NONPARAMETRIC BAYES METHODS FOR HIGH DIMENSIONAL DATA AND GROUP SEQUENTIAL DESIGN FOR LONGITUDINAL TRIALS

Jing Zhou

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.

Chapel Hill
2014

Approved by:

Amy H. Herring

David B. Dunson

Gary Koch

Keaven Anderson

Andrew Olshan

ABSTRACT

JING ZHOU: Nonparametric Bayes Methods for High Dimensional Data AND Group Sequential Design for Longitudinal Trials (Under the direction of Amy H. Herring)

High-dimensional unordered categorical data appear in a number of areas ranging from epidemiology, behavioral and social sciences, etc. Such data can be placed into a large contingency table with cell counts defined as the number of subjects with a given combination of variables values. The contingency table is often sparse in practice in the sense that only a few cells have more than a few counts, with most cells being empty. Traditional approaches for contingency table analysis fail to scale up to moderate dimensions, and alternative approaches based on tensor decomposition are promising. This motivates us to develop sparse tensor decompositions for multivariate categorical variables where the number of variables can be potentially larger than the sample size. The methods are shown to have excellent performance in simulations, and results in various data sets are presented.

In paper 2, we consider such high-dimensional data in case-control studies, with the main goal being detection of the sparse subset of predictors having a significant association with disease. We propose a new approach based on a nonparametric Bayesian low rank tensor factorization to model the retrospective likelihood. Our model allows a very flexible structure in characterizing the distribution of multivariate variables as unknown and without any linearity assumptions as in logistic regression. Predictors are excluded only if they have no impact on disease risk, either directly or through interactions with other predictors. Hence, we obtain an omnibus approach for screening for important predictors. Computation relies on an efficient Gibbs sampler. The methods are shown to have higher power and lower false discovery rates in simulation studies relative to existing methods, and we consider an application to an epidemiologic study of birth defects.

In paper 3, our goal is to design a longitudinal trial using group sequential design. We propose an information-based sample size re-estimation method to update the sample size at each interim analysis, which maintains the target power while controlling the type-I error rate. We illustrate our strategy by data analysis examples and simulations and compare the results with those obtained using fixed design and group-sequential design without sample size re-estimation.

ACKNOWLEDGMENTS

First and foremost I want to thank Professor Amy Herring for all of her valuable guidance and support through my years at UNC-Chapel Hill. I do not think I could have had a better example of how to perform quality research and carry oneself as a professional.

I am grateful to Prof. David Dunson for his suggestions and feedback on multiple projects over the course of my studies, and to Anirban Bhattacharya for being a great research colleague.

I also want to thank Prof. Gary Koch, Dr. Keaven Anderson, and Prof. Andrew Olshan for serving on my Ph.D. committee and for their helpful input and suggestions.

I appreciate Dr. Keaven Anderson for providing a very interesting research topic to expand my research horizons.

My research would not have been possible without the financial support of the National Birth Defects Prevention Study grant. I am also grateful to the research team in providing great scientific support.

To my family, I am ever thankful for your support through these many years. And to Richard, I cannot express how much your love and support has meant to me, I could not have done this without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Literature Review

1.1 Bayesian Methodologies Introduction(Paper 1 and 2)

The most important aspects of epidemiologic research are uncovering dependencies among multiple closely-related exposures and health outcomes, discovering risk factors, and developing accurate predictive models. This is particularly true given that it is crucial in many settings to account for interactions. Multivariate unordered categorical data are routinely encountered in such areas. For example, the categorical variables may correspond to a sequence of A, C, G, T nucleotides in candidate genes or responses to questionnaire data on race, religion and demographic information for an individual (Bhattacharya and Dunson 2012).

In particular, consider the National Birth Defects Prevention Study (NBDPS), the largest population-based study ever conducted in the United States on the etiology of birth defects. The study was designed to evaluate environmental, behavioral, biomedical, sociodemographic, and genetic factors associated with the occurrence of congenital malformations. If one considers placing individual pregnant women in a multi-way table defined by levels of these potentially important variables, only a few cells in the table will have more than a few women, with most cells having no women. Regular maximum likelihood estimation will result in unstable estimates and inferences due to collinearity among variables and sparsity of the cells. Hence, it is not possible simultaneously to obtain an estimate (e.g., odds ratio) for each cell of the table without borrowing information strongly. To address this, one could typically apply dimensionality reduction techniques such as (1) preselect variables that seem likely to be important and ignore interactions to reduce dimensionality; (2) group exposures into class-specific summaries; or (3) include variables one at a time as predictors in a low-dimensional logistic regression model, with interactions only added for predictors with main effects, and significance thresholds on p-values adjusted for false discovery rate control. However, such approaches can lead to (1) overlooking important risk factors; (2) discarding valuable information on variability in the effect within a class; or (3) producing misleading results by not adjusting for correlated exposures. Furthermore, it lacks a probabilistic characterization of uncertainty, which is important in making inferences and representing uncertainty in predictions.

There have been literatures on graphical models (Dawid and Lauritzen 1993) that can model the complicated dependence structure for categorical data (Whittaker 1990; Madigan et al. 1995). Although graphical models are popular due to their flexibility and interpretability, computation is difficult since the size of the model space grows exponentially with p. In parallel to graphic models, factor models have been advocated by West (2003) and Carvalho et al. (2008) to model high-dimensional data with emphasis on dimension reduction. Factor models provide a framework for regularized covariance matrix estimation for normal data, and have been extended to allow binary and ordered categorical data through an underlying Gaussian latent factor structure (Muthén 1983). Multivariate probit models (Ashford and Sowden 1970; Chib and Greenberg 1998; Ochi and Prentice 1984; Zhang et al. 2008) have been used for analyzing multivariate binary or nominal variables of interest via thresholding of a vector of latent variables. For further review of existing graphic and factor models, see Bhattacharya and Dunson

(2012).

For unordered categorical variables, factor models, however, run into computation problems due to their complex model and estimation nature. But the data could be alternatively presented in the form of a high-dimensional contingency table for which there is a vast literature. Fienberg and Rinaldo (2007) provide an overview of the development of log-linear models, maximum likelihood estimation and asymptotic tests for goodness of fit. While log-linear models provide a framework to model interactions among related categorical variables, the number of model parameters becomes large even in the case of a two-way interaction model for moderate to large number of variables. Asymptotic tests based on log-linear models face multiple difficulties in the case of sparse contingency tables –refer to the discussion in section 3 of Fienberg and Rinaldo (2007). Although such problems can be alleviated using a Bayesian approach, posterior model search using traditional Markov chain Monte Carlo (MCMC) methods tends to slow down quickly as the dimension increases. Moreover, even with highly efficient search algorithms (Jones et al. 2005; Carvalho and Scott 2009; Dobra and Massam 2010), it is only feasible to visit a small subset of the model space even for moderate p and accurate model selection is a difficult task. This motivates the development of a new class of models for high-dimensional unordered categorical data in the form of a contingency table.

Dunson and Xing (2009) employed this idea and developed a nonparametric Bayes approach using Dirichlet process (Ferguson 1973; 1974) mixtures of product multinomials to directly model the joint distribution of multivariate unordered categorical data. The modeling of the joint distribution of the category probabilities in a sparse manner enables efficient posterior computation, thereby allowing their method to efficiently scale up to high dimensions. This approach extends latent structure analysis (Lazarsfeld and Henry 1968; Goodman 1974) to the infinite mixture case and is conceptually related to non-negative tensor decompositions (Shashua and Hazan 2005; Kim and Choi 2007).

Likewise, Yang and Dunson (2013) proposed a different nonparametric Bayes model on the conditional distribution of the category probabilities. By choosing a carefully-structured Tucker factorization, another popular tensor decomposition method, they defined a model that can characterize any conditional probability, while facilitating variable selection and modeling of high-order interactions. They suggested a two-stage algorithm which first identifies a model with a set of important predictors in the first stage and then learns the posterior distribution for this model via the Gibbs sampler for other unknown parameters.

Although both the Dunson and Xing (2009) and Yang and Dunson (2013) can handle fairly large contingency tables and reduces the number of parameters from exponential in p to linear in p, the estimation problem is still challenging when p is proportional to or larger than n in that it is not possible to estimate the joint/conditional distribution and the corresponding association of interest without further sparsity assumptions. My dissertation (Paper 1 and 2) will be focusing on building up new methodologies that can model high-dimensional multivariate unordered data in the case of $p \propto$ or $\geq n$ while maintaining attractive statistical properties. The new approaches will be mainly based on, and extended from, the models of Dunson and Xing (2009) and Yang and Dunson (2013). Before discussing our approaches, we first introduce the basics of Dirichlet process, tensor decomposition methods and Bayes nonparametrics followed by the detailed specifications of Dunson and Xing (2009) and Yang and Dunson (2013).

1.1.1 Dirichlet Distribution and Process

Dirichlet Distribution

The basic building tool for Bayesian non-parametric methods is called the Dirichlet Process (DP). To discuss the Dirichlet process, we first need to discuss the Dirichlet distribution. The Dirichlet distribution is the multivariate generalization of the beta distribution. Let $z_1, \ldots, z_k$ be independent random variables with $z_j \sim \text{Gamma}(a_j, 1)$, $j = 1, \ldots, k$. Define

$$u = \sum_{j=1}^{k} z_j,$$

and

$$y_j = \frac{z_j}{u} = \frac{z_j}{\sum_{j=1}^{k} z_j}.$$

Since $\sum_{j=1}^{k} y_j = 1$, we say $(y_1, \ldots, y_k)$ have a $k-1$ dimensional Dirichlet distribution $D_{k-1}(a_1, \ldots, a_k)$ with density

$$f(y_1, \ldots, y_k) = \left(\frac{\Gamma(\sum_{j=1}^{k} a_j)}{\prod_{j=1}^{k} \Gamma(a_j)} \right) \left(\prod_{j=1}^{k} y_j^{a_j - 1} \right).$$

The Dirichlet distribution is a distribution over possible vectors for a multinomial distribution. It is in fact a 'distribution over distributions' and hence can be used as a conjugate prior for the multinomial family. That is, if

$$(x_1, \ldots, x_k) \sim \text{Multinomial}(p_1, \ldots, p_k), \text{ where} \sum_{j=1}^{k} p_j = 1,$$

then the conjugate prior for $(p_1, \ldots, p_k)$ is $D_{k-1}(a_1, \ldots, a_k)$. The posterior, as a result, has the form,

$$(p_1, \ldots, p_k | x) \sim D_{k-1}(a_1 + x_1, \ldots, a_k + x_k),$$

where $x = (x_1, \ldots, x_k)$.

Dirichlet Process

A Dirichlet process, $DP(\alpha G_0)$, expressed as G is with base distribution G_0 and scale parameter α. G is a random probability measure that has the same support as G_0. It is also a distribution over distributions. Ferguson (1973) introduced the Dirichlet process as a class of prior distributions for which the support is large, and the posterior distribution is analytically manageable. The idea of using a Dirichlet process as the prior for the mixing proportions of a simple distribution (e.g., Gaussian) was first introduced by Antoniak (1974).

Consider a model with a parametric likelihood: $y_i \sim N(\theta_i, \tau_i^{-1})$. Instead of assuming $\theta_i \sim G_0$, we could specify $\theta_i \sim G$, and $G \sim DP(\alpha G_0)$, where G_0 is the base distribution such as a normal distribution and α is a precision parameter determining how closely G follows G_0. The Dirichlet process (DP) model is simplified in practice by the Polya urn representation (Blackwell and MacQueen 1973). It relies on marginalizing out G to obtain

$$(\theta_i | \theta_1, \ldots, \theta_{i-1}) \sim (\frac{\alpha}{\alpha + i - 1}) G_0 + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + 1} \right) \delta_{\theta_j}. \tag{1.1}$$

It can be seen that the θ_i's are distributed as the base measure along with the added property that $P(\theta_i = \theta_j) > 0$ for $i \neq j$. The Dirichlet process prior results in what MacEachern (1994) calls a cluster structure among the θ_i's. This cluster structure partitions the n θ_i's into k sets or clusters, $0 < k \leq n$. All of the observations in a cluster share an identical value of θ and subjects in different clusters have different values of θ.

The Chinese restaurant representation can be viewed as an analogy. Say, for instance, that we have a restaurant with infinitely many tables with X_n representing the patrons of the restaurant. Let

1. 1^{st} customer sits at table with dish ζ_1;

2. 2^{nd} customer sits at first table with probability $\alpha/(1+\alpha)$ or new table with dish ζ_2 with probability $1/(1+\alpha)$;

3. process encourages later customers to sit at well occupied tables.

We can see from (1.1) that a customer is more likely to sit at a table if there are already many people sitting there. However, with probability proportional to α, the customer will sit at a new table.

Another popular way of presenting DP, introduced by Sethuraman (1994), is called the stick-breaking process:

$$\theta_i \sim G = \sum_{h=1}^{\infty} \nu_h \delta_{\theta_h},$$

where

$$\nu_h = V_h \prod_{l<h}(1 - V_l), V_h \sim Beta(1, \alpha), \theta_h \sim G_0,$$

for $h = 1, \ldots, \infty$, with δ_θ denoting the degenerate distribution with all its mass at θ. One can illustrate this by starting from a unit probability stick,

1. Break off a random piece (V_1) and allocate this to a random value (θ_1);

2. From the remaining $1 - V_1$, break off a proportion V_2 and allocate to θ_2;

3. Repeat infinitely many times.

Compared with the Polya urn scheme, the stick-breaking process is more attractive in the sense that it provides the ability to conduct inference on G by avoiding marginalization of G.

Use of Dirichlet process mixture models in Bayesian non-parametrics has become computationally feasible with the development of Markov chain methods for sampling from the posterior distribution of the parameters of the component distributions and/or of the associations of mixture components with observations. Methods based on Gibbs sampling can easily be implemented for models based on conjugate prior distributions.

1.1.2 Tensor Decomposition

Let $T_{d_1 \dots d_p}$ denote the set of all tensors of dimension $d_1 \times \dots \times d_p$, and $\mathbf{\Pi}_{d_1 \dots d_p} \subset T_{d_1 \dots d_p}$ denote the set of all probability tensors, so that $\boldsymbol{\pi} \in \mathbf{\Pi}_{d_1 \dots d_p}$ implies

$$\boldsymbol{\pi} = \left\{ \pi_{c_1 \dots c_p} \geq 0, \ c_j = 1, \dots, d_j, j = 1, \dots, p : \ \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \pi_{c_1 \dots c_p} = 1 \right\}.$$

One way of tensor generalization of the matrix singular value decomposition is called PARAFAC decomposition (Harshman 1970; Harshman and Lundy 1994; Zhang and Golub 2001). Kolda (2001) used the notation

$$\mathbf{D} = \sum_{h=1}^{k} \lambda_h \mathbf{U}_h, \ \mathbf{U}_h = \mathbf{u}_h^{(1)} \otimes \mathbf{u}_h^{(2)} \otimes \cdots \otimes \mathbf{u}_h^{(p)},$$

where $\lambda_1 \geq \cdots \geq \lambda_k > 0$, $\mathbf{U}_h$ is a decomposed tensor with $\mathbf{u}_h^{(j)} \in \Re^{d_j}$, and $\otimes$ denotes the outer product, so that

$$D_{c_1 \dots c_p} = \sum_{h=1}^{k} \lambda_h u_{hc_1}^{(1)} \dots u_{hc_p}^{(p)}.$$

One definition of the rank of a tensor is the minimal k such that $\mathbf{D}$ can be expressed as a sum of k decomposed (or rank one) tensors.

Tucker (1966) proposed a different decomposition for three-way data, which was later extended to arbitrary tensors by De Lathauwer et al. (2000). The Tucker decomposition, or higher-order singular value decomposition (HOSVD) aims to decompose a tensor $\mathbf{D} \in T_{d_1 \dots d_p}$ as

$$D_{c_1 \dots c_p} = \sum_{h_1=1}^{d_1} \cdots \sum_{h_p=1}^{d_p} g_{h_1 \dots h_p} u_{h_1 c_1}^{(1)} \dots u_{h_p c_p}^{(p)},$$

where $G = \{g_{h_1 \dots h_p}\} \in T_{d_1 \dots d_p}$ is called a core tensor and its entries control interaction between the different components. Wang and Ahuja (2005); Kim and Choi (2007) and Yang and Dunson (2013) empirically noted that the HOSVD achieves better data compression and requires fewer components compared to the PARAFAC model as it uses all combinations of the mode vectors $u_{h_j}^{(j)}$'s, $h = 1, \dots, k$.

With an interest in decomposing a probability matrix/tensor, the non-negative matrix factorization (NMF) (Gregory and Pullman 1983; Cohen and Rothblum 1993) seeks the best approximation of a non-negative matrix $A \in \Re_+^{m \times n}$ as a product of non-negative matrices $W \in \Re_+^{m \times k}$ and $V \in \Re_+^{k \times n}$ for some $k \leq \min\{m, n\}$, and finds the so-called non-negative rank as the minimal k such that a non-negative matrix can be written as a sum of rank one non-negative matrices. The non-negative versions of the PARAFAC and HOSVD decompositions for tensors are discussed in Kim and Choi (2007) and Shashua and Hazan (2005).

1.1.3 Bayes Nonparametrics

Before turning to the non-parametric model, first consider the fully parametric situation. Suppose y_i is an $n_i \times 1$ random vector indexed by the $p \times 1$ parameter vector θ_i, for each $i = 1, \dots, n$. Suppose the θ_i have a prior distribution with hyperparameter θ_0. That is, $\theta_i \overset{i.i.d.}{\sim} G(\cdot|\theta_0)$. If $G(\cdot|\theta_0)$ is a specified function, then this corresponds to the fully parametric situation. The fully parametric situation can be described by two stages:

- Stage 1: $(y_i|\theta_i) \sim$ (parametric likelihood function) $(i = 1, \dots, n)$,

- Stage 2: $(\theta_i|\theta_0) = G(\cdot|\theta_0)$,

where $G(\cdot|\theta_0)$ is a specified prior distribution, such as a normal, gamma, exponential, beta, etc..

In many parametric likelihood models, we often wish to relax the assumption of a parametric prior on the parameters. A common method is to set the prior distribution to be random such as a Dirichlet process prior, which leads to mixtures of Dirichlet processes (MDP). MDP removes the parametric assumption on $G(\cdot|\theta_0)$, that is $G(\cdot|\theta_0)$ is not known and thus no functional form is specified for G. Thus the MDP model has 3 stages

- Stage 1: $(y_i|\theta_i) \sim$ (parametric likelihood),
- Stage 2: $\theta_i \overset{i.i.d.}{\sim} G$ (G unknown),
- Stage 3: $G|\alpha_0, G_0 \sim DP(\alpha_0 G_0)$.

Thus the MDP model has 3 stages with the last stage being the DP specification. The specification given above is semi-parametric in the sense that a parametric likelihood specification is given in stage 1, and a non-parametric specification is given in stages 2 and 3. Some examples of Bayes semi-parametric methods can be found in MacLehose et al. (2007); Dunson et al. (2008). If we further relax the stage 1 to model the data using a probabilistic characterization of uncertainty while accounting for interaction, such as a tensor factorization, it becomes a Bayes nonparametric model because both the distribution of the data, y_i, and the distribution of the parameter θ are nonparametric.

1.1.4 Joint and Conditional Probabilistic Modeling

Our focus is on sparse non-parametric modeling of the cell probabilities, $\boldsymbol{\pi} = \{\pi_{c_1 \ldots c_p}\}$ with $\pi_{c_1 \ldots c_p} = \Pr(x_{i1} = c_1, \ldots, x_{ip} = c_p)$. Dunson and Xing (2009) incorporated a latent structure model (Lazarsfeld and Henry 1968; Goodman 1974) with a probabilistic version of PARAFAC tensor decomposition by representing π as

$$\pi_{c_1 \dots c_p} = \Pr(x_{i1} = c_1, \dots, x_{ip} = c_p) = \sum_{h=1}^{k} \nu_h \lambda_{hc_1}^{(1)} \dots \lambda_{hc_p}^{(p)}, \tag{1.2}$$

where $\boldsymbol{\nu} = (\nu_1, \dots, \nu_k)'$ is a vector of mixture probabilities, $z_i \in \{1, \dots, k\}$ is a latent class index, $\Pr(x_{ij} = c_j | z_i = h) = \lambda_{hc_j}^{(j)}$ is the probability of $x_{ij} = c_j$ given allocation of individual i to class h. $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})'$ are assumed to be conditionally independent given z_i. Marginalizing over the distribution of z_i induces dependence among the p variables. Note that it is different from a usual PARAFAC decomposition because of the non-negativity constraints on $\boldsymbol{\nu}$ and the $\boldsymbol{\lambda}_h^{(j)}$'s. In this paper, it is proved that any multivariate categorical data distribution can be characterized as a finite mixture of product-multinomial distributions as in (1.2).

However, it is not straightforward to obtain a well-justified approach for estimation of k. Regular methods like maximum likelihood estimation would fail to converge due to the sparsity of the data even for a modest k, or otherwise would provide biased results if small k is chosen. These issues provide motivation for utilizing a Dirichlet process, which avoids selection of a single finite k, allowing the number of components that are occupied by individuals in the sample to grow with sample size. One can specify priors

$$\begin{aligned} \boldsymbol{\lambda}_h^{(j)} &\sim \text{Dirichelet}(a_{j1}, \dots, a_{jd_j}), \\ \boldsymbol{\nu} &\sim \text{Dirichlet process}, \end{aligned} \tag{1.3}$$

using stick-breaking representation for $\boldsymbol{\nu}$, (1.2) and (1.3) can be expressed in the following hierarchical form:

$$\begin{aligned} x_{ij}|z_i = h &\sim \text{Multinomial}\big((1,\ldots,d_j); \lambda_{h1}^{(j)},\ldots,\lambda_{hd_j}^{(j)}\big), \\ z_i &\sim \sum_{h=1}^{\infty} V_h \prod_{l<h}(1-V_l)\delta_h, \\ V_h &\sim \text{Beta}(1,\alpha), \\ \boldsymbol{\lambda}_h^{(j)} \equiv (\lambda_{h1}^{(j)},\ldots,\lambda_{hd_j}^{(j)}) &\sim \text{Diri}(a_{j1},\ldots,a_{jd_j}). \end{aligned} \tag{1.4}$$

The Gibbs sampling algorithm can be performed in a straightforward fashion. Bhattacharya and Dunson (2012) instead applied a different decomposition method to model the joint probability for multivariate categorical data. In comparison, rather than investigating the dependence structure among variables, Yang and Dunson (2013) established a conditional probabilistic Tucker factorization with the goals of classifying the response of interest as well as identifying a sparse subset of important predictors. That is,

$$\Pr(y_i = c \,|\, x_{i1} = c_1, \ldots, x_{ip} = c_p) = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \lambda_{h_1 \cdots h_p}(c) \prod_{j=1}^{p} \pi_{h_j}^{(j)}(c_j), \tag{1.5}$$

with constraints $\sum_{c=1}^{d_0} \lambda_{h_1 \ldots h_p}(c) = 1$ and $\sum_{h=1}^{k_j} \pi_h^{(j)}(c_j) = 1$. The value of $k_j \in \{1,\ldots,d_j\}$ controls the number of parameters characterizing the impact of the j^{th} predictor on the conditional probability, with $k_j = 1$ implying that the j^{th} predictor is excluded from the model. We can simplify the representation by introducing p latent class indicators $z_{i1},\ldots,z_{ip}$ for $x_{i1},\ldots,x_{ip}$. The model can be rewritten as

$$y_i|z_{i1},\ldots,z_{ip} \sim \text{Multinomial}(\{1,\ldots,d_0\}; \lambda_{z_{i1},\ldots,z_{ip}}), \tag{1.6}$$

$$z_{ij}|x_{ij} = c_j \sim \text{Multinomial}(\{1,\ldots,k_j\}; \pi_1^{(j)}(c_j),\ldots,\pi_{k_j}^{(j)}(c_j)), \tag{1.7}$$

where $\lambda_{z_{i1},\ldots,z_{ip}} = \{\lambda_{z_{i1},\ldots,z_{ip}}(1), \ldots, \lambda_{z_{i1},\ldots,z_{ip}}(d_0)\}$. Integrating out the latent class indicators, the conditional probability of y_i given $(x_{i1}, \ldots, x_{ip})$ matches the form in (1.5). The Dirichlet distribution priors are chosen for $\boldsymbol{\lambda}_{h_1 \ldots h_p}$ and $\boldsymbol{\pi}^{(j)}(c_j)$ to maintain conjugacy, while some well-specified discrete distribution are specified for k_j favoring sparsity. Refer to Yang and Dunson (2013) for deriving the corresponding posteriors. Although it seems to have better prediction performance than existing methods and it has the capability of interpreting the relationship between predictors and the outcome, it is worth noting that the approximation of marginal likelihood of $k = \{k_1, \ldots, k_p\}$ was not justified in the paper and there are some computational issues when the number of variables with $(k_j > 1)$ is bigger than seven.

