27.3)	(0.36 – 0.48)	WEEKEND	11:38:14	03:52:14	00:26:24	00:03:06
7	3	1.00	WEEKDAY	11:23:30	03:42:54	00:53:13	00:00:21
	(0.2)	(1.00 – 1.00)	WEEKEND	12:13:40	03:21:07	00:25:11	00:00:00
8	37	0.70	WEEKDAY	10:01:32	05:35:44	00:21:18	00:01:24
	(1.2)	(0.50 – 0.84)	WEEKEND	14:03:22	01:52:12	00:04:11	00:00:13
9	11	0.71	WEEKDAY	05:47:10	09:53:34	00:18:43	00:00:31
	(0.5)	(0.30 – 0.93)	WEEKEND	05:59:09	09:52:49	00:07:56	00:00:05

Table 5.30: Convex combination centroid summary for the Bayesian mixture of product kernels (BMPK) by latent class identified displayed as time (HH:MM:SS) in activity intensity out of 16 hours

	Sample	Adiposity		Time in Activity Intensity (out of 16 hrs)			
Cluster	**n (%)**	**% (95% CI)**	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
1	13	0.89	WEEKDAY	14:49:04	01:08:48	00:02:06	00:00:00
	(0.4)	(0.67 – 0.97)	WEEKEND	15:11:23	00:48:01	00:00:35	00:00:00
2	96	0.85	WEEKDAY	13:58:30	02:01:18	00:00:11	00:00:00
	(3.1)	(0.73 – 0.92)	WEEKEND	14:06:56	01:52:59	00:00:04	00:00:00
3	22	0.46	WEEKDAY	13:41:33	02:17:40	00:00:46	00:00:00
	(0.9)	(0.22 – 0.73)	WEEKEND	13:00:55	02:50:29	00:06:39	00:01:56
4	44	0.90	WEEKDAY	13:32:52	02:27:03	00:00:04	00:00:00
	(1.3)	(0.76 – 0.96)	WEEKEND	13:30:41	02:26:51	00:02:26	00:00:00
5	228	0.74	WEEKDAY	13:04:22	02:51:01	00:04:35	00:00:00
	(7.2)	(0.63 – 0.82)	WEEKEND	13:26:56	02:33:00	00:00:02	00:00:00
6	79	0.79	WEEKDAY	13:00:28	02:58:34	00:00:57	00:00:00
	(3.5)	(0.63 – 0.90)	WEEKEND	13:25:14	02:32:40	00:02:05	00:00:00
7	427	0.68	WEEKDAY	12:33:54	03:17:56	00:08:08	00:00:00
	(13.1)	(0.60 – 0.75)	WEEKEND	12:30:31	03:22:23	00:07:05	00:00:00
8	250	0.68	WEEKDAY	12:20:08	03:32:33	00:07:18	00:00:00
	(7.6)	(0.58 – 0.76)	WEEKEND	12:46:34	03:11:50	00:01:34	00:00:00
9	79	0.56	WEEKDAY	12:20:00	03:23:16	00:15:35	00:01:07
	(2.5)	(0.38 – 0.73)	WEEKEND	13:14:58	02:44:42	00:00:18	00:00:00
10	121	0.62	WEEKDAY	12:10:47	03:33:16	00:15:56	00:00:00
	(5.2)	(0.45 – 0.77)	WEEKEND	11:57:04	03:40:50	00:20:56	00:01:08
11	174	0.51	WEEKDAY	12:08:26	03:39:34	00:11:17	00:00:42
	(6.2)	(0.42 – 0.61)	WEEKEND	12:28:42	03:27:49	00:03:27	00:00:00
12	262	0.66	WEEKDAY	11:58:48	03:38:18	00:22:53	00:00:00
	(7.8)	(0.59 – 0.72)	WEEKEND	12:03:20	03:35:26	00:21:12	00:00:00
13	247	0.47	WEEKDAY	11:56:20	03:44:26	00:18:28	00:00:44
	(8.4)	(0.39 – 0.54)	WEEKEND	11:54:57	03:44:20	00:19:07	00:01:34
14	284	0.52	WEEKDAY	11:48:39	03:49:38	00:21:09	00:00:32
	(8.9)	(0.45 – 0.60)	WEEKEND	12:08:02	03:36:13	00:15:43	00:00:00
15	207	0.36	WEEKDAY	11:33:06	03:49:29	00:31:55	00:05:28
	(7.7)	(0.27 – 0.46)	WEEKEND	12:18:54	03:27:10	00:13:55	00:00:00
16	316	0.32	WEEKDAY	11:22:52	03:53:00	00:37:37	00:06:29
	(12.1)	(0.26 – 0.39)	WEEKEND	11:28:24	03:57:51	00:30:09	00:03:34
17	88	0.34	WEEKDAY	10:21:01	04:12:56	00:58:24	00:27:37
	(3.7)	(0.20 – 0.50)	WEEKEND	10:16:00	04:27:30	00:52:02	00:24:25
18	17	0.75	WEEKDAY	09:51:45	05:15:47	00:52:02	00:00:25
	(0.5)	(0.46 – 0.92)	WEEKEND	10:15:08	05:04:00	00:40:50	00:00:00