In summary, both joint and conditional modeling have advantages of (i) allowing the distribution of multiple categorical variables to be unknown; (ii) a full probabilistic characterization of uncertainty accounting for any possible interaction among predictors; (iii) favoring a sparse structure that allows efficient computation without the problem of overfitting. Note that the joint model aims to infer the dependence structure among variables, while the conditional model focuses on the classification.

1.2 Sample Size Re-estimation Introduction(Paper 3)

Clinical trials with longitudinal endpoints are very common. A key issue in designing such a trial is to determine how large of a study is necessary to detect a clinically important difference with a desired power. A traditional approach of sample size calculation for fixed design requires the investigator to specify a clinically meaningful difference to be detected, the significance level, a desired level of power and any additional nuisance parameters (e.g. the error variance for continuous data, the control group response rate for binary data). As for repeated measure endpoints, Lu et al. (2008; 2009) generalized a formula for calculating the sample size with nuisance parameters containing

(1) correlation among longitudinal visits; (2) standard deviation within longitudinal measurements for each subject and (3) retention rates in both treatment groups. For planning purposes, best guesses are made for the value of the nuisance parameters.

However, there is a great concern that these assumptions of nuisance parameters based on previous studies are often unreliable because of differences in the study population, changes in medical practice, or the measurement techniques. Since incorrect assumptions can lead to substantial underpowering or overpowering to detect the clinically important difference, it may be prudent to check the validity of those assumptions using interim data from the study. There is a rich literature Coffey and Kairalla (2008); Chuang-Stein et al. (2006) discussing the sample size re-estimation methods to rescue the power. Wittes and Brittain (1990) introduced the concept of an internal pilot design, which re-estimates the sample size in the mid-course of the study with no interim testing involved. Internal pilot designs have also been extended to different settings, besides normally distributed outcomes, such as repeated measures. Shih and Gould (1995) described a method to re-estimate sample size in the repeated measure framework. However it is only for a simplified setting, where the parameter of interest is the rate of change (slope) of a continuous measurement. Zucker and Denne (2002) extended Shih and Gould's model to a general setting in which missing and dropout are allowed and a linear combination of treatment effect over time can be set as the meaningful difference to be detected.

Group sequential design Jennison and Turnbull (2000) promises to be more efficient because we are given an opportunity to terminate the study before the planned completion if there is strong evidence that the treatment effect is meaningfully large or the treatment is unlikely to be better than the control group. This design can benefit plenty of longitudinal trials. For instance, suppose we are doing a trial of weight loss, and the primary endpoint is weight loss at one year, with other measures at 3, 6 and 9 months.

At the time of an interim analysis, some patients will have less than full follow-up, but will have some follow-up measurements indicating a trend in their weight. If no weight loss is seen early in follow-up, it may be reasonable to stop a trial for futility. On the other hand, if substantial weight loss is observed and maintained, a convincing efficacy finding may be resulted prior to the final planned analysis. Another example could be a trial of Alzheimers disease in which an endpoint indicating cognitive decline such as Alzheimers Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) may be used. While the primary timepoint of interest may be after 18 months of treatment, intermediate measures may be taken at 6 and 12 months of follow-up. At the time of an interim analysis, some patients will have less than full follow-up, but will have some follow-up measures indicating a trend in cognitive function that may be useful. An interim analysis may be able to stop a trial for futility or, if done later in the trial, may provide convincing results prior to the planned final analysis. While bounds may be set to be broad in order to avoid a premature stop, having bounds may provide useful guidance to a Data Monitoring Committee to help them avoid an early stop that may be due to a spurious finding. Both Galbraith and Marschner (2003) and Kittelson et al. (2005) discussed sequential methods when monitoring trials with longitudinal endpoints as well as making use of people who have not completed the study. Kittelson et al. (2005) also provided a nice discussion when the outcomes are not measured according to the pre-trial schedule. However, to adjust for the sequential monitoring stopping rules, both of them used the estimated information at the end of the study in computing the information-timing rather than the fixed maximal information from the pre-trial design. Moreover, the two papers did not address the potential problem of insufficient power due to the incorrect initial sample size calculation if the variance assumption is incorrect. Burington and Emerson (2003) focused on making flexible group sequential stopping rules when the actual interim analyses deviate from the design with respect to the number and timing. One can either choose to maintain the power or maintain the maximal sample size. But it did not cover the case where the primary endpoint of interest is longitudinal. Thus, with the goal of designing and analyzing a longitudinal trial using group sequential design along with the concern of insufficient power, it is natural to combine internal pilot designs into group sequential design in the longitudinal framework. Mehta and Tsiatis (2001) and Tsiatis (2006) initiated the use of information-based monitoring for implementing internal pilot designs in conjunction with group sequential methods, but only for normal and binary endpoints. The counterpart for the longitudinal setting is missing, yet not trivial. A new design for longitudinal trials, namely the information-based sample size re-estimation method, will be developed in paper 3 of my dissertation with a thorough discussion through simulations and application.

CHAPTER 2

Sparse Tensor Factorizations for Big Contingency Tables

2.1 Introduction

Sparsely observed big tabular data sets are commonly collected in many applied domains. One example corresponds to recommender systems in which the dimensions of the table correspond to users, items and different contexts (Karatzoglou et al. (2010)), with a tiny proportion of the cells filled in for users providing rankings. The task is to fill in the rest of the huge table in order to make recommendations to users of which items they may prefer in each context. This extends the widely studied matrix completion problem (Candès and Recht (2009)) of which the Netflix challenge was one example. Another setting corresponds to contingency tables in which multivariate categorical data are collected for each individual, and the cells of the table contain counts of the number of individuals having a particular combination of values. In contingency table analyses, the focus is typically on inferring associations among the different variables, but challenges arise when there are many variables, so that the number of cells in the table is vastly bigger than the sample size.

Suppose that the tensor of interest is $\pi \in \Pi_{d_1 \times \cdots \times d_p}$, with $\Pi_{d_1 \times \cdots \times d_p}$ a space of p-way tensors having d_j rows in the jth direction. Often there are constraints on the elements of the tensor. For recommender systems, ratings are non-negative so that one is faced with a non-negative tensor factorization problem (Paatero and Tapper (1994);

Lee and Seung (1999); Friedlander and Hatz (2005); Lim and Comon (2009); Liu et al. (2012)). For contingency tables, the tensor corresponds to the joint probability mass function for multivariate categorical data, so that the elements are non-negative and add to one across all the cells (Dunson et al. (2008); Bhattacharya and Dunson (2012)). Let Y denote the data collected on tensor π. For recommender systems, Y consists of ratings for a small subset of the $\prod_{j=1}^{p} d_j$ cells in the tensor, while for contingency tables Y includes response vectors $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$ for subjects $i = 1, \ldots, n$, with $y_{ij} \in \{1, \ldots, d_j\}$ for $j = 1, \ldots, p$. In both cases, data are extremely sparse, with no observations in the overwhelming majority of cells.

To combat this data sparsity, it is necessary to substantially reduce dimensionality in estimating π. The usual way to accomplish this is through a low rank assumption. Unlike for matrices, there is no unique definition of rank but the most common convention is to define the rank k of a tensor π as the smallest value of k such that π can be expressed as

$$\pi = \sum_{h=1}^{k} \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)}, \tag{2.1}$$

which is sum of k rank one tensors, each an outer product of vectors[1] for each dimension (Kolda and Bader 2009). Expression (1) is commonly referred to as parallel factor analysis (PARAFAC) (Harshman (1970); Bro (1997)). For k small, the number of parameters is massively reduced from $\prod_{j=1}^{p} d_j$ to $k \sum_{j=1}^{p} d_j$; as the low rank assumption often holds approximately, this leads to an effective approach in many applications, and a rich variety of algorithms are available for estimation.

However, the decrease in degrees of freedom from exponential in p to linear in p is not sufficient when p is big. Large p small n problems arise routinely, and a usual solution

[1] For $p = 2$, $\psi^{(1)} \otimes \psi^{(2)} = \psi^{(1)}\psi^{(2)\mathrm{T}}$. In general, $(\psi^{(1)} \otimes \cdots \otimes \psi^{(p)})_{c_1 \ldots c_p} = \psi_{c_1}^{(1)} \ldots \psi_{c_p}^{(p)}$

outside of tensor settings is to incorporate sparsity. For example, in linear regression, many of the coefficients are set to zero (Tibshirani 1996; Scott and Berger 2010), while in estimation of large covariance matrices, sparse factor models are used that assume few factors and many zeros in the factor loadings matrices (West (2003); Carvalho et al. (2008)). In the matrix factorization literature, there has been consideration of low rank plus sparse decompositions (Chartrand (2012)), but this approach does not solve our problem of too many parameters. Including zeros in the component vectors $\{\psi_h^{(j)}\}$ is not a viable solution, particularly as we do not want to enforce exact zeros in blocks of the tensor π but require an alternative notion of sparsity.

Our notion is as follows. For component h ($h = 1, \ldots, k$), we partition the dimensions into two mutually exclusive subsets $S_h \cup S_h^c = \{1, \ldots, p\}$. The proposed sparse PARAFAC (sp-PARAFAC) factorization is then

$$\pi = \sum_{h=1}^{k} \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)}, \quad \psi_h^{(j)} = \psi_0^{(j)} \text{ for } j \in S_h^c. \tag{2.2}$$

Hence, instead of having to introduce a separate vector $\psi_h^{(j)}$ for every h and j, we allow there to be more degrees of freedom used to characterize the tensor structure in certain directions than in others. Consider the recommender systems application and suppose we have three dimensions, including users ($j = 1$), items ($j = 2$) and context ($j = 3$). If we let $\psi_h^{(3)} = \psi_0^{(3)}$ for $h = 1, \ldots, k$,

$$\pi_{c_1 c_2 c_3} = \psi_{0c_3}^{(3)} \sum_{h=1}^{k} \psi_{hc_1}^{(1)} \psi_{hc_2}^{(2)}, \tag{2.3}$$

so that we factorize the user-item matrix as being of rank k, and then include a multiplier specific to each level of the context factor. This assumes that users rank systematically higher or lower depending on context but there is no interaction. In the contingency table application, $\Pr(y_{i1} = c_1, \ldots, y_{ip} = c_p) = \pi_{c_1 \cdots c_p}$. If $j \in S_h^c$ for $h = 1, \ldots, k$, then the jth variable is independent of the other variables with $\Pr(y_{ij} = c_j) = \psi^{(j)}_{0c_j}$. By including $j \in S_h^c$ for some but not all $h \in \{1, \ldots, k\}$ one can use fewer degrees of freedom in characterizing the interaction between the jth factor and the other factors. In practice, we will learn $\{S_h\}$ using a Bayesian approach, as the appropriate lower dimensional structure is typically not known in advance.

We conjecture that many tensor data sets can be concisely represented via (2.2), with results substantially improved over usual PARAFAC factorizations due to the second layer of dimension reduction. For concreteness and brevity, we focus on contingency tables, but the methods are easily modified to other settings. Contingency table analysis is routine in practice; refer to Agresti (2002); Fienberg and Rinaldo (2007). However, in stark contrast to the well developed literature on linear regression and covariance matrix estimation in big data settings, very few flexible methods are scalable beyond small tables. Throughout the rest of the paper, we assume that the observed data $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}, i = 1, \ldots, n$, is multivariate unordered categorical, with $y_{ij} \in \{1, \ldots, d_j\}$. Our interest is in situations where the dimensionality p is comparable or even larger than the number of samples n.

2.2 Sparse Factor Models for Tables

2.2.1 Model and prior

We focus on a Bayesian implementation of sp-PARAFAC in (2.2). Let $\mathcal{S}^{r-1} = \{x \in \Re^r : x_j \geq 0, \sum_{j=1}^r x_j = 1\}$ denote the $(r-1)$-dimensional probability simplex. In the contingency table case, Dunson et al. (2008) proposed the following probabilistic PARAFAC factorization.

$$\Pr(y_{i1} = c_1, \ldots, y_{ip} = c_p) = \pi_{c_1 \cdots c_p} = \sum_{h=1}^{k} \nu_h \prod_{j=1}^{p} \lambda^{(j)}_{hc_j}, \tag{2.4}$$

where $\nu = \{\nu_h\} \in \mathcal{S}^{k-1}$ and $\lambda_h^{(j)} = (\lambda_{h1}^{(j)}, \ldots, \lambda_{hd_j}^{(j)}) \in \mathcal{S}^{d_j-1}$ is a vector of probabilities of $y_{ij} = 1, \ldots, d_j$ in component h. Introducing a latent sub-population index $z_i \in \{1, \ldots, k\}$ for subject i, the elements of y_i are conditionally independent given z_i with $\Pr(y_{ij} = c_j \mid z_i = h) = \lambda_{hc_j}^{(j)}$, and marginalizing out the latent index z_i leads to a mixture of product multinomial distribution for y_i. Placing Dirichlet priors on the component vectors leads to a simple and efficient Gibbs sampler for posterior computation. We will refer to this model (2.4) as standard PARAFAC.

This approach has excellent performance in small to moderate p problems, but as p increases there is an inevitable breakdown point. The number of parameters increases linearly in p, as for other PARAFAC factorizations, so problems arise as p approaches the order of n or $p \gg n$. For example, we are particularly motivated by epidemiology studies collecting many categorical predictors, such as occupation type, demographic variables, and single nucleotide polymorphisms. For continuous response vectors $y_i \in \Re^p$, there is a well developed literature on Gaussian sparse factor models that are adept at accommodating $p \gg n$ data (West (2003); Lucas et al. (2006); Carvalho et al. (2008); Bhattacharya and Dunson (2011)). These models include many zeros in the loadings matrices to induce additional dimension reduction on top of the low rank assumption. Pati et al. (2013) provided theoretical support through characterizing posterior concentration.

Our sp-PARAFAC factorization provides an analog of sparse factor models in the tensor setting. Modifying for the categorical data case, we let

$$\pi_{c_1 \ldots c_p} = \sum_{h=1}^{k} \nu_h \prod_{j \in S_h} \lambda_{hc_j}^{(j)} \prod_{j \in S_h^c} \lambda_{0c_j}^{(j)}, \tag{2.5}$$

where $|S_h| \ll p$ ($|S|$ denotes the cardinality of a set S) and the $\lambda_0^{(j)}$ vectors are *fixed in* *advance*; we consider two cases:

$$(i)\ \lambda_0^{(j)} = \left(\frac{1}{d_j}, \ldots, \frac{1}{d_j}\right)^{\mathrm{T}} \quad \text{and} \quad (ii)\ \lambda_0^{(j)} = \left(\frac{1}{n}\sum_{i=1}^{n} 1(y_{ij} = 1), \ldots, \frac{1}{n}\sum_{i=1}^{n} 1(y_{ij} = d_j)\right)^{\mathrm{T}},$$

corresponding to a discrete uniform and empirical estimates of the marginal category probabilities. By fixing the baseline dictionary vectors $\{\lambda_0^{(j)}\}$ in advance, and allocating a large subset of the variables within each cluster h to the baseline component, we dramatically reduce the size of the model space. In particular, the probability tensor π in (2.5) can be parameterized as $\theta_\pi = l(\nu, \{S_h\}_{1\le h\le k}, \{\lambda_h^{(j)}\}_{1\le h\le k, j\in S_h})$, where $\nu \in \mathcal{S}^{k-1}, S_h \subset \{1, \ldots, p\}, \lambda_h^{(j)} \in \mathcal{S}^{d_j-1}$. Thus, the effective number of model parameters is now reduced to $(k-1)+\sum_{h=1}^{k}|S_h|+\sum_{h=1}^{k}\sum_{j\in S_h}(d_j-1)$, which is substantially smaller than the $(k-1)+\sum_{j=1}^{p} k(d_j-1)$ parameters in the original specification, provided $|S_h| \ll p$ for all $h = 1, \ldots k$. The size of S_h is penalized via a sparsity favoring prior on $|S_h|$ in (2.6) below. We will illustrate that this can lead to huge differences in practical performance.

Completing a Bayesian specification with priors for the unknown parameter vectors and expressing the model in hierarchical form, we have[2]

$$\begin{gathered}
y_{ij} \sim \mathrm{Mult}\big(\{1, \ldots, d_j\}; \lambda_{z_i 1}^{(j)}, \ldots, \lambda_{z_i d_j}^{(j)}\big), \\
\lambda_h^{(j)} \sim (1-\tau_h)\delta_{\lambda_0^{(j)}} + \tau_h \mathrm{Diri}(a_{j1}, \ldots, a_{jd_j}), \\
\Pr(z_i = h) = \nu_h = V_h \textstyle\prod_{l<h}(1-V_l), \\
V_h \sim \mathrm{Beta}(1, \alpha), \quad \alpha \sim \mathrm{Gamma}(a_\alpha, b_\alpha), \quad \tau_h \sim \mathrm{Beta}(1, \gamma).
\end{gathered} \tag{2.6}$$

It is evident that the hierarchical prior in (2.6) is supported on the space of probability tensors with a sp-PARAFAC decomposition as in (2.5), since (2.6) is equivalent to

[2] $\mathrm{Mult}\big(\{1, \ldots, d\}; \lambda_1, \ldots, \lambda_d\big)$ denotes a discrete distribution on $\{1, \ldots, d\}$ with probabilities $\lambda_1, \ldots, \lambda_d$ associated to each atom.

letting the subset-size $|S_h| \sim \text{Binom}(p, \tau_h)$ and drawing a random subset S_h uniformly from all subsets of $\{1, \ldots, p\}$ of size $|S_h|$ in (2.5). A stick-breaking prior (Sethuraman 1994) is chosen for the component weights $\{\nu_h\}$, taking a nonparametric Bayes approach that allows $k = \infty$, with a hyperprior placed on the concentration parameter α in the stick-breaking process to allow the data to inform more strongly about the component weights. The probability of allocation τ_h to the *active* (non-baseline) category in component h is chosen as beta$(1, \gamma)$, with $\gamma > 1$ favoring allocation of many of the $\lambda_h^{(j)}$s to the baseline category $\lambda_0^{(j)}$. In the limiting case as $\gamma \to \infty$, the joint probability tensor π becomes an outer product of the baseline probabilities for the individual variables, $\pi = \lambda_0^{(1)} \otimes \cdots \otimes \lambda_0^{(p)}$. On the other hand, as $\gamma \to 0$, one reduces back to standard PARAFAC (2.4).

Line 2 of expression (2.6) is key in inducing the second level of dimensionality reduction in our Bayesian sparse PARAFAC factorization. The inclusion of the baseline component that does not vary with h massively reduces the number of parameters, and can additionally be argued to have minimal impact on the flexibility of the specification. The $\lambda_h^{(j)}$s are incorporated within $\prod_{j=1}^{p} \lambda_{hc_j}^{(j)}$, which for large p is highly concentrated around its mean since the $\lambda_h^{(j)}$'s are independent across j. This is a manifestation of the concentration of measure phenomenon (Talagrand 1996), which roughly states that a random variable that depends in a smooth way on the influence of many independent variables, but not too much on any one of them, is essentially constant. For example, if $\theta_j \overset{iid}{\sim} U(0, 1)$ and $\Theta = \prod_{j=1}^{p} \theta_j$, then $\text{E}(\Theta) = (1/2)^p$ and $\text{var}(\Theta) = (1/3)^p$, which rapidly converges to zero. This implies that replacing a large randomly chosen subset of the $\lambda_h^{(j)}$s by $\lambda_0^{(j)}$ should have minimal impact on modeling flexibility.

2.2.2 Induced prior in log-linear parameterization

An important challenge is accommodating higher order interactions, which play an important role in many applications (e.g., genetics), but are typically assumed to equal zero for tractability. As p grows, it is challenging to even accommodate two-way interactions in traditional categorical data models (log-linear, logistic regression) due to an explosion in the number of terms. In contrast, the tensor factorization does not explicitly parameterize interactions, but indirectly induces a shrinkage prior on the terms in a saturated log-linear model. One can then reparameterize in terms of the log-linear model in conducting inferences in a post model-fitting step. We illustrate the induced priors on the main effects and interactions below.

For ease of exposition, we first focus on a case where $p = 3$ and $d_j = d = 2$ for $j = 1, \dots, 3$. We generate $10,000$ random probability tensors $\pi^{(t)} = (\pi^{(t)}_{c_1c_2c_3}), t = 1, \dots, 10000$, distributed according to (2.6), where we fix the baseline $\lambda_0^{(j)} = (1/2, 1/2)$ for all j. Given a $2 \times 2 \times 2$ tensor π, we can equivalently characterize π in terms of its log-linear parameterization

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{13}, \beta_{23}, \beta_{123})^{\mathrm{T}},$$

consisting of 3 main effect terms $\beta_1, \beta_2, \beta_3$, three second-order interaction terms $\beta_{12}, \beta_{13}, \beta_{23}$ and one third order interaction term β_{123}; refer to §5.3.5 of Agresti (2002). Given each prior sample $\pi^{(t)}$, we equivalently obtain a sample $\boldsymbol{\beta}^{(t)}$ from the induced prior on $\boldsymbol{\beta}$, which allows us to estimate the marginal densities of the main effects and interactions and also their joint distributions. In particular, since γ plays an important role in placing weights on the baseline component, we would like to see how our induced priors differ with different γ values.

In our simulation exercise, we fix three values of γ, namely, $\gamma = 1, 5, 20$. Note that

$\gamma = 1$ corresponds to a $U(0,1)$ prior on τ_h. For different values of γ, we show the histograms of one main effect term β_1, one two-way interaction β_{12} and the three-way interaction β_{123} in Figure 2.1. Table 2.1 additionally reports summary statistics.

In high-dimensional regression, $y_i = x_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$, there has been substantial interest in shrinkage priors, which draw β_j *a priori* from a density concentrated at zero with heavy tails. Such priors strongly shrink the small coefficients to zero, while limiting shrinkage of the larger signals (Park and Casella 2008; Carvalho et al. 2010; Polson and Scott 2010; Hans 2011; Armagan et al. 2013). In Figure 2.1, the induced prior on any of the log-linear model parameters is symmetric about zero, with a large spike very close to zero, and heavy tails. Thus, we have indirectly induced a continuous shrinkage prior on the main effects and interactions through our tensor decomposition approach. In addition, the prior automatically shrinks more aggressively as the interaction order increases. Such greater shrinkage of interactions is commonly recommended (Gelman et al. 2008). Importantly, we do not zero out small interactions but allow many small coefficients, which is an important distinction in applications, such as genomics, having many small signals.

2.3 Posterior Computation

Under model (2.6), we can easily proceed to draw posterior samples from a Gibbs sampler since all the full conditionals have recognizable forms. The algorithm iterates through the following steps:

1. For variable $j = 1, \ldots, p$ and latent class $h = 1, \ldots, k^*$, where $k^* = \max\{z_1, \ldots, z_n\}$, update $\lambda_h^{(j)} \equiv (\lambda_{h1}^{(j)}, \ldots, \lambda_{hd_j}^{(j)})$ from a two component mixture distribution, having

a point mass at the baseline probability:

$$\begin{aligned}(\lambda_h^{(j)}|-) &= w_{0h}^{(j)}\delta_{\lambda_0^{(j)}} + w_{1h}^{(j)}\text{Diri}\bigg(a_{j1} + \sum_{i=1}^{n} 1(y_{ij} = 1, z_i = h), \\ &\qquad \ldots, a_{jd_j} + \sum_{i=1}^{n} 1(y_{ij} = d_j, z_i = h)\bigg), \qquad (2.7)\end{aligned}$$

where $w_{0h}^{(j)}$ and $w_{1h}^{(j)}$ are the mixture weights:

$$\begin{aligned}w_{0h}^{(j)} &= \frac{(1-\tau_h)\prod_{c=1}^{d_j}\lambda_{0c}^{(j)\sum_{i=1}^{n}1(z_i=h,y_{ij}=c)}}{(1-\tau_h)\prod_{c=1}^{d_j}\lambda_{0c}^{(j)\sum_{i=1}^{n}1(z_i=h,y_{ij}=c)} + \tau_h\frac{\Gamma(\sum_{c=1}^{d_j}a_{jc})}{\prod_{c=1}^{d_j}\Gamma(a_{jc})}\cdot\frac{\prod_{c=1}^{d_j}\Gamma\left(a_{jc}+\sum_{i=1}^{n}1(z_i=h,y_{ij}=c)\right)}{\Gamma\left(\sum_{c=1}^{d_j}a_{jc}+\sum_{i=1}^{n}1(z_i=h)\right)}}, \\ w_{1h}^{(j)} &= 1 - w_{0h}^{(j)}.\end{aligned}$$

2. Let $\eta_{hj} \in \{0, 1\}$ be a binary allocation variable indicating the component $\lambda_h^{(j)}$ is drawn from in (2.7), with $\eta_{hj} = 0$ if $\lambda_h^{(j)}$ is updated from the baseline component. Update τ_h, $h = 1, \ldots, k^*$ from a Beta full conditional:

$$\tau_h|- \sim \text{Beta}\bigg(1 + \sum_{j=1}^{p} 1(\eta_{hj} = 1), \gamma + \sum_{j=1}^{p} 1(\eta_{hj} = 0)\bigg). \qquad (2.8)$$

3. The full conditional of V_h, $h = 1, \ldots, k^*$ only requires the updated information on latent class allocation for all subjects:

$$V_h|- \sim \text{Beta}\bigg(1 + \sum_{i=1}^{n} 1(z_i = h), \alpha + \sum_{i=1}^{n} 1(z_i > h)\bigg). \qquad (2.9)$$

4. Sample z_i, $i = 1, \ldots, n$ from the multinomial full conditional with:

$$\Pr(z_i = h|-) = \frac{\nu_h \prod_{j=1}^{p} \lambda_{hy_{ij}}^{(j)}}{\sum_{l=1}^{k^*} \nu_l \prod_{j=1}^{p} \lambda_{ly_{ij}}^{(j)}}, \qquad (2.10)$$

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

5. Update α from the Gamma full conditional:

$$\alpha|- \sim \text{Gamma}\bigg(a_\alpha + k^*, b_\alpha - \sum_{h=1}^{k^*} \log(1 - V_h)\bigg). \tag{2.11}$$

These steps are simple to implement and we gain efficiency by updating the parameters in blocks. For example, instead of updating $\lambda_h^{(j)}$ one at a time, we sample $\boldsymbol{\lambda} \equiv \{\lambda_h^{(j)}, h = 1, \ldots, k^*, j = 1, \ldots, p\}$ jointly with corresponding parameters in matrix form. In all our examples, we ran the chain for $25,000$ iterations, discarding the first $10,000$ iterations as burn-in and collecting every fifth sample post burn-in to thin the chain. Mixing and convergence were satisfactory based on the examination of trace plots and the run time scaled linearly with n and p. We also carried out sensitivity analysis by multiplying and dividing the hyperparamaters a_α, b_α and γ in (2.6) by a factor of 2, with the conclusions remained unchanged from the default setting $a_\alpha = b_\alpha = 1$ and $\gamma = 0.2\ p$.

2.4 Simulation Studies

2.4.1 Estimating sparse interactions

We first conduct a replicated simulation study to assess the estimation of sparse interactions using the proposed sp-PARAFAC model. We simulated 100 dependent binary variables $y_{ij} \in \{0, 1\}, j = 1, \ldots, p = 100$ ($d_j = d = 2$) for $i = 1, \ldots, n = 100$ subjects from a log-linear model having up to three-way interactions:

$$\log\left(\frac{\pi_{c_1 \ldots c_p}}{\pi_{0 \ldots 0}}\right) = \sum_{s=1}^{3} \sum_{S \subset \{1, \ldots, p\} : |S| = s} \beta_S 1_{(c_S = 1)}. \tag{2.12}$$

For example, if $S = \{1, 2, 4\}$, then $\beta_S = \beta_{1,2,4}$ and $1_{(c_S=1)} = 1_{(c_1=1, c_2=1, c_4=1)}$ with $1_{(\cdot)}$ denoting the indicator function. To mimic the situation where only a few interactions are present, we restrict to $S \subset S^* = \{2, 4, 12, 14\}$ and set all interactions except

$$\boldsymbol{\beta} = (\beta_2, \beta_4, \beta_{12}, \beta_{14}, \beta_{2,4}, \beta_{2,12}, \beta_{4,12}, \beta_{4,14}, \beta_{12,14}, \beta_{2,4,12}, \beta_{4,12,14})^{\mathrm{T}}$$

to zero. This data generating mechanism induces dependence among the variables in S^*, while rendering the other variables to be marginally independent. Figure 2.2 reports the posterior means and 95% credible intervals for all main effects and interactions for the variables in S^* averaged across 100 simulation replicates along with the true coefficients. As illustrated in Figure 2.2, averaging across the simulation replicates and different parameters, the 95% credible intervals cover the true parameter values 80% of the time.

Next, we study performance in estimating the dependence structure. Cramer's V is a popular statistic measuring the strength of association or dependence between two (nominal) categorical variables in a contingency table, ranging from 0 (no association) to 1 (perfect association). Let $\rho_{jj'}$ denote the Cramer's V statistics for variables j and j', so that

$$\rho^2_{jj'} = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{(\pi^{(jj')}_{c_j c_{j'}} - \pi^{(j)}_{c_j} \pi^{(j')}_{c_{j'}})^2}{\pi^{(j)}_{c_j} \pi^{(j')}_{c_{(j')}}}, \tag{2.13}$$

where $\pi^{(jj')}_{ll'} = \Pr(y_{ij} = l, y_{ij'} = l')$ and $\pi^{(j)}_{l} = \Pr(y_{ij} = l)$. Under the log-linear model (2.12), $\rho = (\rho_{jj'})$ is a sparse matrix with the Cramer's V for all pairs except those in $S^* \times S^*$ being zero. This is an immediate consequence of the fact that if $(j, j') \notin S^* \times S^*$, then y_{ij} and $y_{ij'}$ are independent.

We compare estimation of the off-diagonal entries of ρ under the sp-PARAFAC model with the empirical Cramer's V matrix $\hat{\rho}$. We can clearly convert posterior samples for the model parameters to posterior samples for $\rho_{jj'}$ through (2.13). The empirical estimator is obtained by replacing $\pi^{(jj')}_{c_j c_{j'}}$ and $\pi^{(j)}_{c_j}$ by their empirical estimators. The left panel in Figure 2.3 shows the posterior summaries (averaged across simulation replicates) of the Cramer's V values for all possible dependent pairs along with the true Cramer's V values (which can be calculated from (2.12)). In the right panel of Figure 2.3, we overlay kernel density estimators of posterior samples (in grey) and the empirical estimators (in red) of the Cramer's V values for all null pairs across all simulation replicates. Note the axes are also marked in grey and red for the respective cases. The sp-PARAFAC method clearly outperforms the empirical estimator convincingly, with the posterior density for the null pairs highly concentrated near zero while the empirical estimator has a mean Cramer's V value of 0.08 across the null pairs.

Furthermore, we can obtain power and type I error rates for the non-null and null variables respectively by computing the percentage of detected significance over the simulation replicates, with a coefficient declared significant if the 95% credible interval doesn't contain zero. Focusing on the power and type I error of the main effects and interactions in S^*, most of the error rates are appealing barring a few cases (see Table 2.2 and Table 2.3). It is not surprising that the approach may face difficulty assessing the exact interaction structure among a set of associated variables based on limited data. Further, given the Cramer's V results in the right panel of Figure 2.3, the type I error for any variable not in S^* should be very small or zero. As an example, we tested the main effects and all possible interactions for positions 20, 30, 40 and 50. The type I error rates are zero for all of them.

2.4.2 Comparison with standard PARAFAC

We now conduct a simulation study to compare estimation of the Cramer's V matrix ρ under the proposed approach to the standard PARAFAC model in (2.4). We considered 100 simulation replicates, with data in each replicate consisting of $p = 100$ categorical variables for $n = 100$ subjects, with each variable having 4 possible levels ($d_j = d = 4$). Two simulation settings were considered to induce dependence between the variables in $S^* = \{2, 4, 12, 14\}$: (i) via multiple subpopulations as in the simulation study in Dunson et al. (2008), and (ii) via a nominal GLM model $Pr(y_{ij} = c) = \frac{exp(\boldsymbol{y}_{i(j)}\boldsymbol{\beta}_c)}{1+\sum_{c=2}^{4} exp(\boldsymbol{y}_{i(j)}\boldsymbol{\beta}_c)}$ for $j \in S^*$, where $\boldsymbol{y}_{i(j)}\boldsymbol{\beta}_c$ is a linear combination of all variables that are associated with the j^{th} variable excluding the j^{th} variable. The remaining variables were independently generated from a discrete uniform distribution.

The color plot on the left in Figure 2.4 shows the true pairwise Cramer's V values under simulation setting (i) (only the top-left 20×20 sub matrix of ρ is shown for clarity). Figure 2.4 (right) and Figure 2.5 represent one of the replicates, in which the right plot in Figure 2.4 shows the Cramer's V under the standard non-sparse PARAFAC method, while Figure 2.5 shows the Cramer's V using our method with the two different choices (i) and (ii) of the baseline components. It is obvious that our approach has much better estimates for not only the true dependent pairs but also the true nulls. Results for simulation (ii) shown in Figure 2.6 again show superiority of our sparse improvement to PARAFAC.

2.5 Application

2.5.1 Splice-junction Gene Sequences

We applied the method to the Splice-junction Gene Sequences, abbreviated as splice data below. The dataset is publicly available at the UCI machine learning repository.

Splice junctions are points on a DNA sequence at which 'superfluous' DNA is removed during the process of protein creation in higher organisms. These data consist of A, C, G, T nucleotides at $p = 60$ positions for $N = 3,175$ sequences. Since the sample size is much larger than the number of variables, we compared our approach with the standard PARAFAC in two scenarios, first a small randomly selected subset (of size $n = 2p = 120$) of the full data set, and second, the full data set itself. Using two different sample sizes in this manner allows for a study of the new and existing methods and a comparison to a gold standard (a sufficiently large data set). We ran the analysis to estimate the pairwise positional dependence structure under the standard PARAFAC method and the proposed approach with discrete uniform baseline component. As is apparent in Figure 2.8, both methods have similar performance when $n \gg p$. However, when the sample size is modest compared to the dimensionality, Figure 2.7 clearly demonstrates the advantage of our proposed method in identifying the dependence structure and pushing the independent pairs to zero, thereby obtaining a closer approximation to the gold standard (Figure 2.8).

2.5.2 The Public Use Microdata Sample (PUMS)

The PUMS data contains a sample of actual responses to the American Community Survey. The dataset includes behavioral, sociodemographic and sociological variables in which 44 categorical variables are derived from the original survey data. There are 38,549 valid subjects without missing values. We used a similar strategy to that used for the splice data to compare the performance with the standard PARAFAC method under a small sample case and a full sample case. 100 subjects were first randomly selected to determine the association among the 44 social variables. Empirical marginal probabilities with a Dirichlet(1,...,1) prior were used in our model, because we believe that the underlying independent variables are not following the discrete uniform distribution and we need to avoid the zero count problem in some categories. Comparing Figure 2.10 with Figure 2.9, the sp-FARAFAC again proves its advantage in detecting more true signals and shrinking the noise.

2.5.3 National Birth Defects Prevention Study

The National Birth Defects Prevention Study is a national case-control study with over 35,000 participants to date, making it the largest study of its kind ever conducted. There are 9 states currently participating in this study: Arkansas, California, Georgia, Iowa, Massachusetts, New York, North Carolina, Texas, and Utah. The study population area covers roughly 10% of all births in the United States. The subjects are comparable to that of the general U.S. population with respect to maternal age, race, ethnicity, and education level. We employ our SP Bayesian methods to investigate (1) the association between 37 different types of heart defects and 80 potentially important covariates, and (2) the association between cleft lip/palate defects and the same factors.

Before conducting association analysis, examining the correlations within the 80 predictors is useful. We use Cramer's V statistic to quantify the associations. The significant pairs are selected if $\Pr(\text{Cramer's V} > 0.05|-) > 0.95$. The upper panel of Figure 2.11 identifies the strong associations among all solvents, the significant relationships between fertility procedures/medications, and tendency for partners to be of the same race/ethnicity.

We then determine the associations within defects outcomes and outcome-predictor associations using odds ratios (OR) as the reported measure of association. Our Bayesian procedure selects significant dependent pairs by choosing the ones with the 2.5% percentile of the Gibbs samples greater than 1 or the 97.5% percentile smaller than 1. The significantly associated pairs within 37 heart defects shown in the bottom panel of Figure 2.11 suggest that several heart defects were strongly related to each other; specifically, left ventricular outflow tract obstruction has strong positive associations with isolated coarctation of the aorta, aortic stenosis, hypoplastic left heart syndrome, and coarctation with ventricular septal defects.

The upper panel of Figure 2.12 shows relationships between heart defects and covariates. Double outlet right ventricle and pulmonary atresia are both affected by solvents of all types (benzene, toluene, and xylene, carbon tetrachloride, chloroform, methylene chloride, perchloroethylene, trichloroethane, trichloroethylene, and stoddard) with odds ratios all around 3, while double outlet right ventricle is also associated with gestational diabetes (OR: around 2). However, only two solvents (benzene and carbon tetrachloride) have an impact on conoventricular ventricular septal defects with odds ratios around 2.2. Moreover, left ventricular outflow defects, hypoplastic left heart syndrome, coarctation of the aorta, and aortic stenosis are associated with the pharmaceuticals sulfamethoxazole, trimethoprim, and thyroid/antithyroid drugs with moderate odds ratios around 1.8. Cleft palate is positively related to the use of fertility medications/procedures and whether the mother had surgery to restore fertility (Figure 2.12 bottom plot). The corresponding odds ratios are around 1.5.

2.6 Discussion

We have proposed a sparse modification to the widely-used PARAFAC tensor factorization, and have applied this in a Bayesian context to improve analyses of ultra sparse huge contingency tables. Given the compelling success in this application area, we hope that the proposed notion of sparsity will have a major impact in other areas, including tensor completion problems in machine learning. There is an enormous literature on low rank and sparse matrix factorizations, and the sp-PARAFAC should facilitate scaling of such approaches to many-way tables while dealing with the inevitable curse of dimensionality. Although we take a Bayesian approach, we suspect that frequentist penalized optimization methods can also exploit our same concept of sparsity in learning a compressed characterization of a huge array based on limited data.

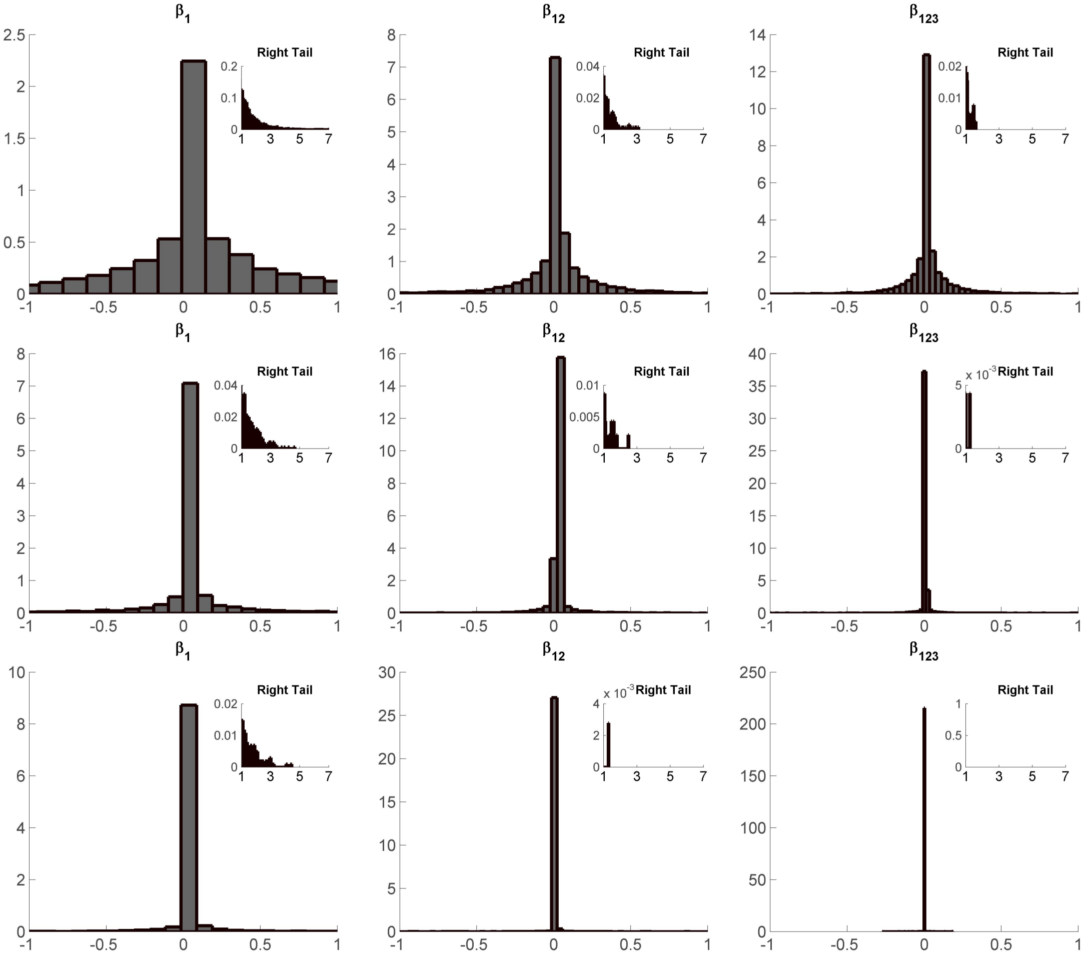

Figure 2.1: Histograms of induced priors for one main effect β_1, one two-way interaction β_{12}, and the three-way interaction β_{123} - Top Row: $\gamma = 1$; Middle Row: $\gamma = 5$; Bottom Row: $\gamma = 20$.

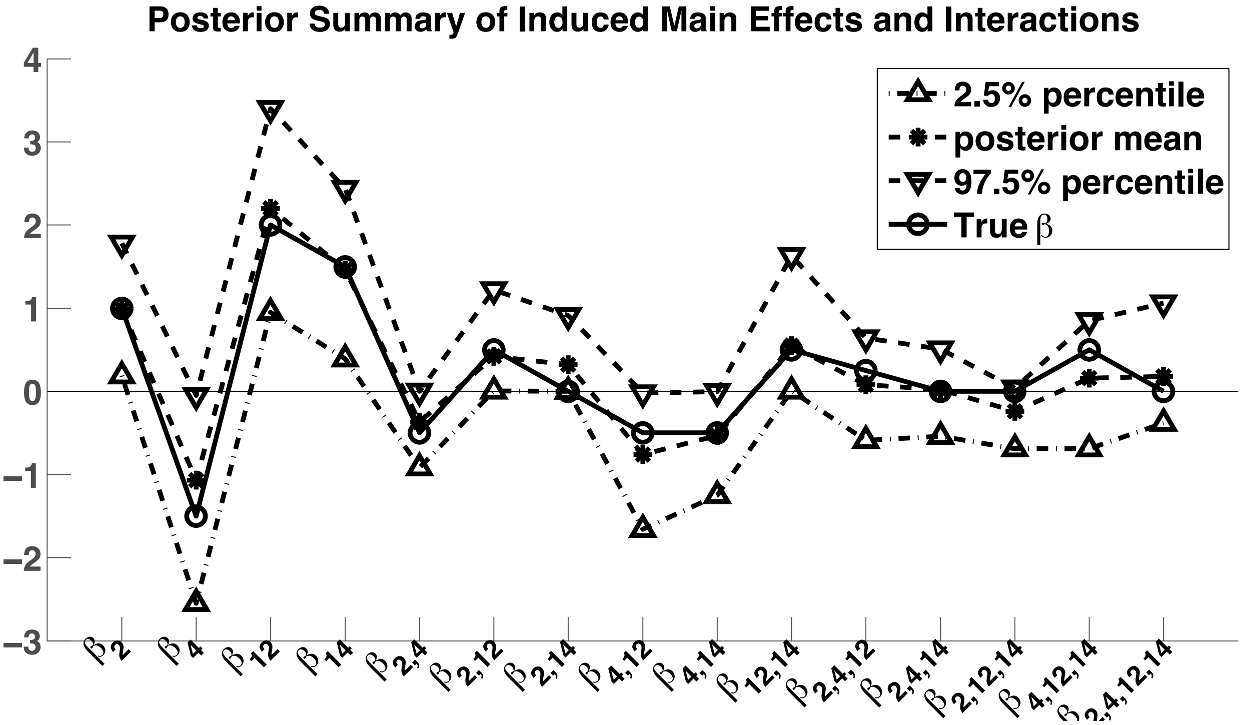

Figure 2.2: Posterior means and 95% credible intervals for all main effects and interactions in S^* compared with the true coefficients.