Table 5.31: Convex combination centroid summary for the supervised Bayesian mixture of product kernels (BMPK) by latent class identified displayed as time (HH:MM:SS) in activity intensity out of 16 hours

	Sample	Adiposity		Time in Activity Intensity (out of 16 hrs)			
Cluster	**n (%)**	**% (95% CI)**	**Time of Week**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
1	19	0.99	WEEKDAY	15:00:05	00:59:13	00:00:41	00:00:00
	(1.2)	(0.88 – 1.00)	WEEKEND	15:28:16	00:31:19	00:00:24	00:00:00
2	87	0.84	WEEKDAY	13:58:45	02:01:05	00:00:08	00:00:00
	(2.6)	(0.69 – 0.92)	WEEKEND	14:04:48	01:55:08	00:00:02	00:00:00
3	18	0.54	WEEKDAY	13:40:14	02:19:08	00:00:36	00:00:00
	(0.8)	(0.24 – 0.81)	WEEKEND	13:15:24	02:34:54	00:07:18	00:02:22
4	47	0.91	WEEKDAY	13:23:39	02:36:13	00:00:06	00:00:00
	(1.5)	(0.79 – 0.97)	WEEKEND	13:24:49	02:32:45	00:02:24	00:00:00
5	225	0.74	WEEKDAY	13:02:58	02:52:44	00:04:16	00:00:00
	(7.2)	(0.64 – 0.82)	WEEKEND	13:23:23	02:36:34	00:00:02	00:00:00
6	417	0.70	WEEKDAY	12:26:40	03:25:37	00:07:42	00:00:00
	(12.4)	(0.61 – 0.77)	WEEKEND	12:26:53	03:26:11	00:06:55	00:00:00
7	180	0.52	WEEKDAY	12:21:07	03:26:51	00:11:14	00:00:46
	(6.2)	(0.42 – 0.63)	WEEKEND	12:42:10	03:14:43	00:03:06	00:00:00
8	67	0.56	WEEKDAY	12:20:34	03:23:15	00:15:03	00:01:06
	(2.2)	(0.36 – 0.74)	WEEKEND	13:13:31	02:46:14	00:00:14	00:00:00
9	256	0.65	WEEKDAY	12:19:47	03:32:53	00:07:19	00:00:00
	(7.4)	(0.55 – 0.74)	WEEKEND	12:47:54	03:10:33	00:01:32	00:00:00
10	265	0.64	WEEKDAY	12:14:03	03:22:43	00:23:13	00:00:00
	(8.6)	(0.55 – 0.73)	WEEKEND	12:13:18	03:27:07	00:19:33	00:00:00
11	84	0.77	WEEKDAY	12:13:58	03:44:50	00:01:10	00:00:00
	(2.9)	(0.61 – 0.88)	WEEKEND	12:31:59	03:25:23	00:02:36	00:00:00
12	123	0.61	WEEKDAY	12:12:25	03:32:15	00:15:19	00:00:00
	(5.2)	(0.43 – 0.77)	WEEKEND	11:54:13	03:44:32	00:20:06	00:01:08
13	232	0.48	WEEKDAY	11:52:53	03:48:07	00:18:17	00:00:41
	(7.9)	(0.40 – 0.57)	WEEKEND	11:54:22	03:45:21	00:18:44	00:01:31
14	284	0.51	WEEKDAY	11:43:29	03:54:54	00:21:03	00:00:32
	(9.2)	(0.42 – 0.60)	WEEKEND	12:02:56	03:41:41	00:15:22	00:00:00
15	210	0.37	WEEKDAY	11:29:11	03:53:29	00:31:48	00:05:31
	(7.5)	(0.28 – 0.47)	WEEKEND	12:16:37	03:29:16	00:14:06	00:00:00
16	335	0.32	WEEKDAY	11:28:00	03:48:23	00:37:30	00:06:05
	(13.2)	(0.25 – 0.39)	WEEKEND	11:29:36	03:56:03	00:30:40	00:03:39
17	85	0.34	WEEKDAY	10:07:07	04:20:28	00:59:13	00:33:11
	(3.2)	(0.22 – 0.48)	WEEKEND	10:12:59	04:27:19	00:51:56	00:27:44
18	17	0.53	WEEKDAY	09:22:56	05:48:56	00:47:42	00:00:24
	(0.6)	(0.23 – 0.80)	WEEKEND	09:34:45	05:40:02	00:45:11	00:00:00