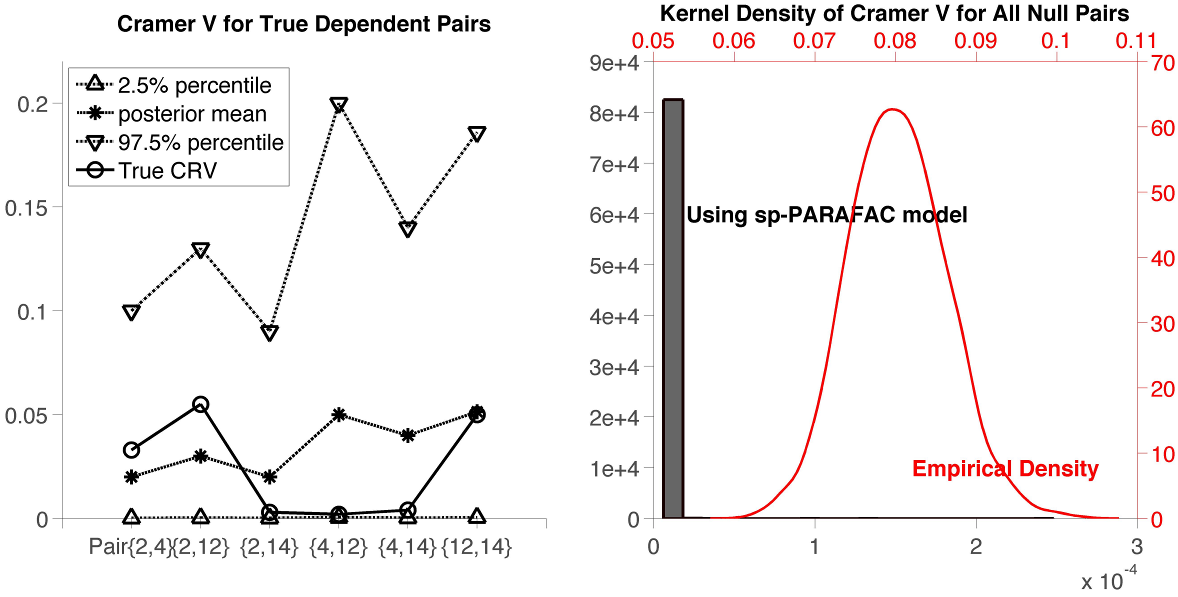

Figure 2.3: Left: Posterior summaries of the Cramer's V values for all dependent pairs vs. the true Cramer's V values; Right: Estimated density of Cramer's V combining all null pairs under sp-PARAFAC vs. empirical estimation.

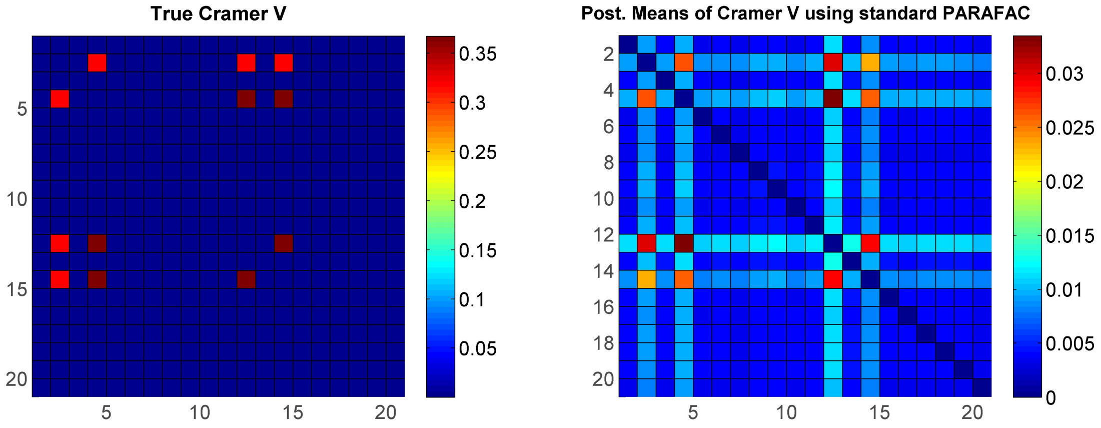

Figure 2.4: Simulation setting (i) – Left: True Cramer's V matrix; Right: Posterior means of Cramer's V using standard PARAFAC. Top 20×20 sub-matrix shown.

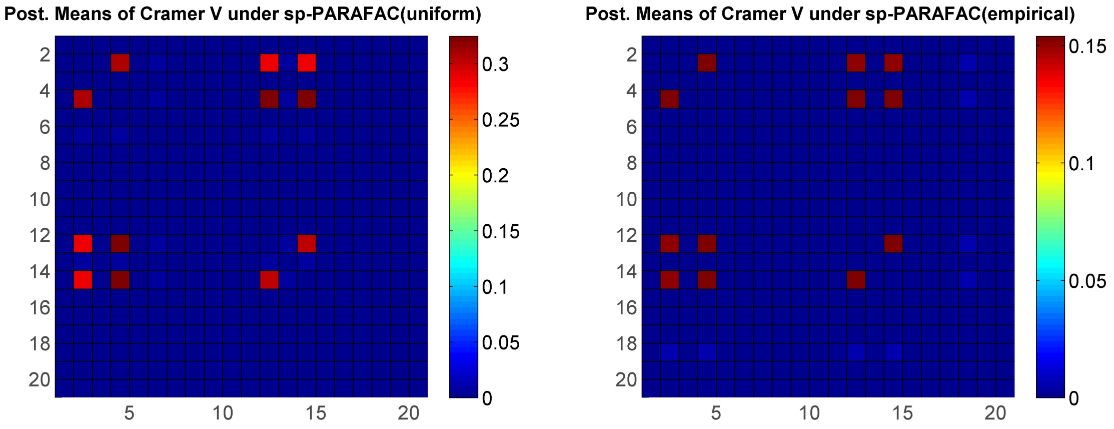

Figure 2.5: Posterior means of Cramer's V under simulation setting (i) using proposed method – Left: with $\lambda_0^{(j)}$ being discrete uniform; Right: with $\lambda_0^{(j)}$ being empirical estimates of the marginal category probabilities. Top 20 × 20 sub-matrix shown.

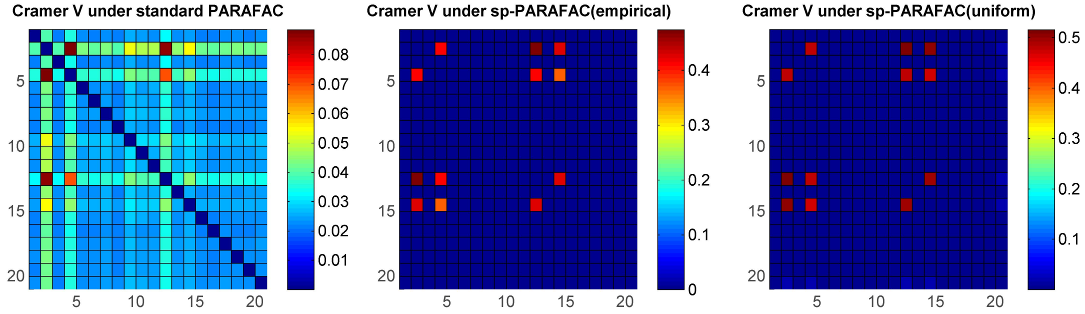

Figure 2.6: Posterior means of Cramer's V under simulation setting (ii) – Left: using standard PARAFAC; Middle: under proposed method using empirical marginal with Diri(1,...,1) prior for λ_0; Right: using proposed method with discrete uniform λ_0. Top 20×20 sub-matrix shown.

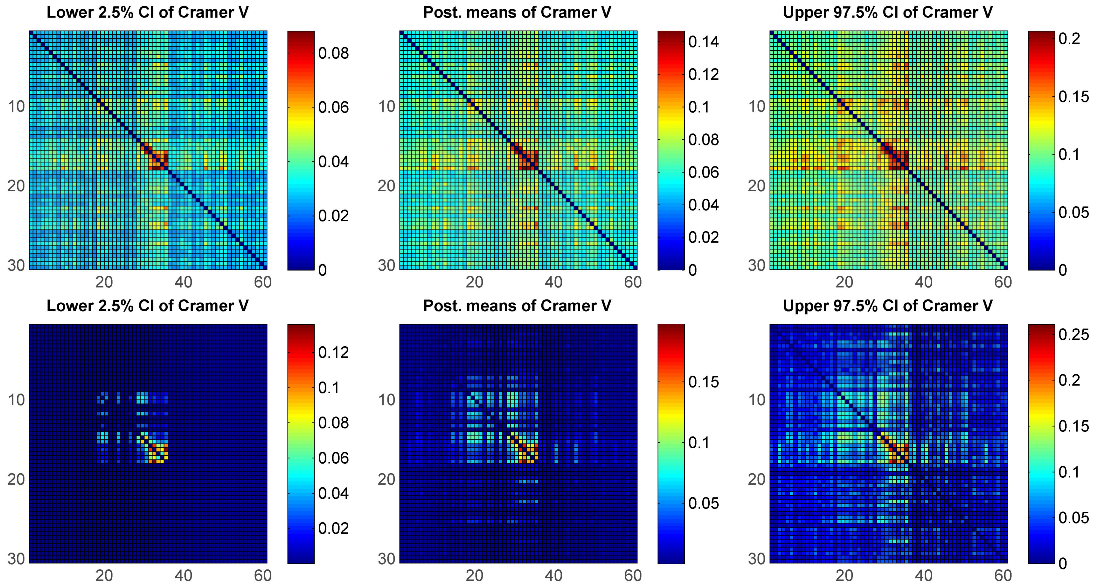

Figure 2.7: Posterior quantiles of Cramer's V with 120 sequences of splice data – Upper panel: under standard PARAFAC; Bottom panel:under proposed method.

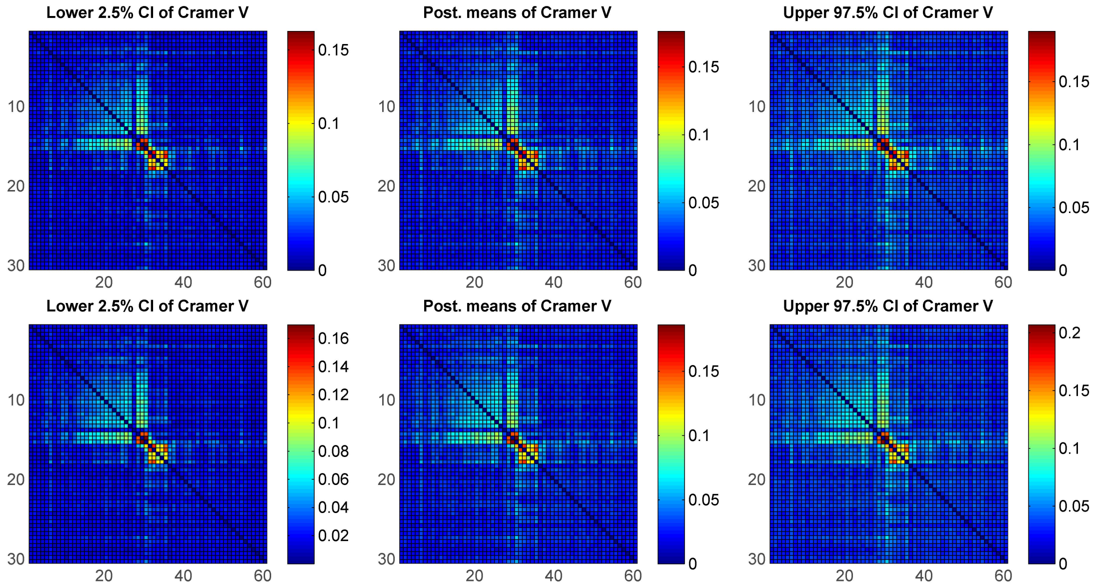

Figure 2.8: Posterior quantiles of Cramer's V with 3,175 sequences of splice data – Upper panel: under standard PARAFAC; Bottom panel:under proposed method.

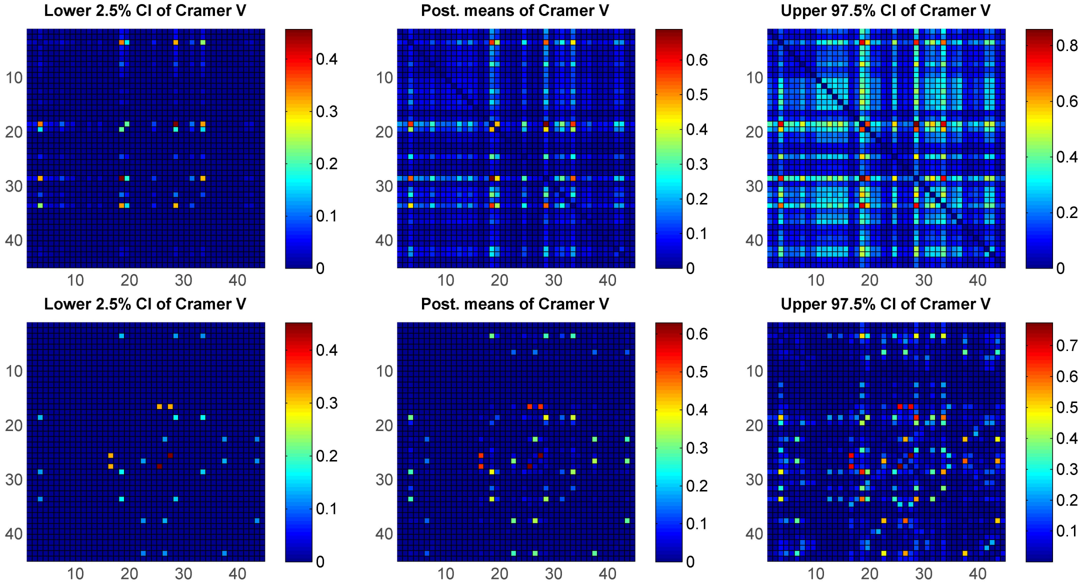

Figure 2.9: Posterior quantiles of Cramer's V with 100 subjects of PUMS – Upper panel: under standard PARAFAC; Bottom panel: under proposed method.

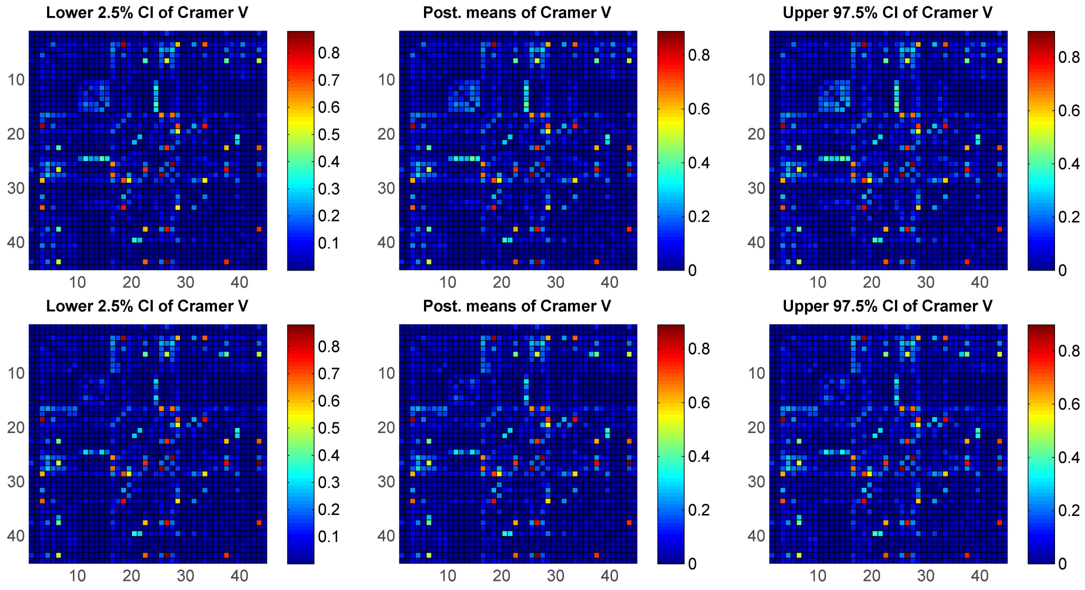

Figure 2.10: Posterior quantiles of Cramer's V with 38,549 subjects of PUMS – Upper panel: under standard PARAFAC; Bottom panel:under proposed method.

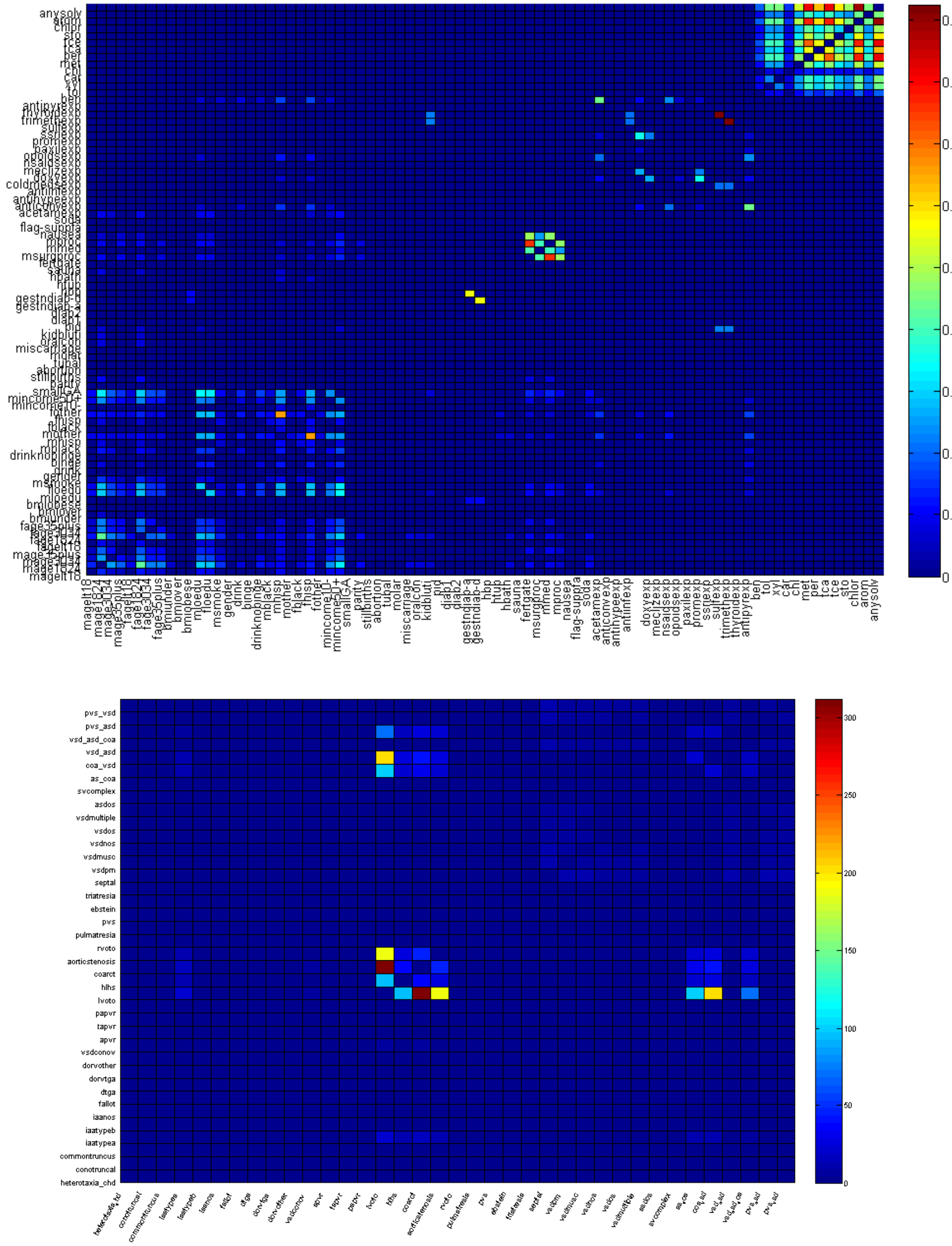

Figure 2.11: Upper panel:Posterior mean of Cramer's V for 80 potential factors; Bottom panel: Posterior mean of significant odds ratios within 37 heart related birth defects.

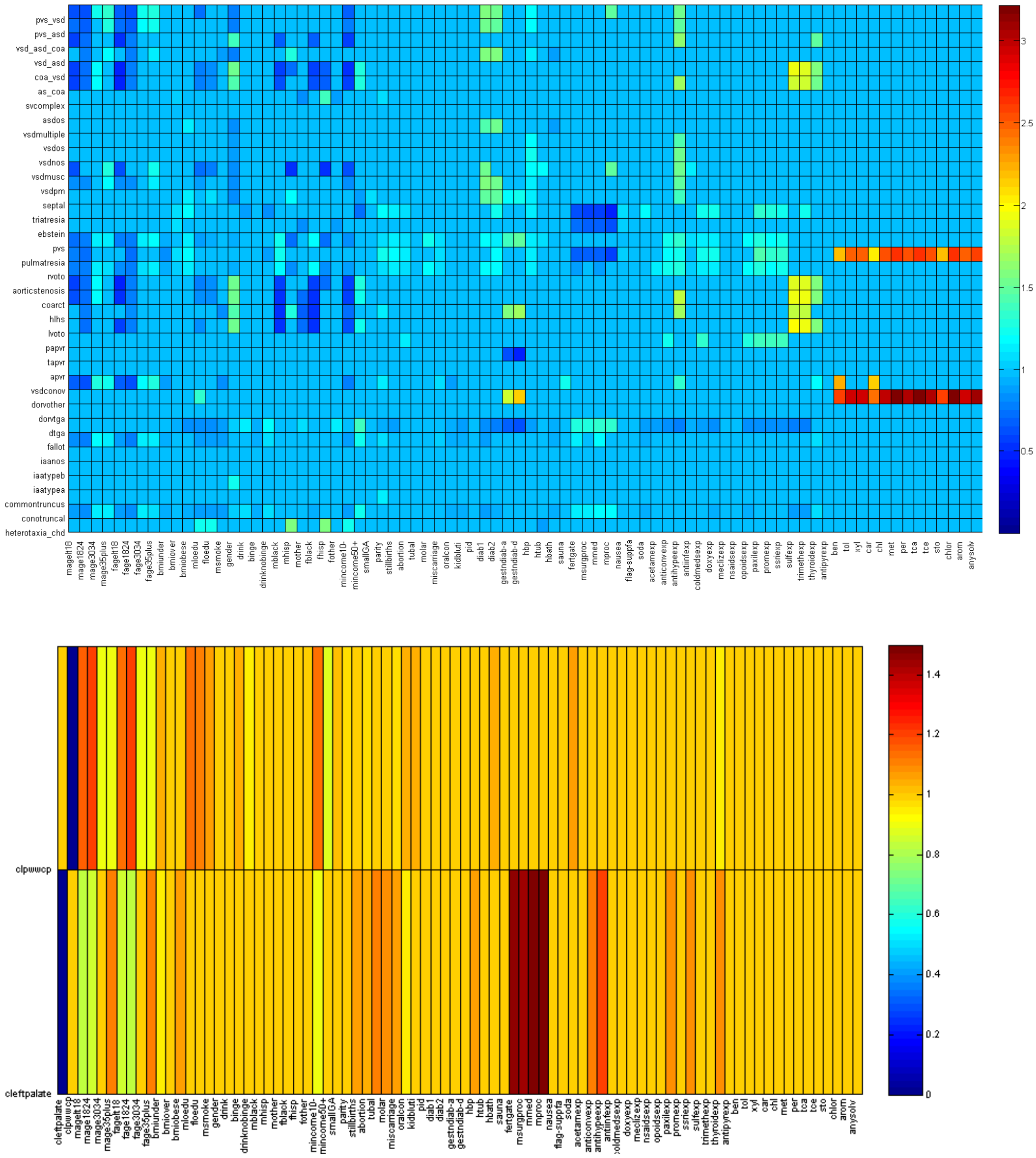

Figure 2.12: Upper panel: Posterior mean of odds ratios between 37 heart related birth defects and 80 potential factors; Bottom panel: Posterior mean of odds ratios between 2 cleft defects and 80 potential factors.