APPENDIX E: HCHS/SOL RESULTS FIGURES FOR CHAPTER 5

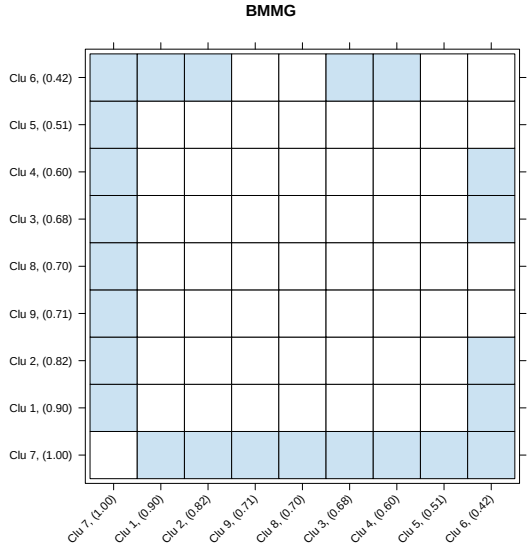

Figure 5.31: Pairwise comparisons of adiposity by latent class membership for the Bayesian mixture of multivariate Gaussians model. Color filled cells denote Tukey adjusted p-values < 0.05 for each latent class pairwise comparison of adiposity. Cluster labeling within each method is ordered by estimated weekday sedentary behavior time budget proportion (sedentary component of estimated convex combination latent class centroid). The estimated percentage of adiposity for each latent class is enclosed in parenthesis.

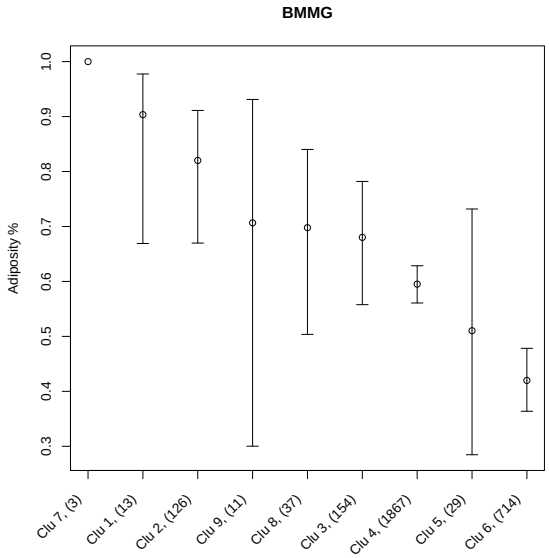

Figure 5.32: Confidence interval plots of estimated adiposity proportion by latent class for the Bayesian mixture of multivariate Gaussians model.

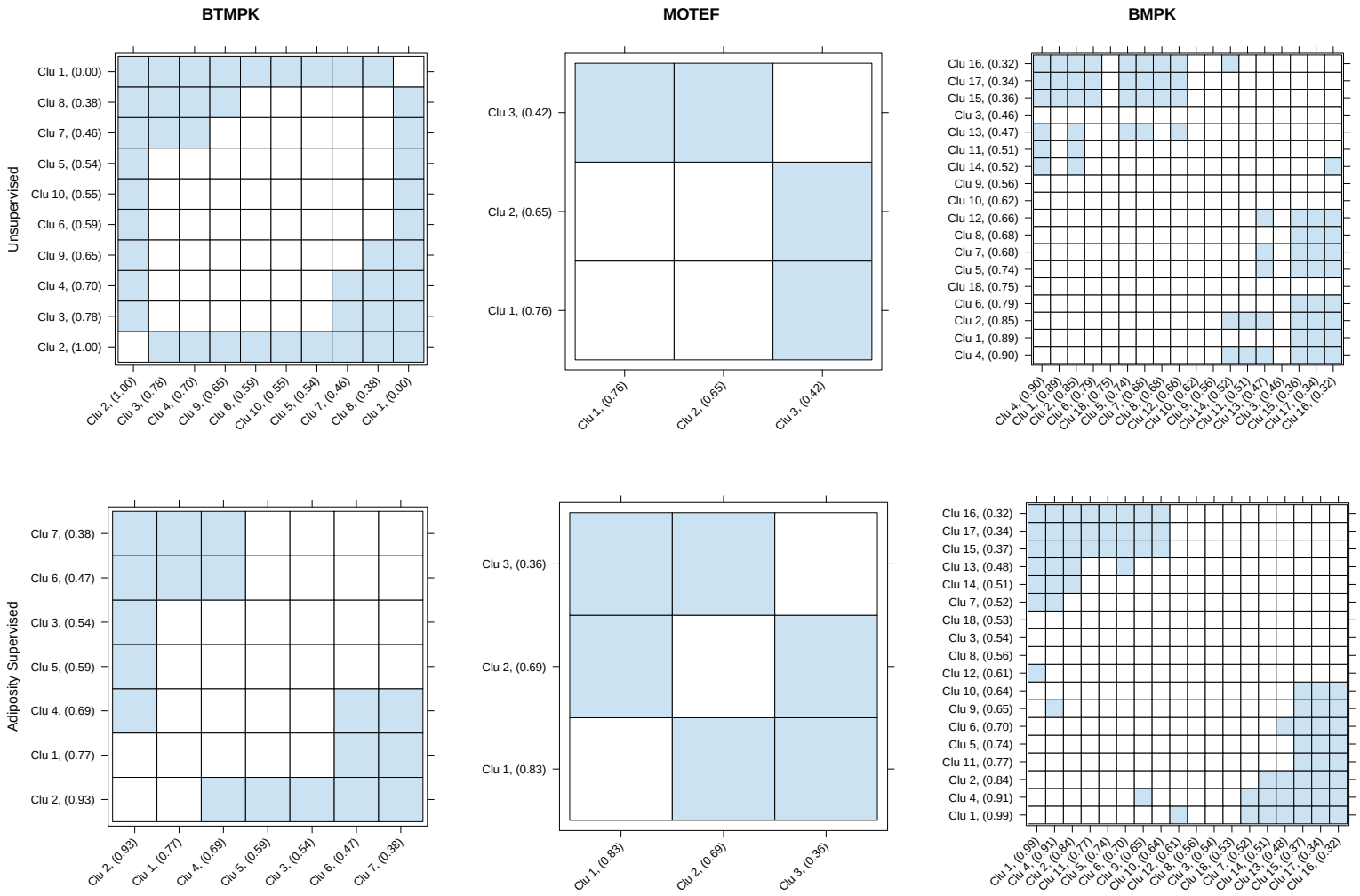

Figure 5.33: Panel displaying pairwise comparisons of adiposity by latent class membership and method implemented. Color filled cells denote Tukey adjusted p-values < 0.05 for each latent class pairwise comparison of adiposity. Cluster labeling within each method is ordered by estimated weekday sedentary behavior time budget proportion (sedentary component of estimated convex combination latent class centroid). The estimated percentage of adiposity for each latent class is enclosed in parenthesis.