Table 2.1: Summary statistics of induced priors on coefficients in log-linear model parameterization.

γ	Coefficient	Mean	Std.dev	Min	Max	Skewness	Kurtosis
1	β_1	0.014	0.831	-6.765	6.389	0.210	9.109
1	β_{12}	-0.002	0.340	-2.895	3.105	-0.025	16.583
1	β_{123}	0.002	0.196	-2.223	2.632	0.525	24.686
5	β_1	-0.002	0.485	-5.648	5.433	0.031	27.980
5	β_{12}	0.000	0.124	-2.085	2.244	0.495	93.438
5	β_{123}	0.000	0.051	-1.214	0.745	-3.701	159.360
20	β_1	0.002	0.246	-3.109	5.669	2.474	99.554
20	β_{12}	0.000	0.042	-1.126	1.819	9.488	632.790
20	β_{123}	0.000	0.009	-0.664	0.214	-44.051	3014.000

Table 2.2: Power for Non-null Variables Based on 100 Simulations

	β_2	β_4	β_{12}	β_{14}	$\beta_{2,4}$	$\beta_{2,12}$	$\beta_{4,12}$	$\beta_{4,14}$	$\beta_{12,14}$	$\beta_{2,4,12}$	$\beta_{4,12,14}$
Power	0.97	0.9	1	1	0.95	0.99	0.98	0.97	0.99	0	0
True coefficient	1	-1.5	2	1.5	-0.5	0.5	-0.5	-0.5	0.5	0.25	0.5

Table 2.3: Type I Error for Null Variables Based on 100 Simulations

	$\beta_{2,14}$	$\beta_{2,4,14}$	$\beta_{2,12,14}$	$\beta_{2,4,12,14}$
Type I error	0.97	0	0.68	0
True coefficient	0	0	0	0

CHAPTER 3

Nonparametric Bayes Modeling for Case Control Studies

3.1 Introduction

Retrospective case-control studies are common in epidemiologic research because they are much more cost effective than prospective studies, particularly for rare diseases. However, retrospective studies only model exposure given disease, presenting some challenges in analysis and interpretation of the results. In prospective studies, logistic models are widely used to estimate adjusted odds ratios for each of multiple risk factors. A primary concern when analyzing case-control data is whether prospective inferences can be made. In the frequentist framework, there is a rich literature (Anderson 1972; Prentice and Pyke 1979) demonstrating that one can ignore the study design and use estimation and inference based on a logistic regression. That is, it has been shown that odds ratios for prospective and case control data are equivalent.

Consider the National Birth Defects Prevention Study (NBDPS), the largest case-control study ever conducted in the United States on the etiology of birth defects (Yoon et al. 2001). Data are collected on many different defects along with hundreds of potentially associated factors, including environmental, behavioral, biomedical and occupational variables. Typically these variables are categorized, leading to a huge sparse contingency table having mostly zero counts. There is strong prior reason to suspect interactions, and logistic regression is clearly not appropriate. Although there is a recent Bayesian literature on analysis of high-dimensional contingency tables (Dunson and Xing (2009), Bhattacharya and Dunson (2011), Zhou et al. (2013)), these methods view the data as multivariate categorical arising from a prospective design. Our focus is on addressing the question of whether we can adapt these approaches to case control settings.

There is a rich literature on Bayesian analysis of case-control data in low-dimensional settings. Zelen and Parker (1986); Nurminen and Mutanen (1987); Marshall (1988) and Ashby et al. (1993) all consider identical Bayesian formulations of a case-control model with a binary exposure X. Let ϕ and γ be the probabilities of exposure in control and case populations respectively. The retrospective likelihood is

$$l(\phi, \gamma) \propto \phi^{n_{01}}(1-\phi)^{n_{00}}\gamma^{n_{11}}(1-\gamma)^{n_{10}}, \tag{3.1}$$

where n_{01} and n_{00} are the number of exposed and unexposed observations in the control population, whereas n_{11} and n_{10} denote the same for the case population. Independent conjugate prior distributions for ϕ and γ are chosen as $Beta(u_1, u_2)$ and $Beta(\nu_1, \nu_2)$ respectively. After reparametrization one obtains the posterior distribution of the log odds ratio parameter, $\beta = log\{\gamma(1-\phi)/\phi(1-\gamma)\}$ as

$$l(\beta|n_{11}, n_{10}, n_{01}, n_{00}) \propto \exp\{(n_{11}+\nu_1)\beta\} \int_0^1 \frac{\phi^{n_{11}+n_{01}+\nu_1+u_2-1}(1-\phi)^{n_{10}+n_{00}+\nu_2+u_1-1}}{\{1-\phi+\phi\exp(\beta)\}^{n_{11}+n_{10}+\nu_1+\nu_2}} d\phi \tag{3.2}$$

The above references used different methods to approximate the posterior distribution of β shown in (3.2) as well as discussing different prior elicitations based on historical studies.

An alternative is to induce a retrospective likelihood by starting with a model for the prospective likelihood and using Bayes rule. For each subject i, let d_i be a binary response observed together with covariates X_i. Assume a binary response logistic regression for the conditional likelihood of d_i given covariates, with β the coefficients, and let θ denote parameters in a model for the marginal distribution of X_i. Assuming X_i is continuous, Müller and Roeder (1997) proposed a semiparametric Bayes approach. They factorize the joint posterior as

$$\Pr\big(\beta, \theta | \mathbf{X}, \mathbf{D}\big) \propto \Pr(\beta, \theta) \prod_{i=1}^{n} \Pr(X_i | d_i, \beta, \theta), \tag{3.3}$$

where under conditional independence assumptions they let,

$$\Pr(X_i | d_i, \beta, \theta) = \frac{\Pr(d_i | X_i, \beta) \Pr(X_i | \theta)}{\Pr(d_i | \beta, \theta)}. \tag{3.4}$$

Problems arise in approximating the denominator in (3.4), as this involves an analytically intractable high-dimensional integral.

Seaman and Richardson (2001) extended these two types of models by allowing more than one categorical exposure variable and employing Markov chain Monte Carlo methods to sample the posterior of β. Müller et al. (1999) modeled the retrospective likelihood directly for continuous exposures, also allowing binary covariates via a probit model. Ghosh and Chen (2002); Sinha et al. (2004; 2005) developed general Bayesian methods for matched case-control studies in the presence of one or more exposure variables, missing exposures, and multiple disease states. None of the above approaches can accommodate more than a modest number of categorical predictors. As the number of covariates increases, the algorithms either fail to implement or have highly biased estimates.

There has also been research establishing the equivalence of prospective and retrospective Bayesian models. Seaman and Richardson (2004) obtained equivalence through carefully chosen priors. Staicu (2010) extended the class of priors, while still relying on logistic regression. As motivated above, logistic models are too inflexible for our motivating application. Byrne and Dawid (2013) established an equivalence of learning odds ratios whether using retrospective or prospective likelihood and Bayesian approach. However this equivalence only holds with particular conditions satisfied for the models and priors. Unfortunately their method is impractical for large number of covariates in Bayes analysis.

With this motivation, we develop a nonparametric Bayes method based on directly modeling the retrospective likelihood building on existing methods for high-dimensional categorical data. The basic framework is proposed in Section 3.2. Section 3.3 outlines a Gibbs sampler for posterior computation. Section 3.4 compares performance with competitors in a simulation study. Section 3.5 analyzes data from the motivating birth defect study, and Section 3.6 contains a discussion.

3.2 Conditional Sparse Parallel Factor Analysis Model

3.2.1 Model and prior

The general form of the retrospective likelihood is:

$$l(\theta_1, \theta_0) = \prod_{i:d_i=1} \Pr(x_i|d_i = 1, \theta_1) \prod_{i:d_i=0} \Pr(x_i|d_i = 0, \theta_0), \tag{3.5}$$

where $\Pr(x_i|d_i = d, \theta_d)$ is the conditional likelihood of the high-dimensional categorical predictors $x_i = (x_{i1}, \ldots, x_{ip})'$, with $x_{ij} \in \{1, \ldots, d_j\}$ for $j = 1, \ldots, p$, given disease status d ($0 =$ control, 1=case). When p is moderate to large (say in the dozens to 100s or more), problems arise in defining a *flexible* model for these high-dimensional categorical predictors. Potentially log-linear models can be used, but unless the vast majority of the interactions are discarded *a priori*, one obtains an unmanageably enormous number of terms to estimate, store and process. These bottlenecks are freed by the use of Bayesian low rank tensor factorizations, which have had promising performance in practice (Dunson and Xing (2009); Bhattacharya and Dunson (2011); Kunihama and Dunson (2013); Zhou et al. (2013)). Johndrow, Bhattacharya and Dunson (2014) recently showed that sparse log-linear models have low rank tensor factorizations, providing support for the use of tensor factorizations as a computationally convenient alternative.

We build upon the sparse parallel factor analysis (SPFA) method of Zhou et al. (2013), motivated by their strong theory and exceptional practical performance. Conditional on disease status, the SPFA factorization of the joint p.m.f. of x_i can be expressed as

$$\Pr(x_{i1} = c_1, \ldots, x_{ip} = c_p | d_i = d) = \sum_{h=1}^{k} \nu_{dh} \prod_{j=1}^{p} \lambda_{dhc_j}^{(j)}, \tag{3.6}$$

with sparsity assumptions:

$$\lambda_{dhc_j}^{(j)} = \begin{cases} \lambda_{dhc_j}^{(j)} = \Pr(x_{ij} = c_j | d_i = d, z_i = h), & \text{if } j \in S_{dh} \\ \lambda_{0c_j}^{(j)} = \Pr(x_{ij} = c_j), & \text{if } j \in S_{dh}^c \end{cases}, \tag{3.7}$$

where in (3.6), $\nu_{dh} = \Pr(z_i = h | d_i = d)$ is a mixture probability for latent class variable $z_i \in \{1, \ldots, k\}$ under disease d, and $\sum_{h=1}^{k} \nu_{dh} = 1$. $\boldsymbol{\lambda}_{dh}^{(j)} = (\lambda_{dh1}^{(j)}, \ldots, \lambda_{dhd_j}^{(j)})$ is a vector of the multinomial probabilities of $x_{ij} = 1, \ldots, d_j$ given disease d and latent class component h. This model, ignoring the sparsity assumptions for the moment, is motivated by latent structure analysis (Lazarsfeld and Henry 1968) which provides meaningful interpretation. Suppose we have two categorical covariates for example, given disease outcome d, model (3.6) becomes

$$\begin{aligned}\Pr(x_{i1} = c_1, x_{i2} = c_2 | d_i = d) &= \sum_{h=1}^{k} \nu_{dh} \lambda_{dhc_1}^{(1)} \lambda_{dhc_2}^{(2)} \qquad (3.8)\\ &= \sum_{h=1}^{k} \Pr(z_i = h | d_i = d) \prod_{j=1}^{2} \Pr(x_{ij} = c_j | z_i = h, d_i = d).\end{aligned}$$

With the introduction of the latent class z_i for all subjects in outcome group d, any covariates x_{i1} and x_{i2} that are possibly dependent can be assumed conditionally independent. But marginalizing out the latent index z_i produces a mixture of product multinomial distributions for x_i and hence leads to a possible dependence structure within x_i in outcome group d. Any joint probability of $x_i = (x_{i1}, x_{i2})'$ for all subjects in each group d can always be decomposed as in (3.8) for some sufficiently big k (Dunson and Xing 2009). The extension to the multivariate covariates case is straightforward. A nonparametric Bayes approach can be used to deal with uncertainty in k.

The effective number of parameters can be massively reduced by choosing a prior that favors independence between many of the predictors. This can be instantiated via the sparsity assumption in (3.7).In particular, in each disease group d and component h, we partition the p dimensions of covariates into two mutually exclusive subsets $S_{dh} \cup S_{dh}^c = \{1, \ldots, p\}$, and for the variables within subset S_{dh}^c, we allocate $\lambda_{dhc_j}^{(j)}$ to its baseline category $\lambda_{0c_j}^{(j)}$, which is not dependent on the latent class or the outcome group. A Bayes approach is used to learn the allocation of the subsets for each variable. This dramatically reduces the number of parameters needed to learn the distribution of x_i by sharing parameters between disease group and latent class levels for a large number of variables.

Consider a simple case of three covariates. If we let $\lambda_{dhc_3}^{(3)} = \lambda_{0c_3}^{(3)}$ for $h = 1, \ldots, k$

and $d = 0, 1$, we have

$$\begin{aligned}\Pr(x_{i1} = c_1, x_{i2} = c_2, x_{i3} = c_3 | d_i = d) &= \lambda^{(3)}_{0c_3} \sum_{h=1}^{k} \nu_{dh} \lambda^{(1)}_{dhc_1} \lambda^{(2)}_{dhc_2} \\ &= \Pr(x_{i3} = c_3) \cdot \Pr(x_{i1} = c_1, x_{i2} = c_2 | d_i = d),\end{aligned}$$

implying the third covariate is independent of the outcome and does not have any interaction with the other two variables. However, the sparsity assumption has the flexibility in allowing $j \in S^c_{dh}$ for some but not all $h \in \{1, \ldots, k\}$, which leads to some interactions/collinearity between the j^{th} factor and the other factors. This implicitly indicates the j^{th} covariate can be associated with the disease through the other factors correlated with the disease. Moreover, if a variable j is independent of the other covariates, a marginal association between the j^{th} variable and the outcome can be introduced by having $j \in S^c_{dh}$ for all h but not for all d . In practice, the cardinality of S_{dh} (denoted as $|S_{dh}|$) is unknown but can be estimated by a Bayesian approach which will be discussed later. $\boldsymbol{\lambda}^{(j)}_0 = \{\lambda^{(j)}_{01}, \ldots, \lambda^{(j)}_{0d_j}\}$ vectors are *fixed in advance*; one natural choice is: $\boldsymbol{\lambda}^{(j)}_0 = \left(\frac{1}{d_j}, \ldots, \frac{1}{d_j}\right)'$ corresponding to a discrete uniform. Furthermore, our model also allows subjects in different outcome groups to have a different mixture probability to a specific class h (i.e. ν_{dh}), which results in a more flexible distribution structure for x_i for each outcome group.

Compared with Zhou et al. (2013) whose aim is to model the joint distribution of outcomes and predictors, we are now modeling the retrospective likelihood by conditioning on disease and having two different groups. In order to study the dependence between outcomes and covariates, it would be extremely inefficient to estimate the high-dimensional distribution of the covariates completely separately in the two groups, which would be effectively acting as if all the predictors are important. It would also be inappropriate to pool the two groups, as that would assume there was no impact of the covariates on disease implicitly. Hence, our primary modeling contribution is in allowing uncertainty in what attributes are similar between the groups; in particular, we would like to adaptively learn which parameters are common and which are different. This adaptive learning is key to inferring the prospective impact of the predictors on disease risk.

Our proposed model (3.6) with assumptions (3.7) can be expressed in a hierarchical form with priors specified for the unknown parameter vectors: for $d = 0$ or 1,

$$\begin{gathered} x_{ij} | d_i = d, z_i = h \sim \text{Mult}\Big(\{1, \ldots, d_j\}; \lambda_{dh1}^{(j)}, \ldots, \lambda_{dhd_j}^{(j)}\Big), \\ \boldsymbol{\lambda}_{dh}^{(j)} \equiv \Big(\lambda_{dh1}^{(j)}, \ldots, \lambda_{dhd_j}^{(j)}\Big) \sim (1 - \tau_{dh})\delta_{\boldsymbol{\lambda}_0^{(j)}} + \tau_{dh}\text{Diri}(a_{j1}, \ldots, a_{jd_j}), \\ \Pr(z_i = h | d_i = d) = \nu_{dh} = V_{dh} \textstyle\prod_{l<h}(1 - V_{dl}), \\ V_{dh} \sim \text{Beta}(1, \alpha), \quad \alpha \sim \text{Gamma}(a_\alpha, b_\alpha), \quad \tau_{dh} \sim \text{Beta}(1, \gamma). \end{gathered} \tag{3.9}$$

Expression (3.9) is equivalent to letting the subset-size $|S_{dh}| \sim \text{Binom}(p, \tau_{dh})$ and drawing a random subset S_{dh} uniformly from all subsets of $\{1, \ldots, p\}$ of size $|S_{dh}|$ in (3.7). A stick-breaking representation of the Dirichlet process prior (Sethuraman 1994) is chosen for the component weights $\{\nu_{dh}, h = 1, \ldots, k\}$ allowing $k = \infty$, with a hyperprior placed on the concentration parameter α in the stick-breaking process to allow the data to inform more strongly about the component weights. The probability of allocation τ_{dh} to the *active* (non-baseline) category is chosen as beta$(1, \gamma)$, with $\gamma > 1$ favoring allocation of many of the $\boldsymbol{\lambda}_{dh}^{(j)}$s to the baseline category $\boldsymbol{\lambda}_0^{(j)}$ in both outcome groups.

3.2.2 Inference

It is common in practice to learn the prospective odds ratio in case-control studies. To proceed, the odds for any X_A versus X_B given the disease is:

$$odds(X_A\ vs.\ X_B|d) = \frac{\Pr(X_A|d)}{\Pr(X_B|d)}. \tag{3.10}$$

The corresponding retrospective odds ratio can be computed in a straightforward fashion by our proposed model:

$$OR_{retro} = \frac{odds(X_A\ vs.\ X_B|d=1)}{odds(X_A\ vs.\ X_B|d=0)}. \tag{3.11}$$

It is well known that prospective odds ratio is equivalent to retrospective odds ratio which is obvious from Bayes' theorem. That is,

$$OR_{prosp} = \frac{odds(d=1|X_A)}{odds(d=1|X_B)} = \frac{odds(X_A\ vs.\ X_B|d=1)}{odds(X_A\ vs.\ X_B|d=0)} = OR_{retro}. \tag{3.12}$$

As a consequence, it is valid and without difficulty to report prospective odds ratios from case-control studies. The marginal odds ratio for the j^{th} predictor can be easily computed by setting $X_A = \{x_{ij} = 1\}$ and $X_B = \{x_{ij} = 0\}$.

3.3 Posterior Computation

Under model (3.9), we can easily proceed to draw posterior samples from a Gibbs sampler since all the full conditionals have recognizable forms. The algorithm iterates through the following steps:

1. In each disease group d, for variable $j = 1, \dots, p$ and latent class $h = 1, \dots, k_d^*$,

where $k^*_d = \max\{z_1, \ldots, z_{n_d}\}$, update $\boldsymbol{\lambda}^{(j)}_{dh} \equiv (\lambda^{(j)}_{dh1}, \ldots, \lambda^{(j)}_{dhd_j})$ from a two component mixture distribution, having a point mass at the baseline probability vector:

$$\begin{aligned}(\boldsymbol{\lambda}^{(j)}_{dh}|-) &= \frac{w^{(j)}_{0h}}{w^{(j)}_{0h}+w^{(j)}_{1h}}\delta_{\boldsymbol{\lambda}^{(j)}_0} + \frac{w^{(j)}_{1h}}{w^{(j)}_{0h}+w^{(j)}_{1h}}\text{Diri}\Big(a_{j1}+\sum_{i=1}^{n}1(x_{ij}=1, z_i=h, d_i=d), \\ &\qquad \ldots, a_{jd_j}+\sum_{i=1}^{n}1(x_{ij}=d_j, z_i=h, d_i=d)\Big), \end{aligned} \tag{3.13}$$

where $w^{(j)}_{0h}$ and $w^{(j)}_{1h}$ are proportional to the mixture weights:

$$\begin{aligned} w^{(j)}_{0h} &= (1-\tau_{dh})\prod_{c=1}^{d_j}\lambda^{(j)\,\sum_{i=1}^{n}1(z_i=h,d_i=d,x_{ij}=c)}_{0c}, \\ w^{(j)}_{1h} &= \tau_{dh}\frac{\Gamma(\sum_{c=1}^{d_j}a_{jc})}{\prod_{c=1}^{d_j}\Gamma(a_{jc})}\cdot\frac{\prod_{c=1}^{d_j}\Gamma\big(a_{jc}+\sum_{i=1}^{n}1(z_i=h,d_i=d,x_{ij}=c)\big)}{\Gamma\big(\sum_{c=1}^{d_j}a_{jc}+\sum_{i=1}^{n}1(z_i=h,d_i=d)\big)}. \end{aligned}$$

2. Let $S^j_{dh} \in \{0,1\}$ be a binary allocation variable indicating the component $\lambda^{(j)}_{dh}$ is drawn from in (3.13), with $S^j_{dh}=0$ if $\lambda^{(j)}_{dh}$ is updated from the baseline component. Update τ_{dh}, $h=1,\ldots,k^*$ from a Beta full conditional:

$$\tau_{dh}|- \sim \text{Beta}\Big(1+\sum_{j=1}^{p}1(S^j_{dh}=1), \gamma+\sum_{j=1}^{p}1(S^j_{dh}=0)\Big). \tag{3.14}$$

3. The full conditional of V_{dh}, $h=1,\ldots,k^*_d$ only requires the updated information on latent class allocation for the subjects within the disease group d:

$$V_{dh}|- \sim \text{Beta}\Big(1+\sum_{i=1}^{n_d}1(z_i=h), \alpha+\sum_{i=1}^{n_d}1(z_i>h)\Big). \tag{3.15}$$

4. Sample z_i, for $\{i=1,\ldots,n \text{ s.t. } d_i=d\}$ and $d=0,1$ from the multinomial full

conditional with:

$$\Pr(z_i = h|d_i = d, -) = \frac{\nu_{dh} \prod_{j=1}^{p} \lambda_{dhx_{ij}}^{(j)}}{\sum_{l=1}^{k^*} \nu_{dl} \prod_{j=1}^{p} \lambda_{dlx_{ij}}^{(j)}}, \tag{3.16}$$

where $\nu_{dh} = V_{dh} \prod_{l<h}(1 - V_{dl})$. Note that $\boldsymbol{z} = \{z_i, i = 1, \ldots, n_0; z_l, l = 1, \ldots, n_1\}$.