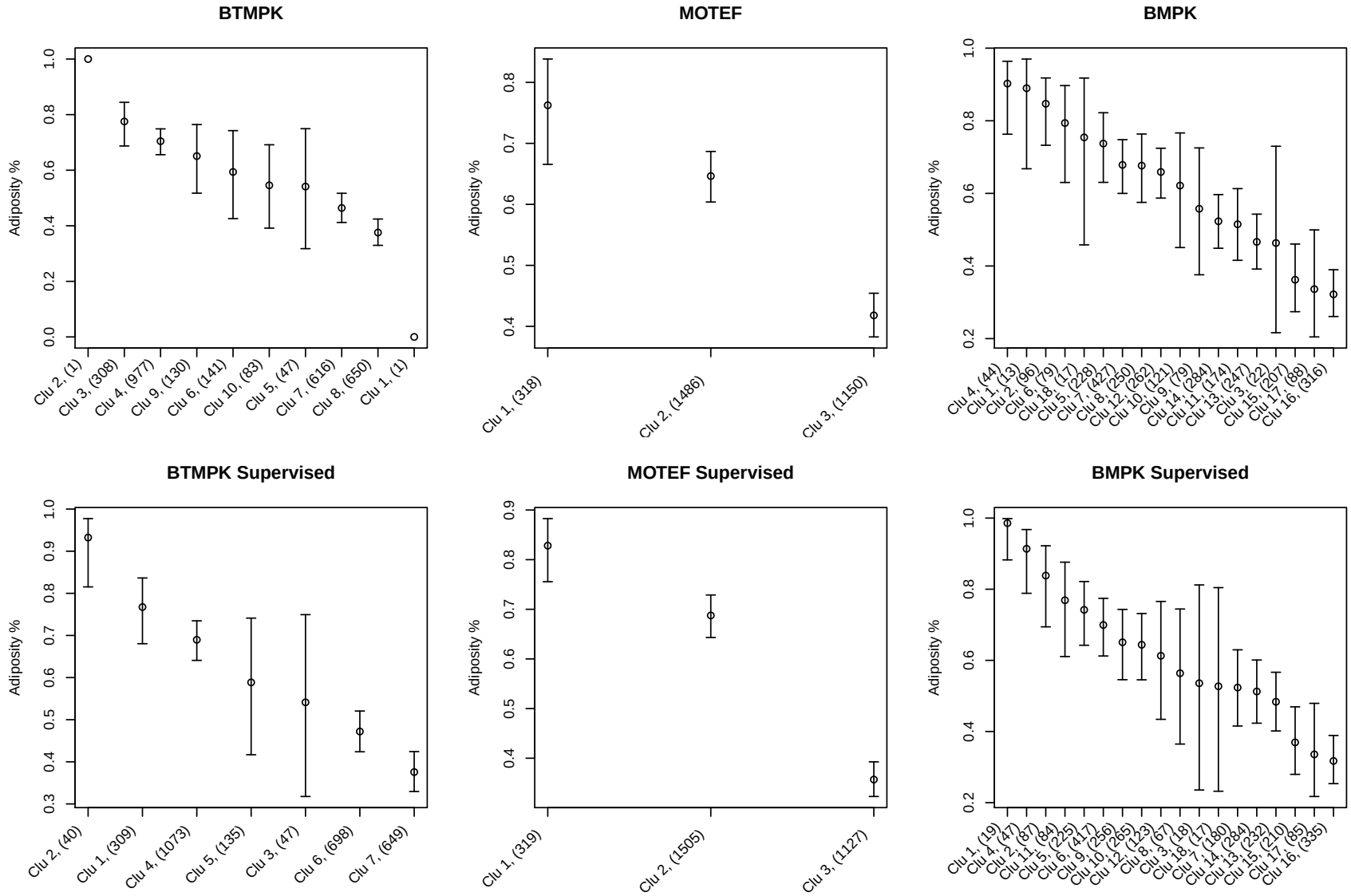

Figure 5.34: Confidence interval plots of estimated adiposity proportion by latent class and method.

BIBLIOGRAPHY

Adams, M. M., Mulinare, J., and Dooley, K. (1989). Risk factors for conotruncal cardiac defects in atlanta. *Journal of the American College of Cardiology*, 14(2):432–442.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, Ltd., London, UK, UK.

Aitchison, J. (2005). A concise guide to compositional data analysis.

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

Amiri, L., Khazaei, M., and Ganjali, M. (2017). A mixture latent variable model for modeling mixed data in heterogeneous populations and its applications. *AStA Advances in Statistical Analysis*, pages 1–21.

Antonelli, J., Mazumdar, M., Bellinger, D., Christiani, D. C., Wright, R., and Coull, B. A. (2017). Estimating the health effects of environmental mixtures using bayesian semiparametric regression and sparsity inducing priors. *arXiv preprint arXiv:1711.11239*.

Banerjee, A., Murray, J., and Dunson, D. (2013). Bayesian learning of joint distributions of objects. In *Artificial Intelligence and Statistics*, pages 1–9.

Barrera-Gómez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J. R., Nieuwenhuijsen, M., et al. (2017). A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environmental Health*, 16(1):74.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2015). Bayesian density estimation for compositional data using random bernstein polynomials. *Journal of Statistical Planning and Inference*, 166:116–125.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111.

Botto, L. D., Panichello, J. D., Browne, M. L., Krikov, S., Feldkamp, M. L., Lammer, E., Shaw, G. M., and Study, N. B. D. P. (2014). Congenital heart defects after maternal fever. *American journal of obstetrics and gynecology*, 210(4):359–e1.

Bouguila, N., Ziou, D., and Vaillancourt, J. (2004). Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Buckley, J. P., Engel, S. M., Mendez, M. A., Richardson, D. B., Daniels, J. L., Calafat, A. M., Wolff, M. S., and Herring, A. H. (2015). Prenatal phthalate exposures and childhood fat mass in a new york city cohort. *Environmental health perspectives*, 124(4):507–513.

Calif, R., Emilion, R., and Soubdhan, T. (2011a). Classification of wind speed distributions using a mixture of dirichlet distributions. *Renewable energy*, 36(11):3091–3097.

Calif, R., Emilion, R., and Soubdhan, T. (2011b). Classification of wind speed distributions using a mixture of dirichlet distributions. *Renewable energy*, 36(11):3091–3097.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539.

Canale, A. and Dunson, D. B. (2015). Bayesian multivariate mixed-scale density estimation. *Statistics and its Interface*, 8(2):195–201.

Carmona, C., Nieto-Barajas, L., and Canale, A. (2016). Model based approach for household clustering with mixed scale variables. *arXiv preprint arXiv:1612.00083*.

Carson, V., Faulkner, G., Sabiston, C. M., Tremblay, M. S., and Leatherdale, S. T. (2015). Patterns of movement behaviors and their association with overweight and obesity in youth. *International journal of public health*, 60(5):551–559.

Chastin, S. F., Palarea-Albaladejo, J., Dontje, M. L., and Skelton, D. A. (2015). Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. *PloS one*, 10(10):e0139984.

Choi, L., Liu, Z., Matthews, C. E., and Buchowski, M. S. (2011). Validation of accelerometer wear and nonwear time classification algorithm. *Medicine and science in sports and exercise*, 43(2):357.

Colley, R. C., Garriguet, D., Janssen, I., Craig, C. L., Clarke, J., and Tremblay, M. S. (2011). Physical activity of canadian adults: accelerometer results from the 2007 to 2009 canadian health measures survey. *Health reports*, 22(1):7.

Comas-Cufí, M., Martín-Fernández, J. A., and Mateu-Figueras, G. (2016). Log-ratio methods in mixture models for compositional data sets. *SORT-Statistics and Operations Research Transactions*, 1(2):349–374.

Cox, D. R. (1984). Interaction. *International Statistical Review/Revue Internationale de Statistique*, pages 1–24.

Cox, D. R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):441–461.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218.

Dahl-Petersen, I. K., Brage, S., Bjerregaard, P., Tolstrup, J., and Jørgensen, M. E. (2017). Physical activity and abdominal fat distribution in greenland. *Medicine and science in sports and exercise*, 49(10):2064.

Davalos, A., Herring, A., and Olshan, A. (2019). Joint modeling of mixed scale variables using modularized tensor factorization. *Submitted*.

Davalos, A. D., Luben, T. J., Herring, A. H., and Sacks, J. D. (2017). Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Annals of Epidemiology*.

de Leon, A. (2007). One-sample likelihood ratio tests for mixed data. *Communications in Statistics—Theory and Methods*, 36(1):129–141.

de Leon, A. and Carriégre, K. (2005). A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis*, 92(1):174–185.

de Leon, A. and Carrièggre, K. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics*, 35(4):533–548.

De Leon, A., Soo, A., and Williamson, T. (2011). Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, 38(5):1021–1032.

De Leon, A. R. and Chough, K. C. (2013). *Analysis of mixed data: Methods & applications*. Chapman and Hall/CRC.

Dumuid, D., Olds, T., Lewis, L. K., Martin-Fernández, J. A., Katzmarzyk, P. T., Barreira, T., Broyles, S. T., Chaput, J.-P., Fogelholm, M., Hu, G., et al. (2017a). Health-related quality of life and lifestyle behavior clusters in school-aged children from 12 countries. *The Journal of pediatrics*, 183:178–183.

Dumuid, D., Stanford, T. E., Martin-Fernández, J.-A., Pedišić, Ž., Maher, C. A., Lewis, L. K., Hron, K., Katzmarzyk, P. T., Chaput, J.-P., Fogelholm, M., et al. (2017b). Compositional data analysis for physical activity, sedentary time and sleep research. *Statistical methods in medical research*, page 0962280217710835.

Dunson, D. B. and Bhattacharya, A. (2011). Nonparametric Bayes regression and classification through mixtures of product kernels. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9*, pages 145–164. Oxford University Press, Oxford.

Dunson, D. B. and Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25.

Dunson, D. B., Herring, A. H., and Engel, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, 103(482):534–546.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Edwards, D. (2012). *Introduction to graphical modelling*. Springer Science & Business Media.

Efromovich, S. (2011). Nonparametric estimation of the anisotropic probability density of mixed variables. *Journal of Multivariate Analysis*, 102(3):468–481.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Evenson, K. R., Wen, F., Hales, D., and Herring, A. H. (2016). National youth sedentary behavior and physical activity daily patterns using latent class analysis applied to accelerometry. *International Journal of Behavioral Nutrition and Physical Activity*, 13(1):55.

Evenson, K. R., Wen, F., Metzger, J. S., and Herring, A. H. (2015). Physical activity and sedentary behavior patterns using accelerometry from a national sample of united states adults. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):20.

Feng, Y., Yu, D., Yang, L., Da, M., Wang, Z., Lin, Y., Ni, B., Wang, S., and Mo, X. (2014). Maternal lifestyle factors in pregnancy and congenital heart defects in offspring: review of the current evidence. *Italian journal of pediatrics*, 40(1):85.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.

Fernández, M., Daunis i Estadella, J., and Mateu i Figueras, G. (2015). On the interpretation of differences between groups for compositional data. *SORT: statistics and operations research transactions, 2015, vol. 39, núm. 2, p. 231-252*.

Filippi, S., Holmes, C. C., and Nieto-Barajas, L. E. (2016). Scalable Bayesian nonparametric measures for exploring pairwise dependence via Dirichlet process mixtures. *Electronic Journal of Statistics*, 10(2):3338–3354.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

Frolov, P., Alali, J., and Klein, M. D. (2010). Clinical risk factors for gastroschisis and omphalocele in humans: a review of the literature. *Pediatric surgery international*, 26(12):1135–1148.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 3. Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5*.

Gilboa, S. M., Correa, A., Botto, L. D., Rasmussen, S. A., Waller, D. K., Hobbs, C. A., Cleves, M. A., Riehle-Colarusso, T. J., and Study, N. B. D. P. (2010). Association between prepregnancy body mass index and congenital heart defects. *American journal of obstetrics and gynecology*, 202(1):51–e1.

Giordan, M. and Wehrens, R. (2015). A comparison of computational approaches for maximum likelihood estimation of the dirichlet parameters on high-dimensional data. *SORT*, 39:109–126.

Green, R. F., Devine, O., Crider, K. S., Olney, R. S., Archer, N., Olshan, A. F., Shapira, S. K., and Study, T. N. B. D. P. (2010). Association of paternal age and risk for major congenital anomalies from the national birth defects prevention study, 1997 to 2004. *Annals of epidemiology*, 20(3):241–249.

Gupta, N., Mathiassen, S. E., Mateu-Figueras, G., Heiden, M., Hallman, D. M., Jørgensen, M. B., and Holtermann, A. (2018). A comparison of standard and compositional data analysis in studies addressing group differences in sedentary behavior and physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 15(1):53.