5. Update α from the Gamma full conditional:

$$\alpha|- \sim \text{Gamma}\left(a_\alpha + k_0^* + k_1^*, b_\alpha - \sum_{d=0}^{1} \sum_{h=1}^{k_d^*} \log(1 - V_{dh})\right). \tag{3.17}$$

The default setting is $a_\alpha = b_\alpha = 1$.

3.4 Simulation Studies

3.4.1 Simulation from log-linear models

We first conduct a replicated simulation study mimicking a case-control design to assess the performance using the proposed model compared with logistic regression with and without the Benjamini and Hochberg correction, CART, random forest, and Lasso. For 50 case and 50 control subjects, we simulated p binary covariates $x_{ij} \in \{0, 1\}$, $j = 1, \ldots, p$, under two scenarios: (i) $p = 20$, and (ii) $p = 100$, among which four variables ($j = 2, 4, 12, 14$) were assumed dependent and generated from a saturated log-linear model with coefficients varying by outcome d:

$$\log\left(\frac{\pi_{c_2,c_4,c_{12},c_{14}}^{d}}{\pi_{0,0,0,0}^{d}}\right) = \sum_{s=1}^{4} \sum_{S^* \subset \{2,4,12,14\}:|S^*|=s} \beta_{S^*}^{d} 1_{(c_{S^*}=1)}, \tag{3.18}$$

where $\pi_{c_2,c_4,c_{12},c_{14}}^{d} = Pr(x_{i2} = c_2, x_{i4} = c_4, x_{i,12} = c_{12}, x_{i,14} = c_{14} \mid d_i = d)$. If $S^* = \{2, 4\}$, for example, then $\beta_{S^*}^{d} = \beta_{2,4}^{d}$ and $1_{(c_{S^*}=1)} = 1_{(c_2=1,c_4=1)}$ with $1_{(\cdot)}$ denoting the indicator function. Different values of $c_j, j = 2, 4, 12, 14$, will lead to different coefficients in the model. One illustration is if $c_2 = 1, c_4 = 1, c_{14} = 1$, model (3.18) becomes

$$\log\left(\frac{\pi^d_{1,1,0,1}}{\pi^d_{0,0,0,0}}\right) = \beta^d_2 + \beta^d_4 + \beta^d_{14} + \beta^d_{2,4} + \beta^d_{2,14} + \beta^d_{4,14} + \beta^d_{2,4,14}. \tag{3.19}$$

All the true coefficients are set as in Table 3.1 and 3.2. Having different main effects given disease outcome in the log-linear model results in association between the outcome and those four variables. All the remaining null variables $j \in \{1, \ldots, p\}, j \neq \{2, 4, 12, 14\}$ were independently generated from a discrete uniform distribution. This data generating mechanism induces dependence among the variables in S^* and their impact on outcome, while rendering the other variables marginally independent.

Simulations were conducted based on 1,000 data replicates for each scenario. In each replicate, the posterior marginal odds ratio for each variable j using (3.10)- (3.12) was computed according to:

$$OR^{(j)} = \frac{P(x_{ij} = 1|d_i = 1)}{P(x_{ij} = 0|d_i = 1)} \bigg/ \frac{P(x_{ij} = 1|d_i = 0)}{P(x_{ij} = 0|d_i = 0)}, \tag{3.20}$$

where

$$\Pr(x_{ij} = c_j|d_i = d) = \begin{cases} \sum_{h=1}^{k} \nu_{dh}\lambda^{(j)}_{dhc_j}, & \text{if } j \in S_{dh} \\ \sum_{h=1}^{k} \nu_{dh}\lambda^{(j)}_{0c_j} = \lambda^{(j)}_{0c_j}, & \text{if } j \in S^c_{dh} \end{cases}. \tag{3.21}$$

The corresponding credible interval of the odds ratio was used to identify whether the variable j was significant. For each data replicate, we ran the chain for $25,000$ iterations, discarding the first $10,000$ iterations as burn-in and collecting every fifth sample post burn-in to thin the chain. Mixing and convergence were satisfactory based on the examination of trace plots.

Receiver Operating Characteristic (ROC) curves were plotted to compare the performance among methods under the two p scenarios respectively. ROC is a plot of the true positive rate (Sensitivity) against the false positive rate (100-Specificity) for different possible cut-off points. In our case, we define sensitivity as the combined power for four true variables, while 100-specificity is the combined type I error rate for all the null variables. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

As illustrated in Figure 3.1, the proposed Bayes case control method is obviously the best among the 6 approaches for both p cases. Our method tends to have much smaller combined type I error and provide better power at all times. Note that the x axis scale is only shown within [0,0.5] for display purposes and because a type I error being too large is usually not acceptable. Some ROC curves are cut off due to the scale limit.

Another case is considered for data that contains covariates which are correlated, but not associated with outcome, in addition to the outcome-dependent variables mentioned above. We added another four variables using a saturated log-linear model similar to (3.18) but having the same coefficients in both disease groups instead. The coefficients are set in Table 3.3. The new mechanism results in the extra four covariates correlated to each other but not impacting disease. All the other $p - 8$ variables are generated independently from a discrete uniform distribution. We then created two new ROC curves for both p cases shown in Figure 3.2. Compared with Figure 3.1, we obtained a slightly inflated false positive rate. However, it still performs much better than the other methods.

3.4.2 Simulation from latent class models

We now perform another simulation study with data generated from a latent class model rather than a log-linear model. We again had 50 case and 50 control subjects having p binary predictors with (i) $p = 20$ and (ii) $p = 100$ for each data replicate.

We assume four predictors are associated with the outcome in the true model, whereas those four variables ($j = 2, 4, 12, 14$) are correlated by introducing the multiple latent classes $z = 1, \ldots, k$, with each latent class having different marginal probabilities for those four variables. We assumed 80% of individuals fell into the first latent class, with the remaining individuals in a second latent class ($k = 2$). Furthermore, the variable dependence on disease outcome can be induced by letting the marginal probabilities under each latent class vary by disease outcome. In particular,

$$Pr(x_{ij} = c_j \mid d_i = d, z_i = h) = \lambda^{(j)}_{dhc_j}, \quad h = 1, 2;\ d = 0, 1 \tag{3.22}$$

where vector $\boldsymbol{\lambda}^{(j)}_{dh}$ only varies by d and h for $j = 2, 4, 12, 14$. All the remaining predictors were generated from a discrete uniform distribution with $\lambda^{(j)}_{dh} = \lambda^{(j)}_{0} \equiv (\frac{1}{2}, \frac{1}{2})'$. Within a latent class and the disease group, all the variables are conditionally independent. However, marginalizing out the latent class indicator, one obtains dependence in those variables that have different marginal probabilities across the latent classes, conditional on the disease outcome. Additionally, it becomes clear that having the probability of those correlated covariates differing by the disease group implies the association between the outcome and those four variables.

We generated 1000 simulated datasets, then ran the Gibbs sampler of Section 3.3 with satisfactory mixing and convergence rates. We also computed the posterior samples of the odds ratios to assess the ROC performance. As in Figure 3.3, our approach outperforms the other methods for both $p = 20$ and 100 with much better combined power and lower type-I error.

To examine the performance of all methods when including correlated covariates not associated with outcome, we generated another 4 correlated variables for both p cases with different marginal probabilities in each latent class but do not vary by the disease group (i.e. $Pr(x_{ij} = c_j \mid z_i = h) = \lambda^{(j)}_{hc_j}, \quad h = 1, 2$). The corresponding ROC curves, based on a new 1,000 simulated data, are provided in Figure 3.4. It shows that the performance is similar to Figure 3.3, but more complex data adding correlated covariates not related to outcome would slightly affect the false positive rate.

The power and type I error rates for screening based on 95% credible intervals for odds ratios not including 1 are provided in Table 3.4 and Table 3.5. In both simulation scenarios, the power does not appear to be affected with respect to adding more correlated covariates. The type I error, as we observed in the ROC curves, is slightly inflated but remains well below 0.05 in a more correlated covariate structure.

3.5 Application to the National Birth Defects Prevention Study

The new method is motivated by the National Birth Defects Prevention Study, which is a U.S. nation-wide case-control study with approximately 26,000 cases and 10,000 controls that was started in 1997. The study was designed to evaluate environmental, behavioral, biomedical, sociodemographic, genetic, and occupational factors associated with the occurrence of congenital malformations. There are 9 states currently participating in this study: Arkansas, California, Georgia, Iowa, Massachusetts, New York, North Carolina, Texas, and Utah. The study population area covers roughly 10% of all births in the United States. The subjects are comparable to that of the general U.S. population with respect to maternal age, race, ethnicity, and education level. There are 54 birth defects and 177 potentially important risk factors of interest.

We employ our case control Bayesian methods to investigate the associations per defect using odds ratios (OR) as the measure of association. 19 defects (any heart defect, conotruncal, left ventricular outflow tract obstruction (LVOTO), right ventricular outflow tract obstruction (RVOTO), ventricular septal defect perimembranous (VSDPM), atrial septal defect not otherwise specified (ASD2NOS), ventricular and atrial septal defect (VSD_ASD), neural tube defects (NTD), ear, cleft palate, cleft lip with cleft palate, cleft lip without cleft palate, esophageal, anorectal, hypospadias, limb, craniosyn, diaphragm, gastroschisis) are analyzed due to the insufficient cases for other specific defects.

Our Bayesian procedure estimates a 95% credible interval for the marginal odds ratio for each factor and selects those factors as significant having 95% interval not including one. We ran $20,000$ iterations with first $5,000$ iterations as burn-in and collecting every fifth sample post burn-in to thin the chain. The effective sample sizes for marginal odds ratios are ranging from 2,783 to 3,000 which suggests the thinned samples are close to independent. We output the results for each defect with 177 predictors as one row, and combine all 19 analyses for different outcomes into a 19×177 matrix which is then separated into three heat maps displayed in the top figures in Figures 3.5 - 3.7.

In the upper panel of Figure 3.5, gastroschisis is positively associated with a parental age less than 24 years old (OR ≈ 5) relative to age 25-30, and mother's low education (OR = 3.6). Hispanic parents (OR= 3.5), molar pregnancy (OR=3.9), and type 1 diabetes (OR=5.1) are risk factors for ear defect. Esophageal atresia is affected by mothers' fertility procedure with OR 3.8, sulfamethoxazole and trimethoprim exposure with both ORs 3.2. Molar pregnancy is also associated with ventricular and atrial septal defect (OR:4.3) and cleft lip without cleft palate (OR:3.6). Pelvic inflamatory disease (PID) is a risk effect for ventricular septal defect perimembranous, cleft palate, craniosyn, and ventricular and atrial septal defect with ORs around 3.5 for the first three defects and 5.3 for the last one. Meclizine exposure has an impact on multiple defects including left ventricular outflow tract obstruction (OR:4.3), ventricular and atrial septal defect (OR:7.0), neural tube defects (OR:4.2), ear (OR:8.0), cleft palate (OR:9.2), craniosyn (OR:6.7), and diaphragm (OR:7.2).

In Figure 3.6 top figure, human immunodeficiency virus (HIV) exposure is strongly associated with most defects with ORs larger than 8 except conotruncal, atrial septal defect, neural tube defects, cleft lip with cleft palate, and esophageal. Human papillomavirus (HPV) exposure plays a role in right ventricular outflow tract obstruction (OR:4.2), ventricular septal defect perimembranous (OR:4.4), atrial septal defect (OR:3.3), and ear (OR: 5.7). Pelvic inflamatory disease medication affects ventricular septal defect perimembranous (OR: 3.8), ventricular and atrial septal defect (OR:5.2), cleft palate (OR:3.4), anorectal and craniosyn with OR 3.6 for the last two. Pregnancy outcome (stillbirth) is related to neural tube defects (OR:15.0), anorectal (OR:3.3), and gastroschisis (OR:5.5). Induced abortion is instead related to craniosyn with OR 4.5, ventricular and atrial septal defect, neural tube defects, and ear all with ORs greater than 8. Gestational age of 32-36 weeks or birth weight more than 4,000 grams are more likely to have esophageal(OR$\approx$ 3.5) and gastroschisis(OR$\approx$ 9). Gastroschisis is additionally associated with parental substance abuse with ORs 3-4.

In Figure 3.7 top panel, in terms of mother's occupation, protective service occupations have a risk in developing ventricular and atrial septal defect, cleft lip without cleft palate, diaphragm with ORs between 3 and 4 in the baby. Construction work can lead to ventricular and atrial septal defect (OR:10.2) and cleft lip without cleft palate (OR:5.4), while installation, maintenance, and repair occupations are associated with right ventricular outflow tract obstruction, atrial septal defect, ventricular and atrial septal defect, cleft palate, cleft lip without cleft palate, and esophageal with large ORs. Furthermore, military occupations have a dramatically high risk (all ORs > 5) in developing defects such as right ventricular outflow tract obstruction, ventricular septal defect perimembranous, ventricular and atrial septal defect, neural tube defects, ear, cleft palate, esophageal, limb, craniosyn, and diaphragm defect. Among all the solvents, benzene is associated with left ventricular outflow tract obstruction, ventricular septal defect perimembranous, ventricular and atrial septal defect, and diaphragm with ORs between 3.5 and 5.7. Carbon tetrachloride, on the hand, has an impact on right ventricular outflow tract obstruction (OR:4.7), ventricular and atrial septal defect (OR:12.3), cleft palate (OR:5.0), and craniosyn (OR:5.1). We also found that right ventricular outflow tract obstruction and limb defect are affected by most solvents with moderate ORs.

From simulations we learned that our approach has better power and lower type-I error when compared with other existing methods. To compare with the proposed approach in application, we chose logistic regression without multiple testing as it performs relatively better than the other existing methods (LASSO, CART, and logistic regression with multiple testing correction) in simulations. The results are shown in the bottom panels of Figures 3.5 - 3.7. We detected more significant associations than logistic regression. The following are some interesting examples of associations selected as significant in the proposed method but not in logistic regression.

Interestingly, in Figure 3.5, our model identified meclizine exposure as a risk factor for left ventricular outflow tract obstruction (OR=4.3), ventricular and atrial septal defect (OR=7.0), neural tube defects (OR=8.0), craniosyn (OR=6.7), and diaphragm (OR=7.2), while logistic regression using p-value < 0.5 did not. Likewise, in Figure 3.6, HIV was discovered to have an association with left ventricular outflow tract obstruction, right ventricular outflow tract obstruction, ventricular septal defect perimembranous, ventricular and atrial septal defect, cleft palate, cleft lip without cleft palate, ear, anorectal, hypospadias, limb, craniosyn, diaphragm with ORs around 10. Some of the occupations in Figure 3.7 drew our attention: mothers who have maintenance/repair jobs are more likely to develop atrial septal defect, ventricular and atrial septal defect, cleft palate (OR$\approx$5). Women who serve in the military have a higher risk in developing right ventricular outflow tract obstruction, ventricular septal defect perimembranous, ventricular and atrial septal defect, neural tube defects, cleft palate, limb, craniosyn, and diaphragm defects with all ORs larger than 5. For a father on the other hand, the risk is higher for heart defect (OR=1.6), ventricular and atrial septal defect (OR=3.4), and craniosyn (OR=2.7). Moreover, logistic regression was not able to detect chemical solvents that could affect the new born babies. We found out that carbon tetrachloride is associated with right ventricular outflow tract obstruction (OR=4.7), cleft palate (OR=5.0), hypospadias (OR=3.5), and craniosyn (OR=5.1). Benzene also has a harmful effect on conotruncal defect (OR=2.9), left ventricular outflow tract obstruction (OR=3.2), ventricular septal defect perimembranous (OR=4.2), and ventricular and atrial septal defect (OR=5.7). More details are in Appendix II.

3.6 Discussion

In this paper, a new method utilizing a sparse parallel factor analysis model has been proposed for case control designs. It has been shown through simulation that it has exceptional performance in identifying true predictors while keeping the type I error rate very small. The outstanding performance, compared to existing methods, is due to flexible distribution modeling for the retrospective likelihood and borrowing information among variables in our model. This method can be applied to any case control study that has many categorical covariates with an interest in investigating the association.

Our paper is focused on building a flexible nonparametric model for the data to improve inferences on marginal associations, but an important next step is to develop approaches for inferences on conditional associations.

As for analyses containing multiple outcomes such as the National Birth Defects Prevention Study, there is clear evidence of dependence over the rows of the figures showing results (e.g. Figure 3.6). This suggests that certain factors are risk factors for multiple different birth defects. It would be interesting to develop a new method to group similar factors effects on multiple outcomes. If we had alternatively used a prospective logistic regression, then it would be very natural to build a hierarchical regression model (Coull et al. 2001). Our nonparametric Bayes extension would be a competitor especially in a high-dimensional case.

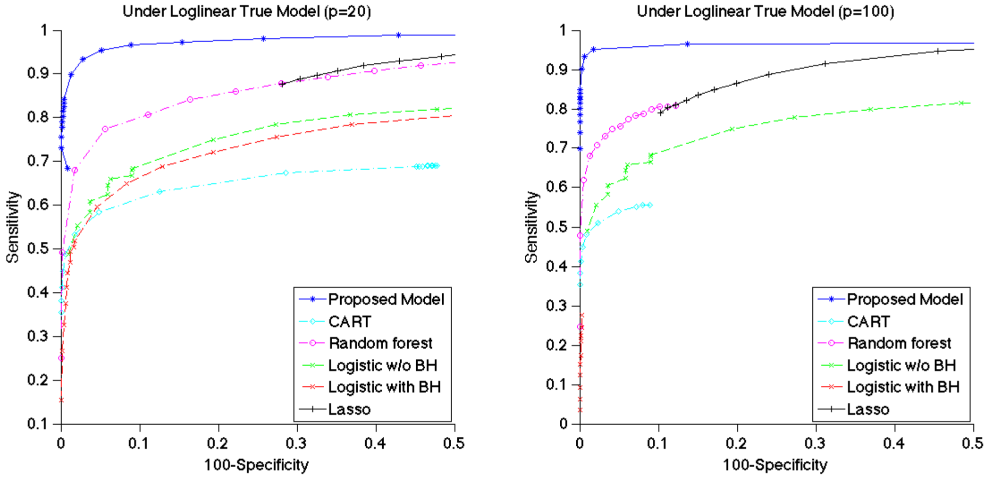

Figure 3.1: ROC curves under loglinear true models – Left: p=20; Right: p=100.

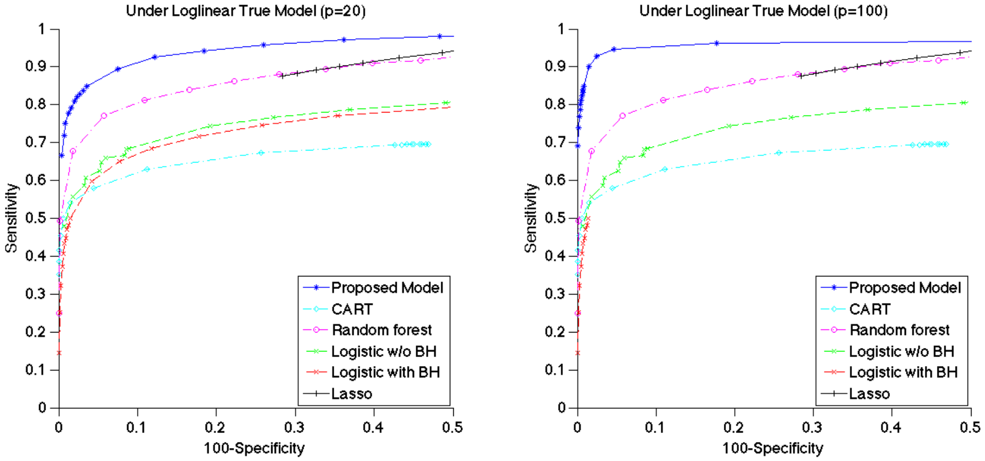

Figure 3.2: ROC curves under loglinear true models with more correlated covariates – Left: p=20; Right: p=100.

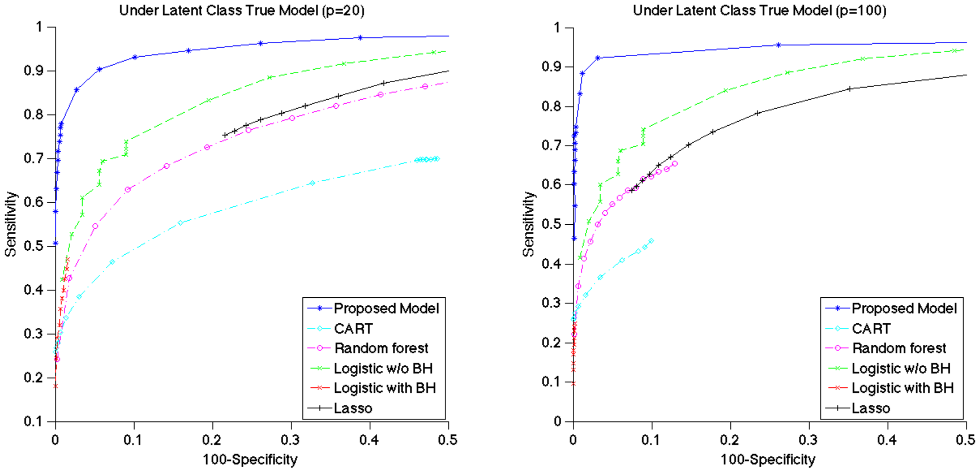

Figure 3.3: ROC curves under latent class true models – Left: p=20; Right: p=100.

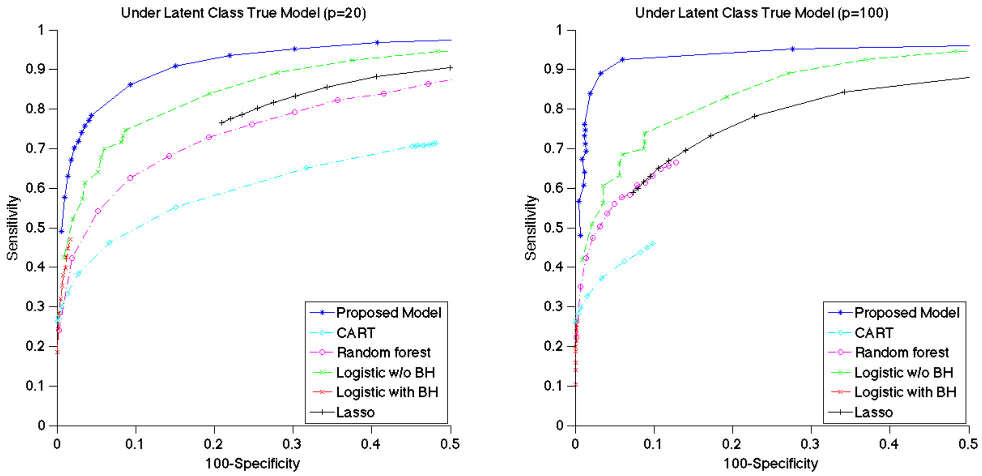

Figure 3.4: ROC curves under latent class true models with more correlated covariates – Left: p=20; Right: p=100.