Hamra, G. B. and Buckley, J. P. (2018). Environmental exposure mixtures: questions and methods to address them. *Current epidemiology reports*, 5(2):160–165.

Haris, A., Witten, D., and Simon, N. (2016). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004.

Harley, K. G., Berger, K., Rauch, S., Kogut, K., Henn, B. C., Calafat, A. M., Huen, K., Eskenazi, B., and Holland, N. (2017). Association of prenatal urinary phthalate metabolite concentrations and childhood bmi and obesity. *Pediatric research*, 82(3):405.

Hastie, D. I., Liverani, S., and Richardson, S. (2015). Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, 25(5):1023–1037.

Herring, A. H. (2010). Nonparametric Bayes shrinkage for assessing exposures to mixtures subject to limits of detection. *Epidemiology*, 21(Suppl 4):S71.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.*, 1(1):265–283.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., and Hothorn, M. T. (2017). Package 'multcomp'. *See https://cran. r-project. org/web/packages/multcomp/index. html*.

Huh, J., Riggs, N. R., Spruijt-Metz, D., Chou, C.-P., Huang, Z., and Pentz, M. (2011). Identifying patterns of eating and physical activity in children: a latent class analysis of obesity risk. *Obesity*, 19(3):652–658.

Institute, S. (2015). *Base SAS 9.4 procedures guide*. SAS Institute.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105.

Kaul, A., Davidov, O., and Peddada, S. D. (2017a). Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics*, 18(3):422–433.

Kaul, A., Mandal, S., Davidov, O., and Peddada, S. D. (2017b). Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:2114.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36:493–496.

Kunihama, T., Halpern, C. T., and Herring, A. H. (2016). Nonparametric Bayes models for mixed-scale longitudinal surveys. *arXiv preprint arXiv:1606.02381*.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1):31–57.

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., et al. (2010). Sample design and cohort selection in the hispanic community health study/study of latinos. *Annals of epidemiology*, 20(8):642–649.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, pages 1222–1235.

Leech, R. M., McNaughton, S. A., and Timperio, A. (2014). The clustering of diet, physical activity and sedentary behavior in children and adolescents: a review. *International Journal of Behavioral Nutrition and Physical Activity*, 11(1):4.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian analysis*, 5(1):151–170.

Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2):266–292.

Lim, M. and Hastie, T. (2013). Glinternet: learning interactions via hierarchical group-lasso regularization. *R package version 0.9. 0*.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Little, R. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72(3):497–512.

Liu, J., Kim, J., Colabianchi, N., Ortaglia, A., and Pate, R. R. (2010). Co-varying patterns of physical activity and sedentary behaviors and their long-term maintenance among adolescents. *Journal of Physical Activity and Health*, 7(4):465–474.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics*, 42(2):413.

Lumley, T. et al. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

Lumley, T. and Lumley, M. T. (2019). Package 'survey'.

Mac Bird, T., Robbins, J. M., Druschel, C., Cleves, M. A., Yang, S., Hobbs, C. A., and Study, N. B. D. P. (2009). Demographic and environmental risk factors for gastroschisis and omphalocele in the national birth defects prevention study. *Journal of pediatric surgery*, 44(8):1546–1551.

MacLehose, R. F. and Dunson, D. B. (2010). Bayesian semiparametric multiple shrinkage. *Biometrics*, 66(2):455–462.

MacLehose, R. F., Dunson, D. B., Herring, A. H., and Hoppin, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology*, 18(2):199–207.

Malik, S., Cleves, M. A., Honein, M. A., Romitti, P. A., Botto, L. D., Yang, S., Hobbs, C. A., et al. (2008). Maternal smoking and congenital heart defects. *Pediatrics*, 121(4):e810–e816.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing*, 26(1-2):303–324.

Malsiner-Walli, G., Pauger, D., and Wagner, H. (2018). Effect fusion using model-based clustering. *Statistical Modelling*, 18(2):175–196.

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9):2688–2704.

Martın-Fernandez, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011). Dealing with zeros. *Compositional data analysis: Theory and applications*, pages 43–58.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2011). *The principle of working on coordinates*. John Wiley & Sons, Chichester.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J.-J. (2013). The normal distribution in some constrained sample spaces. *SORT: statistics and operations research transactions*, 37(1):0029–56.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models*. Routledge.

McGrath, S., Brazel, D., Dugas, L., Cao, G., Durazo-Arvizu, R., and Luke, A. (2017). Physical activity and central adiposity in a cohort of african-american adults. *BMC obesity*, 4(1):34.

Mena, R. H. and Walker, S. G. (2015). On the Bayesian mixture model and identifiability. *Journal of Computational and Graphical Statistics*, 24(4):1155–1169.