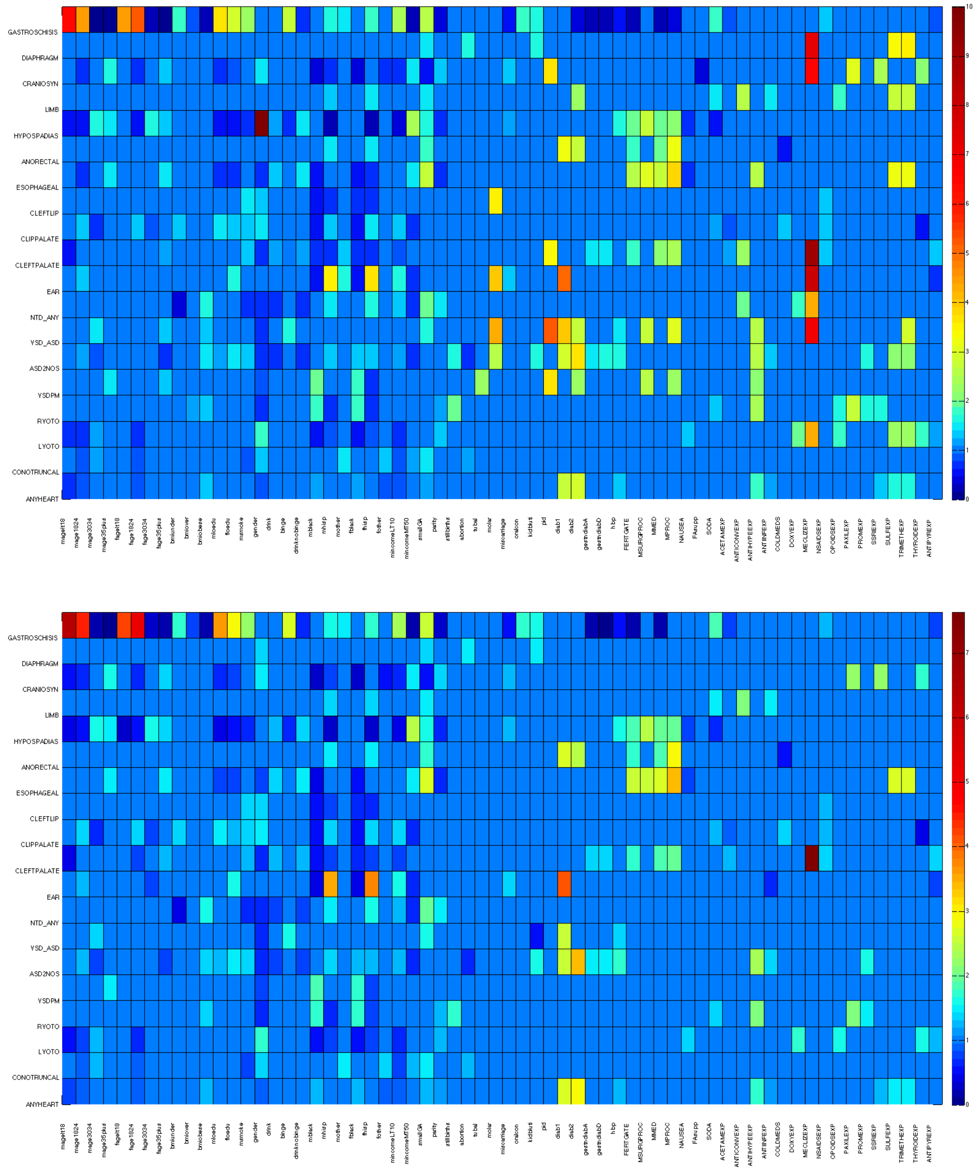

Figure 3.5: Part 1 – significant odds ratios between 19 birth defects and 64 potential factors. Top: using proposed method; Bottom: using 1-to-1 logistic regression.

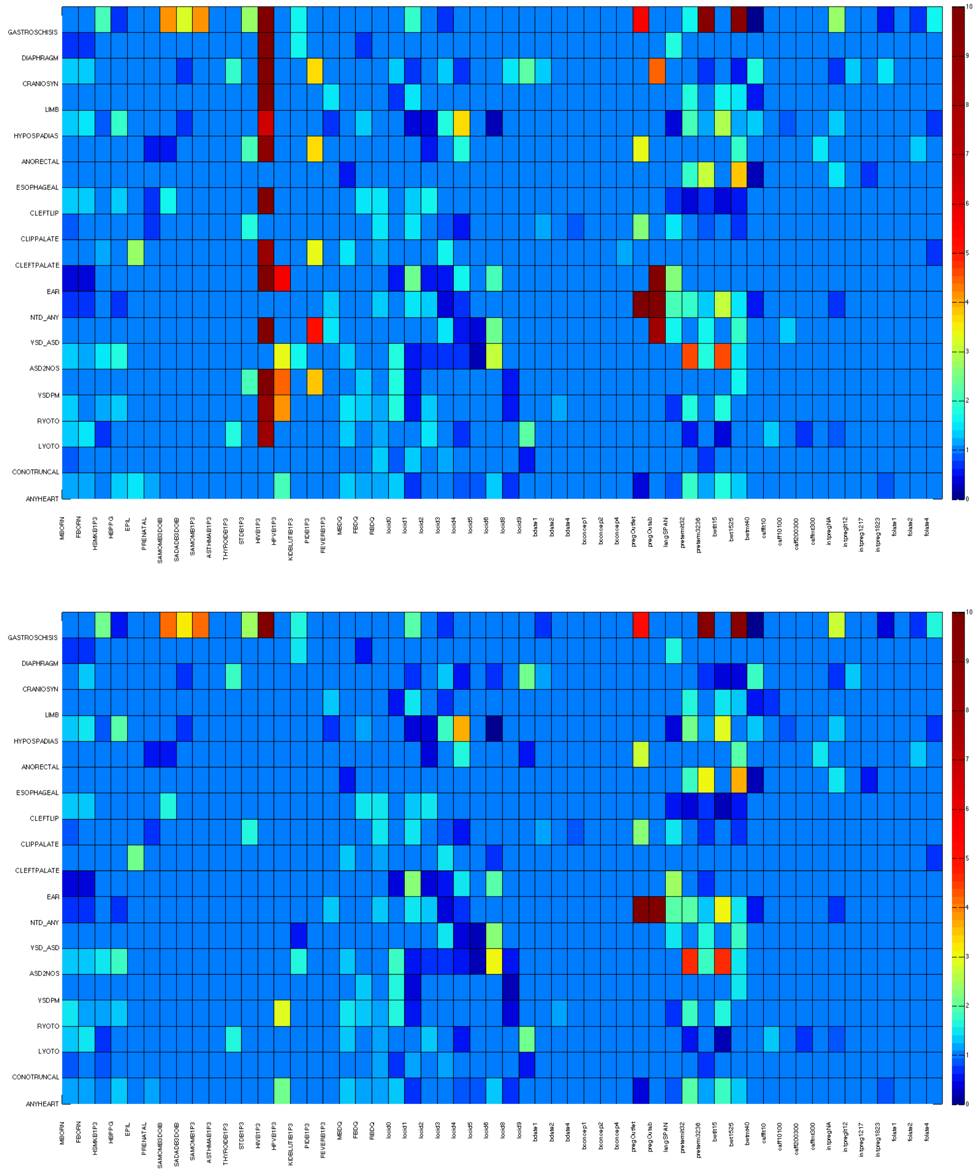

Figure 3.6: Part 2 – significant odds ratios between 19 birth defects and 64 potential factors. Top: using proposed method; Bottom: using 1-to-1 logistic regression.

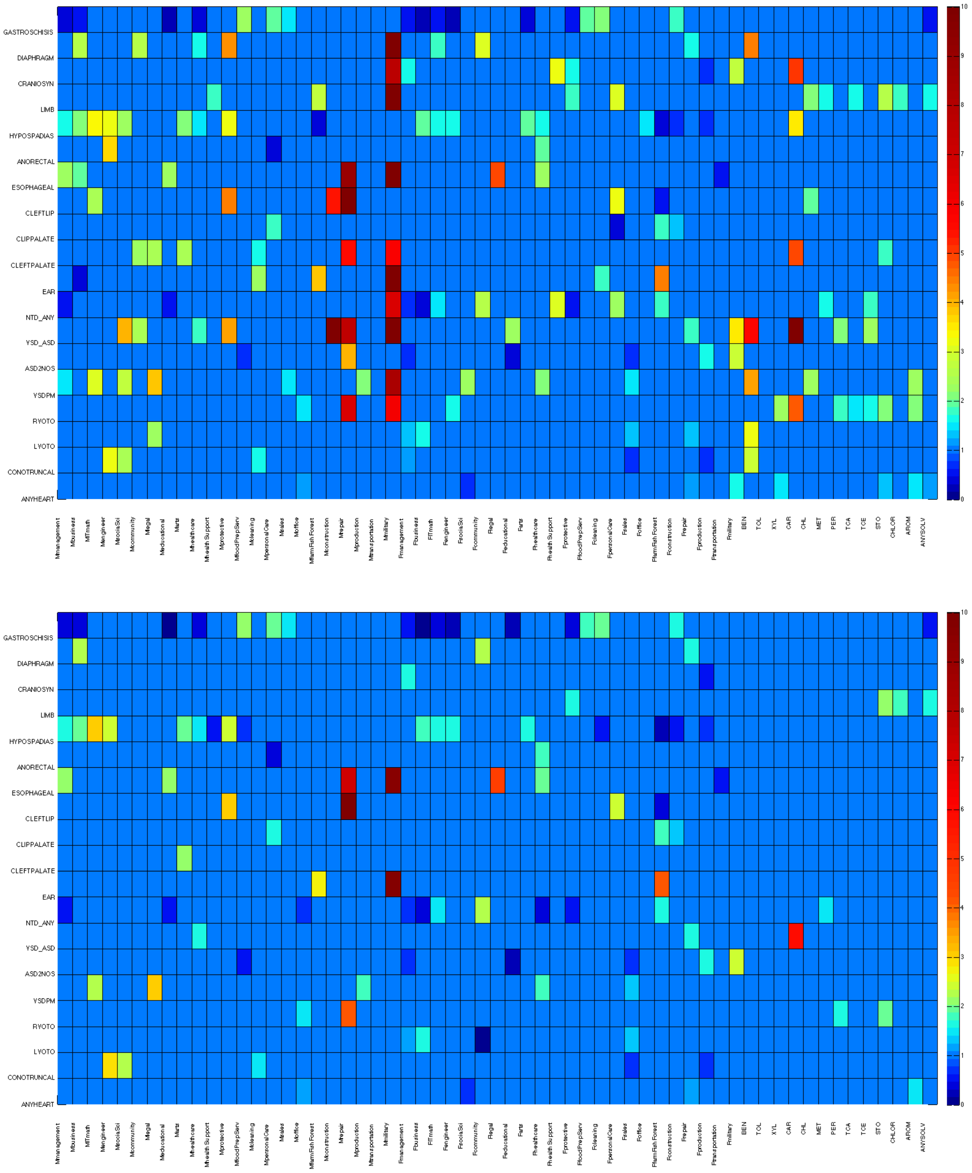

Figure 3.7: Part 3 – significant odds ratios between 19 birth defects and 64 potential factors. Top: using proposed method; Bottom: using 1-to-1 logistic regression.

Table 3.1: True Coefficients

	β_2^d	β_4^d	β_{12}^d	β_{14}^d	$\beta_{2,4}^d$	$\beta_{2,12}^d$	$\beta_{2,14}^d$	$\beta_{4,12}^d$	$\beta_{4,14}^d$	$\beta_{12,14}^d$
d=0	0.5	-1.5	-2	1	-0.5	0.5	0	0.5	-0.5	-0.5
d=1	3	-3	-0.5	4	-0.5	0.5	0	0.5	-0.5	-0.5

Table 3.2: True Coefficients (continued)

	$\beta^d_{2,4,12}$	$\beta^d_{2,4,14}$	$\beta^d_{2,12,14}$	$\beta^d_{4,12,14}$	$\beta^d_{2,4,12,14}$
d=0	0.25	0	0	0.5	0
d=1	0.25	0	0	0.5	0

Table 3.3: True Coefficients for Another Four Dependent Variables

β_1	β_3	β_{11}	β_{13}	$\beta_{1,3}$	$\beta_{1,11}$	$\beta_{1,13}$	$\beta_{3,11}$
1.5	-0.5	2	-1	-0.5	0.25	0	-0.5
$\beta_{3,13}$	$\beta_{11,13}$	$\beta_{1,3,11}$	$\beta_{1,3,13}$	$\beta_{1,11,13}$	$\beta_{3,11,13}$	$\beta_{1,3,11,13}$	
0.5	-0.5	-0.25	0.25	-0.25	0	0	

Table 3.4: Power Using 95% Credible Intervals

Power	Log-Linear Model		Latent Class Model	
	p=20	p=100	p=20	p=100
4 correlated covariates	0.791	0.801	0.697	0.664
8 correlated covariates	0.792	0.801	0.701	0.673

Table 3.5: Type I Error Using 95% Credible Intervals

Type I Error	Log-Linear Model		Latent Class Model	
	p=20	p=100	p=20	p=100
4 correlated covariates	0.002	3.75E-04	0.003	0.003
8 correlated covariates	0.016	0.004	0.022	0.008

CHAPTER 4

Sample Size Re-estimation in Longitudinal Group Sequential Design

4.1 Outline

With the goal of designing and analyzing a longitudinal trial using group sequential design along with the concern of insufficient power, the background information on how to determine the sample size for fixed design and group sequential design is provided in section 4.2. In Section 4.3, we introduce the information-based sample size re-estimation method in group sequential design to be utilized in longitudinal trials. Adaptation rules for updating sample size have been developed which will be described and illustrated by examples. Section 4.4 provides the simulation results for our method compared with fixed design and group sequential design without sample size re-estimation. A simple data analysis is presented in section 4.5. Finally, section 4.6 contains a discussion and summary of the results.

4.2 Sample Size Determination for Longitudinal Analysis

4.2.1 Model Notation

We model the longitudinal data as in Liang and Zeger (2000) which includes the baseline value as part of the response vector. The marginal mean model is given as

$$E(Y_{ijt}) = \mu_0 + \gamma_{jt} I(treatment = j) I(time = t; t > 0), t = 0, 1, \ldots, T, \tag{4.1}$$

where i indexes the subject, j indexes the treatment group ($j = 1$ for the control group, $j = 2$ for the active treatment group), and t indexes the time point ($t = 0$ for the baseline and $t = 1, \ldots, T$ for the post baseline time points). In addition, μ_0 is the mean response at $t = 0$, which is constrained to be the same for both treatment groups due to randomization. As a result, the baseline measurement is not considered as an 'outcome' to treatment although it is included in the response vector together with the post baseline measurements. Hereafter, we will refer to (4.1) as the constrained longitudinal data analysis (cLDA) model (Lu et al. 2009). The parameter γ_{jt} is the effect of change from baseline at time t for treatment j ; hence, $\mu_{jt} = \mu_0 + \gamma_{jt}$ is the mean at time t for treatment j. Let $\theta_j = (\mu_{j1}, \ldots, \mu_{jT})'$ denote the mean vector for post baseline measurements for treatment j. The mean parameters for model (4.1) can be written as $\psi = (\mu_0, \theta_1', \theta_2')'$.

The cLDA model assumes that baseline and post baseline values are jointly multivariate normal with $\Sigma = \{\sigma_{st} : s, t = 0, 1, \ldots, T\}$. This matrix can be represented as a correlation matrix sandwiched by the diagonal matrix of standard deviations where the correlation matrix is given by $R = \{\rho_{st} : s, t = 0, 1, \ldots, T\}$ and the standard deviations are with respect to the pure error within each longitudinal measurement. Let us denote n_{jt} as the subjects retained at time t in treatment j with the assumption of a monotone missing data pattern, and $n_j = n_{j0}$ as the total number of subjects in treatment j at baseline. Define $r_{jt} = n_{jt}/n_j$ as the proportion of enrolled subjects retained at time t in treatment j. Note that the retained people include those that are are still under active follow-up but exclude those who drop out. The drop-out rate, the proportion of enrolled subjects dropped out between time t and $t + 1$, is $p_{jt} = (n_{jt} - n_{j,t+1})/n_j$. It follows immediately that $p_{jt} = r_{jt} - r_{j,t+1}$.

4.2.2 Fixed Design

Suppose we are interested in a linear contrast of the treatment means across time,

$$\delta = c'(\theta_2 - \theta_1), \tag{4.2}$$

where c is a contrast vector of length T corresponding to the T post baseline assessment time points. For instance, $c = (0, \ldots, 0, 1)$ is for treatment comparison at the last time point. If we want to detect the treatment effect δ_a, the Fisher information, I, to be needed based on a two-sided Z-test for $H_0 : \delta = 0$ versus $H_a : \delta = \delta_a$ with power $1 - \beta$ at significant level α can be derived as

$$I = (\frac{Z_{\alpha/2} + Z_\beta}{\delta_a})^2, \tag{4.3}$$

Now we can determine the sample size with the knowledge that

$$I = Var^{-1}(\hat{\delta}), \tag{4.4}$$

where $Var(\hat{\delta})$ is a function of sample size with some nuisance parameters. $\hat{\delta}$ denotes the estimate of δ to be calculated from data. This variance varies among different types of trials. For longitudinal trials, in particular, based on the cLDA model we defined earlier in (4.1), the variance inverse of $\hat{\delta}$ incorporating missingness is given by

$$Var^{-1}(\hat{\delta}) = (\frac{c'S\Lambda_1^{-1}Sc}{n_1} + \frac{c'S\Lambda_2^{-1}Sc}{n_2})^{-1}, \tag{4.5}$$

where n_j is sample size for treatment j, $j = 1, 2$. For simplicity, we assume the randomization ratio is 1, so $n_1 = n_2 = \frac{N}{2}$ though extensions are trivial. The parameter c is denoted as above, $S = diag(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{TT}})$ denotes the standard deviations at post baseline time points. Λ_j is given by

$$\Lambda_j = \sum_{t=1}^{T} p_{jt} \begin{pmatrix} R_{tt\cdot 0}^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \tag{4.6}$$

where $R_{tt\cdot 0}^{-1} = R_{tt} - R_{0t}' R_{0t}$, $R_{tt} = \{\rho_{ij} : i, j = 1, \ldots, t\}$, and $R_{0t} = (\rho_{01}, \ldots, \rho_{0t})$. The proof of the derivation of $Var(\hat{\delta})$ for cLDA is in Lu et al. Lu et al. (2009). We note that the nuisance parameters in (4.5) are essentially R, S, and r_j, where R, S are the correlation matrix and standard deviations respectively, and $r_j = (r_{j1}, \ldots, r_{jT})$ is the retention rate across time for treatment j. Without loss of generality, we assume $r_1 = r_2$ hereafter. Therefore, connecting (4.3), (4.4), and (4.5) while assuming the nuisance parameters (R, S, and r_1), the calculation of sample size is straightforward.

4.2.3 Group Sequential Design

Group sequential design has the advantage to stop early for efficacy or futility. Rather than performing only one analysis at the end of the study, we perform up to K analyses at interim monitoring times $\tau_1, \tau_2, \ldots \tau_K$, respectively, and terminate the study at the first interim look that rejects the null hypothesis. However this flexibility to possibly terminate early comes at a cost. In particular, the type-I error is inflated due to the multiple interim hypothesis tests if we keep the stopping boundaries unchanged. We thus need to adjust the stopping criteria appropriately such that the type-I error is controlled. Moreover, in order to achieve a power of $1 - \beta$ while controlling the type-I error rate, the information has to be inflated by

$$I_{max} = (\frac{Z_{\alpha/2} + Z_{\beta}}{\delta_a})^2 \times IF(\Delta, \alpha, \beta, K), \tag{4.7}$$

where I_{max} is denoted as the information at the final look, which is a counterpart of the I in (4.3) under the framework of group sequential design. The function IF$(\cdot)$ is an inflation factor that depends on α, β, K and Δ. The parameter Δ is defined with respect to the shape of the stopping boundaries over the K repeated tests. Mehta and Tsiatis Mehta and Tsiatis (2001), relying on theoretical results by Scharfstein et al. Scharfstein et al. (1997), has the detailed derivation and examples of the inflation factor. Keep in mind that this maximum information does not require any knowledge of unknown nuisance parameters.

In practice, however, an estimated number of patients is required at the time of study design. The corresponding sample size determination follows the same strategy as that used in fixed design. One can similarly connect equations (4.7), (4.4), and (4.5) to solve the maximum sample size (N_{max}). In short, the required maximum sample size can be computed in the following two steps:

(1) Utilize the `'gsDesign'` R package developed by Anderson (Anderson (2014),Zhu et al. (2011)) to calculate I_{max} once we define the necessary parameters in (4.7),

(2) Convert I_{max} to N_{max} by using (4.5) which is essentially a function of N_{max} with some nuisance parameters R, S and r_1.

These calculations are possible in many other group sequential design packages (e.g., RCTdesign in R, PEST, EaSt) as long as the correct variance and information timing are used. Given accurate assumptions of nuisance parameters, collecting N_{max} subjects will, in the end, result in obtaining I_{max} while achieving the desired power and maintaining the type-I error. For example, if we plan for 4 looks with an equal-spaced information-based design, 25% of I_{max} is expected at each iterim analysis. But in real world studies, the sample size may be incorrect since the nuisance parameters are unknown and it often happens that we do not have good estimates for these parameters at hand. The power will be affected as a consequence. Thus, our method with the aim of tackling this problem is introduced in the next section. It is noted that the inflation factor only allows us to maintain the power and control the type-I error provided that the assumptions of R, S and r_1 are correct. Instead of maintaining power by adjusting the sample size, it is also possible to fix the total sample size at N_{fix} (sample size for a fixed-sample trial) and then evaluate the effect of the monitoring plan on trial power, which is however not a focus of this paper. A design with analyses spaced by equal amounts of information is assumed from now just for simplicity in illustration; it is easy to extend to an unequally-spaced information.

4.3 Information-Based Sample Size Re-Estimation

4.3.1 Sample Size Re-estimation

Define $I(\tau_k)$ as the information at the k^{th} interim analysis. Assuming complete follow-up for observations, it can be easily shown by (4.5) that,

$$\frac{I(\tau_k)}{I_{max}} = \frac{N(\tau_k)}{N_{max}}. \tag{4.8}$$

Thus, one can re-estimate N_{max} at each interim using:

$$N^*_{max} = N(\tau_k) / \frac{I(\tau_k)}{I_{max}}, \tag{4.9}$$

where $N(\tau_k)$ is the sample size of subjects with completed longitudinal visits at the k^{th} look, $I(\tau_k) = Var^{-1}(\hat{\delta}(\tau_k))$ is estimated from the data, and I_{max} is fixed under the design by (4.7). Note the denominator on the right hand side is an evaluated information fraction that is to be compared with the anticipated information fraction (e.g. 25% if currently at 1^{st} interim with 4 looks in total planned). If it is larger than planned, N^*_{max} will be smaller than original N_{max}, and vice versa. Updating the maximum sample size at each look can correct inaccurate assumptions for nuisance parameters and maintain the power while controlling the type-I error rate.