Mirkamali, S. J. and Ganjali, M. (2016). A Bayesian joint modeling using Gaussian linear latent variables for mixed correlated outcomes with possibility of missing values. *Journal of Statistical Theory and Application*, 15(4):373–386.

Molitor, J., Coker, E., Jerrett, M., Ritz, B., and Li, A. (2016). Part 3. modeling of multipollutant profiles and spatially varying health effects with applications to indicators of adverse birth outcomes. *Research Report (Health Effects Institute)*, 183(3):3–47.

Moon, J.-Y., Wang, T., Sofer, T., North, K. E., Isasi, C. R., Cai, J., Gellman, M. D., Moncrieft, A. E., Sotres-Alvarez, D., Argos, M., et al. (2017). Objectively measured physical activity, sedentary behavior and genetic predisposition to obesity in us hispanics/latinos: results from the hispanic community health study/study of latinos (hchs/sol). *Diabetes*, page db170573.

Murabito, J. M., Pedley, A., Massaro, J. M., Vasan, R. S., Esliger, D., Blease, S. J., Hoffman, U., and Fox, C. S. (2015). Moderate-to-vigorous physical activity with accelerometry is associated with visceral adipose tissue in adults. *Journal of the American Heart Association*, 4(3):e001379.

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(502):656–665.

Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479.

Myers, A., Gibbons, C., Butler, E., Dalton, M., Buckland, N., Blundell, J., and Finlayson, G. (2018). Disentangling the relationship between sedentariness and obesity: Activity intensity, but not sitting posture, is associated with adiposity in women. *Physiology & behavior*, 194:113–119.

Nagler, T. (2017). Asymptotic analysis of the continuous convolution kernel density estimator. *arXiv preprint arXiv:1705.05431*.

Norets, A. and Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, 168(2):332–346.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32(2):448–465.

Papageorgiou, G. and Richardson, S. (2016). Bayesian density regression for discrete outcomes. *arXiv preprint arXiv:1603.09706*.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, pages 169–186.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Patel, S. S. and Burns, T. L. (2013). Nongenetic risk factors and congenital heart defects. *Pediatric cardiology*, 34(7):1535–1555.

Patnode, C. D., Lytle, L. A., Erickson, D. J., Sirard, J. R., Barr-Anderson, D. J., and Story, M. (2011). Physical activity and sedentary activity patterns among children and adolescents: a latent class analysis approach. *Journal of Physical Activity and Health*, 8(4):457–467.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2002). Blu estimators and compositional data. *Mathematical Geology*, 34(3):259–274.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.

Poon, W.-Y. and Lee, S.-Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*, 52(3):409–430.

Rasmussen, S. A. and Frías, J. L. (2008). Non-genetic risk factors for gastroschisis. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 148, pages 199–212. Wiley Online Library.

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

Rivera-Pinto, J., Egozcue, J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. (2018). Balances: a new perspective for microbiome analysis. *MSystems*, 3(4):e00053–18.

Rodríguez, C. E. and Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

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

Sotres-Alvarez, D., Siega-Riz, A. M., Herring, A. H., Carmichael, S. L., Feldkamp, M. L., Hobbs, C. A., Olshan, A. F., and the National Birth Defects Prevention Study (2013). Maternal dietary patterns are associated with risk of neural tube and congenital heart defects. *American Journal of Epidemiology*, 177(11):1279–1288.

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

Stephenson, B., Herring, A. H., and Olshan, A. (2017). Robust clustering with subpopulation-specific deviations. *arXiv preprint arXiv:1711.03884*.

Suh, H. H., Zanobetti, A., Schwartz, J., and Coull, B. A. (2011). Chemical properties of air pollutants and cause-specific hospital admissions among the elderly in atlanta, georgia. *Environmental health perspectives*, 119(10):1421–1428.

Sun, Z., Tao, Y., Li, S., Ferguson, K. K., Meeker, J. D., Park, S. K., Batterman, S. A., and Mukherjee, B. (2013). Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health*, 12(1):85.

Tate, R. F. (1954). Correlation between a discrete and continuous variable. *Annals of Mathematical Statistics*, 25:603–607.

Templ, M., Hron, K., Filzmoser, P., and Templ, M. M. (2018). Package 'robcompositions'. *alr*, 10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54.

Wong, S. L., Colley, R., Gorber, S. C., and Tremblay, M. (2011). Actical accelerometer sedentary activity thresholds for adults. *Journal of Physical Activity and Health*, 8(4):587–591.

Wu, Y. and Ghosal, S. (2010). The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419.

Xu, X., Ghosh, M., et al. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936.

Zhang, X., Boscardin, W., Belin, T., Wan, X., He, Y., and Zhang, K. (2015). A Bayesian method for analyzing combinations of continuous, ordinal, and nominal categorical data with missing values. *Journal of Multivariate Analysis*, 135:43–58.

Zhang, Z., Descoteaux, M., and Dunson, D. B. (2016). Nonparametric Bayes models of fiber curves connecting brain regions. *arXiv preprint arXiv:1612.01014*.