This approach is fairly easy to understand and implement without unblinding the treatment effect, however, it has the drawback that only completed and drop-out subjects are contributing to the interim analysis. An alternative way to re-estimate the maximum sample size is to first estimate the nuisance parameters using all available data at the current interim, and then use them as input in the calculation of N_{max} as discussed in section 2.3. It can make use of all data including ongoing patients but it loses the simplicity of the previous method as we need to estimate the many nuisance parameters at each interim and have to use a statistical package to obtain the updated sample size. It may also unblind the trial in that all the nuisance parameters need to be updated at each interim. When the enrollment is slow, the method in (4.9) is more attractive in practice since the additional information provided by the ongoing patients may be neglectable. Otherwise, the latter method is recommended if unblinding is not concerned. The power analysis results based on the second approach are provided later in Section 4.

4.3.2 Adaptation

We develop a sample size adaptation rule in the following based on the practical characteristics of clinical trials. Note that the 'overrun of patients' below in (b) stands for all the subjects enrolled so far including not only the completed/discontinued ones but also those still continuing, while the current sample size is only with regard to completed/discontinued patients.

(a) If current sample size is enough: meaning I_{max} is reached, stop the trial regardless of whether efficacy or futility is detected.

(b) If overrun of patients is enough to provide sufficient information: stop enrolling more patients but keep collecting data for enrolled patients.

(c) If next planned sample size is enough: stop the enrollment when the updated maximum sample size is reached.

(d) If next planned sample size is not enough but original maximum sample size is: continue enrollment to the next planned interim.

(e) If original maximum sample size is not enough: use the updated maximum sample size but with a upper limit depending on practical aspects of certain trials (e.g. two times the previous maximum sample size) and continue enrollment to the interim analysis.

For illustration purposes, the following is an example of adaptation. Suppose the longitudinal study is to be designed for up to four interim monitoring looks including a final analysis; each subject is expected to have four longitudinal visits to the clinic after the baseline measurement and the corresponding N_{max} is 800 with all the necessary parameters assigned. Hence $K = 4$, $T = 4$, and $N_{max} = 800$. Let $k = 2$, $N(\tau_2) = 400$, 480 patients have been enrolled and the current plan for $N(\tau_3) = 600$, the consequent adaptation rule at 2^{nd} interim can be represented as:

- If $N^*_{max} \leq 400$: Stop the trial regardless.
- If $400 < N^*_{max} \leq 480$: Stop enrolling more patients but keep collecting data for enrolled patients, and do one final analysis.
- If $480 < N^*_{max} \leq 600$: Stop the enrollment when N^*_{max} is reached and perform one final analysis.
- If $600 < N^*_{max} \leq 800$: Continue the next planned interim.

- If $N^*_{max} > 800$: Use $\min(2{\times}800,\ N^*_{max})$ as our new maximum sample size and continue the interim analysis with three-fourths of the new maximum sample size.

In actuality, the total sample size only needs to be increased when either the planned sample size has been reached (with incomplete follow-up), when the last interim analysis is performed, or when planning for increased patient enrollment and clinical supplies is needed.

Next, we use Figure 1 to illustrate the sample size adaptation for an entire trial if we use the method in (4.9). Once again, we plan 4 looks at the design stage, and analyze $200, 400, 600$ and 800 patients at a time. Different colors correspond to each interim look. At the 1^{st} interim, we see that instead of anticipated 25% of I_{max}, 30% of I_{max} has been observed. The following re-estimation of N_{max} tells us 667 patients are required to obtain the maximum information in the end rather than the original 800 patients. Because 667 falls into the fourth bullet of the above adaptation rule, we need to collect 200 more patients as planned and perform the 2^{nd} analysis. At that time, another 50% of I_{max} has been gathered. A total of 80% is considerablely larger than what we anticipated (50%), which suggests that the assumption of nuisance parameters is very conservative. Therefore, the next analysis is planned as the final analysis at $N^*_{max} = 500$ and enrollment stops at that point. The final analysis was conducted with 1.03 times the I_{max} observed, that is also evidence that our approach is useful to save time and resources while maintaining all of the good statistical characteristics. In addition to the above sample size planning rules, the trial may stop if an efficacy or futility bound is crossed.

4.3.3 Interim Analysis Procedure

Using what we supposed in the previous subsection ($K = 4$, $N_{max} = 800$), we plan to enroll subjects continuously and conduct each interim analysis cumulatively for every 200 subjects with complete visits that have been gathered. Our interest is to test whether the treatment difference at the final look (δ_a) is 0.25. Using all the completed/drop-out data (for method 1) or the available data (for method 2) to fit the constrained longitudinal model (4.1) assuming unstructured covariance structure Diggle (2002), the testing procedure at interim k is given in the following: $k = 1 \ldots K$,

1. Estimate $I(\tau_k) = Var^{-1}(\hat{\delta}(\tau_k)) = s.e.^{-2}(\hat{\delta}(\tau_k))$.

2. Estimate $T(\tau_k) = \frac{\hat{\delta}(\tau_k)}{s.e.(\hat{\delta}(\tau_k))}$.

3. Update the actual information fraction vector up to current k^{th} interim:

$$(\frac{I(\tau_1)}{I_{max}}, \frac{I(\tau_2)}{I_{max}}, \ldots, \frac{I(\tau_k)}{I_{max}}, \frac{k+1}{K}, \ldots, 1), \tag{4.10}$$

   where $\frac{I(\tau_1)}{I_{max}}, \frac{I(\tau_2)}{I_{max}}, \ldots, \frac{I(\tau_k)}{I_{max}}$ are observed information fraction up to k^{th} interim look, and $\frac{k+1}{K}, \ldots, 1$ are planned information fraction after k^{th} interim. Then one can calculate the corresponding stopping upper and lower boundaries by updating bounds using the methods of Lan and DeMets Lan and DeMets (1983); this can be done, for example using the `'gsDesign'` R package.

4. $$\begin{cases} \text{stop for efficacy,} & \text{if } T(\tau_j) \geq \text{upper bound} \\ \text{continue,} & \text{if lower bound} < T(\tau_j) < \text{upper bound} \\ \text{stop for futility,} & \text{if } T(\tau_j) \leq \text{lower bound.} \end{cases}$$

One extra step is required here if we do not stop at the current analysis. That is to re-estimate N_{max} as in section 4.3.1 and to adapt the new maximum sample size discussed in section 4.3.2. It is noted that besides futility or efficacy, the study is terminated when I_{max} is reached, or at the K^{th} final analysis.

4.4 Simulation Study

As discussed earlier, incorrect assumptions of nuisance parameters will lead to an incorrect sample size which will affect the power. In this section, we verify through simulations that our method works as planned in the sense that the power is maintained while preserving the type-I error rate. The expected sample size (E(n)) is another characteristic of interest. We define n, for group sequential design, as the number of enrolled when stopping early, but as the number of analyzed when stopping at the final look. For the fixed design, however, n is always N_{fix}. In the meantime, we compare performance of our approach with that of fixed design and that of group sequential design without the sample size re-calculation.

The procedure is composed of four parts: design, data generation, testing and results comparison. All the trials are designed for 90% power to detect a treatment difference at the last (4^{th} post baseline) measurement of 0.25 using 4-look one-sided O'Brien-Fleming stopping boundaries with a type-I error 0.025. The assumptions for nuisance parameters (R, S, r_1) are that R_0 = compound symmetry with correlation coefficient 0.579, $S_0 = 0.8$, and $r_{10} = (0.91, 0.84, 0.77, 0.70)$. It is then straightforward to obtain the sample size for a fixed design (N_{fix}) and a group sequential design (N_{max}). Since the sample size only depends on $R, S,$ and r_1 through information, and the true values of these nuisance parameters (R, S, r_1) could be different from what we planned (R_0, S_0, r_{10}). Thus, we generate 1000 datasets under each of 18 different combinations of (R, S, r_1) to see how the power and type-I error behave; we expect that the one scenario with the assumed values will result in good power with type-I error controlled. True R is chosen to be among compound symmetry (cs), toplitz, and AR(1); S is either

0.8 or 0.925, whereas the three options of r_1 are (0.84, 0.71, 0.60, 0.5), (0.91, 0.84, 0.77, 0.7), and (0.97, 0.95, 0.92, 0.9). We simulate data by considering the mean in the control group as (3.0, 2.8, 2.6, 2.4, 2.0). The true treatment effect is (0, 0.13, 0.17, 0.19, 0.25) if under the alternative, or (0, 0, 0, 0, 0) if under the null. A number of patients drop-out at each longitudinal visit and the corresponding measurements are set to be missing given the true retention rate. Because the treatment difference at the last measurement is of interest in the simulation, the contrast vector is (0,0,0,1) excluding baseline. Cases with greater treatment effects would tend to stop early due to crossing the efficacy bound.

Figure 4.2 and Figure 4.3 (left) show the power curves under 18 different combinations of true nuisance parameters. Each circle corresponds to one of the 18 scenarios while x axis is for the 3 different retention rates, different line color stands for the 3 correlation structures, and the line type denotes different standard deviations. The assumed nuisance parameters in both figures are the same. The dot with a cross symbol denotes that the nuisance parameters share the same values in design and in true data, whereas other 17 circles employed various true values of R, S and r_1. As is visible in Figure 4.2, the power using our approach is well maintained around 90%, while the power under fixed design is not satisfactory under some scenarios where the nuisance parameters assumption deviates much from the true values. This is also seen in left plot in Figure 4.3 for group sequential design without sample size re-estimation. When checking the expected sample size in the right plot of Figure 4.3, we noticed that for about half of the cases for which we did not assume nuisance parameters accurately enough, it requires more patients for our method in group sequential design than that in fixed design.

On the other hand, under the null, Figure 4.4 and Figure 4.5 (left) show that all three methods can control the type-I error. Furthermore, the expected sample size in

Figure 4.5 (right) is similar to what we have under the alternative. Table 4.1 provides the standard error for the power and type I error based on 1,000 simulations. Symmetric and asymmetric two-sided tests were both examined as well to assess the performance, and they turn out to have very similar results to that for one-sided test, hence they are omitted here. All the simulations are based on the second method in section 4.3.1, although it is noticed that all the results look very similar to when we instead use the first method (i.e. information fraction method as in (4.9)) with completed and drop-out data only (results are omitted). The reason is that there is not much information gained by having around 10 more ongoing patients at each interim given the assumed slow enrollment.

The correlation coefficient in the above simulaitons was set as 0.579. To check performance when varying the correlation coefficient, we keep the design parameters (R_0, S_0 and r_{10}) the same as above, let the true nuisance parameters $S = S_0$ and $r_1 = r_{10}$, but vary the true correlation structure using a correlation coefficient 0.3 or 0.8 under compound symmetry, Toplitz and AR(1). In Table 4.2, targeted power is observed under our re-estimation approach with various correlation coefficients 0.3 and 0.8. In contrast, the fixed design and the group sequential design without sample size re-estimation laed to power ranging from 0.77 to 1, and from 0.81 to 0.98 respectively. As indicated in Table 4.3, there is no significant problem of controlling type I error due to the simulation error for the three methods. The corresponding expected sample size are (422, 201, 459, 275, 468, 373) for the 6 scenarios given $N_{fix} = 392$ and the original $N_{max} = 398$.

Lastly, since the simulation is designed to detect a treatment difference at the last (4^{th} post baseline) measurement, it would be interesting to implement the fixed design and the group sequential design with and without sample size re-determination to analyze only the last time point as normal data using method by Mehta and Tsiatis

Mehta and Tsiatis (2001). Keeping the correlation coefficient at 0.579 and following the same simulation setup, as expected, in Figure 4.6- and 4.7(left), different correlation structures do not make a difference for all three designs since we only use the last measurement for all patients. It also makes sense that the information-based method by Mehta and Tsiatis Mehta and Tsiatis (2001) is able to maintain the power when the retention rate is not too low. Similar feature is observed for the expected sample size as shown in Figure 4.7(right). Figure 4.8 and 4.9 are the corresponding type I error and expected sample size under the null.

4.5 Example

We build functions in R to formalize our method and to perform a data analysis. Our data is motivated by clinical trials studying change in tumor size over time. Before analyzing data, we need to know the sample size assignment for each interim by designing α, β, δ, one-sided test or two-sided test, number of planned looks, planned information fraction at each look, and nuisance parameters assumption(R_0, S_0, r_{10}). We wish to detect 0.25 treatment difference for the last repeated measurements between two groups of patients by designing a study planning one interim look and one final analysis using a one-sided test with type-I error 0.025, power 90%, and with a planned information fraction of $(0.5, 1)$ for clarity and simplicity. Each patient is expected to have 4 visits to the clinic to get the tumor measured. A monotone missing pattern is assumed for this study. The nuisance parameters assumed here are

$$R_0 = \begin{pmatrix} 1.000 & 0.579 & 0.579 & 0.579 & 0.579 \\ 0.579 & 1.000 & 0.579 & 0.579 & 0.579 \\ 0.579 & 0.579 & 1.000 & 0.579 & 0.579 \\ 0.579 & 0.579 & 0.579 & 1.000 & 0.579 \\ 0.579 & 0.579 & 0.579 & 0.579 & 1.000 \end{pmatrix},$$

$$S_0 = (0.925, 0.925, 0.925, 0.925, 0.925), r_{10} = (0.950.90, 0.85, 0.80),$$

The design and analysis are presented using our proposed information fraction approach in (4.9) due to the ease of implementation and illustration. The R program for the other re-estimation method is also available upon request. The corresponding N_{max} is then 466. The fixed design sample size for this longitudinal study calculated using the strategy in section 4.3.1 is also 466 because in this case the inflation factor in (4.7) is nearly 1. Hence, 233 patients (= 466/2) including completed and dropout shall be collected before analyzing the first interim analysis under group sequential design. A sample of typical clinical data is shown in Table 4.4.

The third column is the time at which each patient is enrolled. The column 'week' is recording the number of the longitudinal visits per person with '$0'$ denoting baseline and $1-4$ denoting post baseline visits. The column 'y' is the response of interest and 'flag' distinguishes patients who are still continuing (1) or not (0). It is noticed that the second patient does not have all four post baseline measurements but he/she is not continuing, implying that this person dropped out at the last visit.

Once the 1^{st} interim data has been collected ($N_{max} \times 1/2$), plugging in all the design parameters including α, β, δ, one-sided test or two-sided test, number of planned looks, planned information fraction at each look, and nuisance parameters assumption(R_0, S_0, r_{10}) as well as available data, the R function employing the strategy introduced earlier generates the result in Table 4.5 (second section):

The first section of the table displays the parameters we defined at the design stage. The rows 'Planned.timing1' and 'Planned.timing2' are the planned information fraction at first and final interim. The number of planned looks is clearly the number of rows for these variables (2 in this case). The second section of the table suggests that we should continue to do a 2^{nd} analysis since neither efficacy nor futility has been detected and that next analysis should be our final analysis according to the adaptation rule. The rows 'update.act.t1' and 'update.act.t2' are the actual information fraction as updated in (4.10), 'orig.Nmax' and 'new.Nmax' are N_{max} calculated from the design and re-estimated at the first interim respectively. The decrease of maximum sample size is because the anticipated information fraction (0.5) is smaller than the actual 0.514 which is produced in the bottom of the section. The underlying reason is that the nuisance parameter assumptions deviate from the truth, and as a result, the estimated covariance or information estimated by the real data does not agree with that using originally assumed nuisance parameters. Next, we collect another 178 ($= 411 - 466 \times \frac{1}{2}$) patients and analyze the final look. At the 2^{nd} look, we need to add the actual information fraction vector $(0.514, 1)$ and the new N_{max} (411) into our function. The result in the bottom section of Table 4.5 shows that the study can be stopped for efficacy and current information fraction is 1.01. The function and its help files are available upon request.

4.6 Discussion

We presented two information-based group sequential sample size re-estimation methods that can be for longitudinal trials which adapts appropriately depending on the true value of unknown nuisance parameters. Whereas previous work by Shih and Gould (1995) andZucker and Denne (2002) only evaluate a single interim and no hypothesis testing is performed until the final analysis; our approach has advantages of early termination and multiple interim looks. The simulation results confirm the method maintains power while controlling the type-I error, while a fixed design or a group sequential design without adjusting for nuisance parameters cannot. The reason is that in some cases where we do not have good historical evidence of the nuisance parameters, we have the ability of correcting it during the interim. In addition, a smaller sample is expected when the assumption is reasonably accurate, however, poor assumption requires more patients to maintain the statistical power. In conclusion, our method will help to both limit investment in treatments that do not work and ensure an appropriate investment to power trials for drugs that do work. For drugs that provide more than a minimally interesting treatment effect, the group sequential efficacy bounds provide a method to bring very effective drugs to market quickly.

We assume equally-spaced information-based design and equal retention rate for control and active treatment group just for simplicity in explanation. It is, however, fairly easy to extend it to a general case. To perform a real data analysis by our methods, we have built functions to calculate the necessary sample size before starting the trial enrollment and to re-estimate the sample size with testing if stopping early at the same time. Although this is being done in an unblinded fashion, our method can certainly be used to re-calculate the sample size and testing as long as the estimated parameter of interest and its variance are provided from the third party. Moreover, our methods work well for a small sample size as long as there are sufficient data to be analyzed in the random effect model at each interim look. However, given the complexity of the problem, it would be difficult to back-calculate the interim treatment effect based on the sample size adaptation as can be done in cases that are simpler than longitudinal data analysis.

Our methods presume that all subjects have measurement at their pre-defined measurement times. It is possible to introduce bias at the interim analyses if measurements occur at times other than the pre-defined follow-up times. Methods such as using a piece-wise linear approximation proposed by Kittelson et al. (2005) may be incorporated for the future work to handle departure from the protocol-defined measurement times. Another extension of our work could be to loosen the assumption of monotone missingness to missing at random. Moreover, subjects who are still under active follow-up may be different from those who drop out. Methods for evaluation of sensitivity to informative dropouts is another potential topic. Gao et al. (2013), Emerson and Fleming (1990), Kim (1989) introduced methods for unbiased estimation following sequential testing, and these methods could be incorporated when reporting results from any group sequential trial.

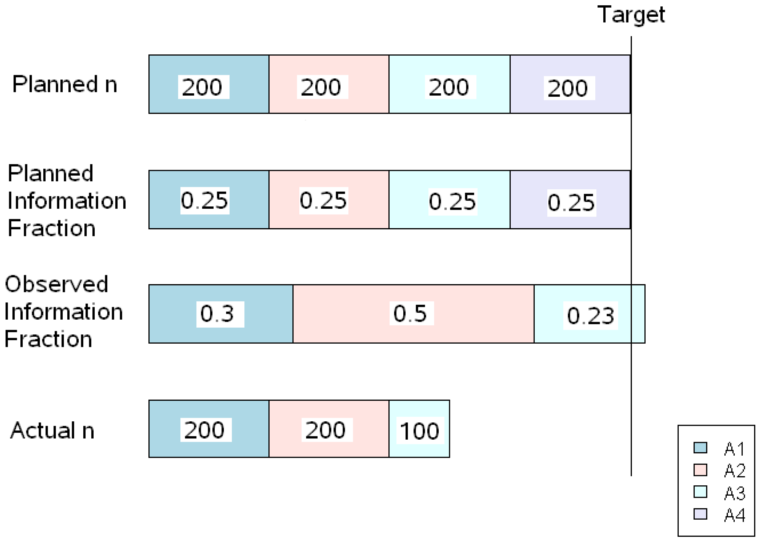

Figure 4.1: Adaptation Example

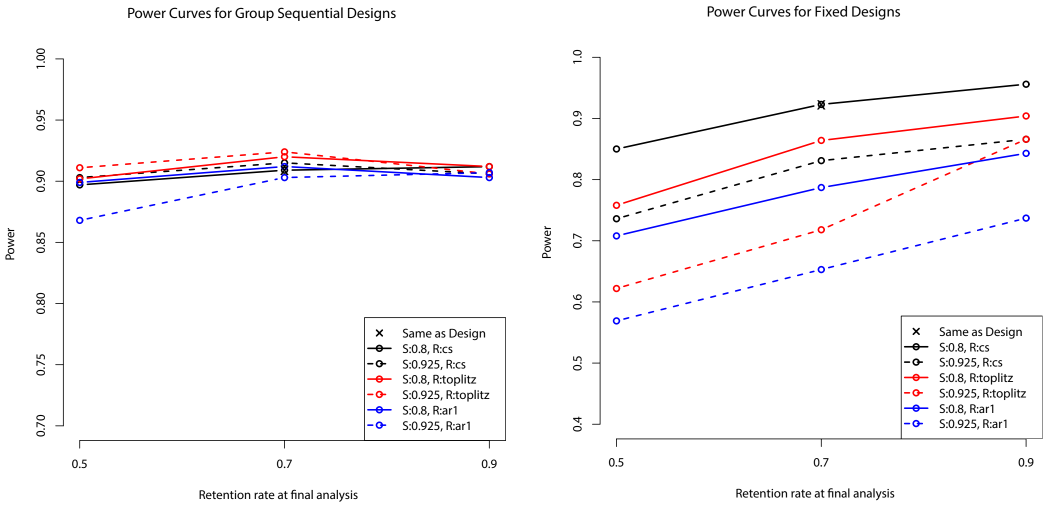

Figure 4.2: Power Curves Using Our Re-estimation Method (Left) v.s. Fixed Design (Right)

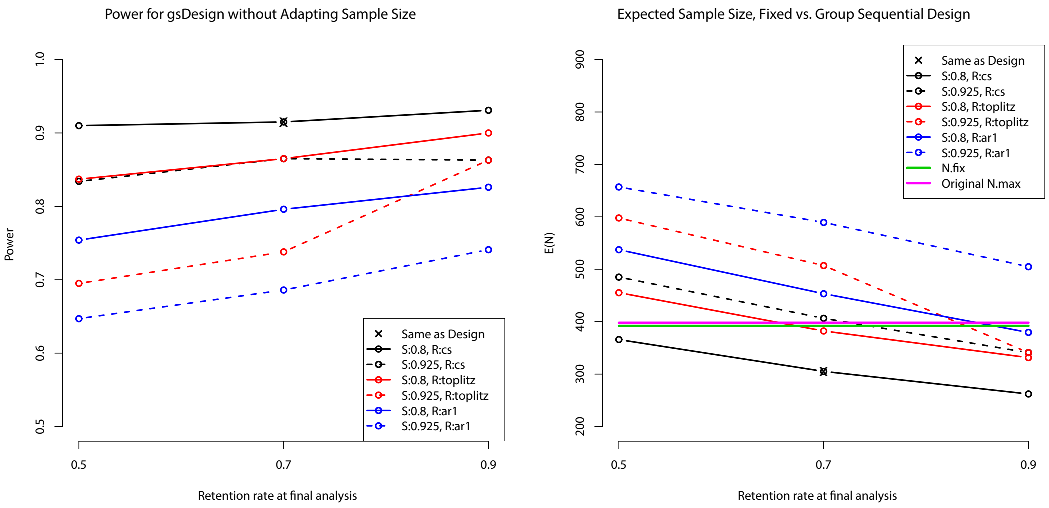

Figure 4.3: Left: Power Curves Using Group-sequential Design without Re-estimation; Right: Expected Sample Size Using Group Sequential Design with Sample Size Re-estimation v.s. Fixed Design

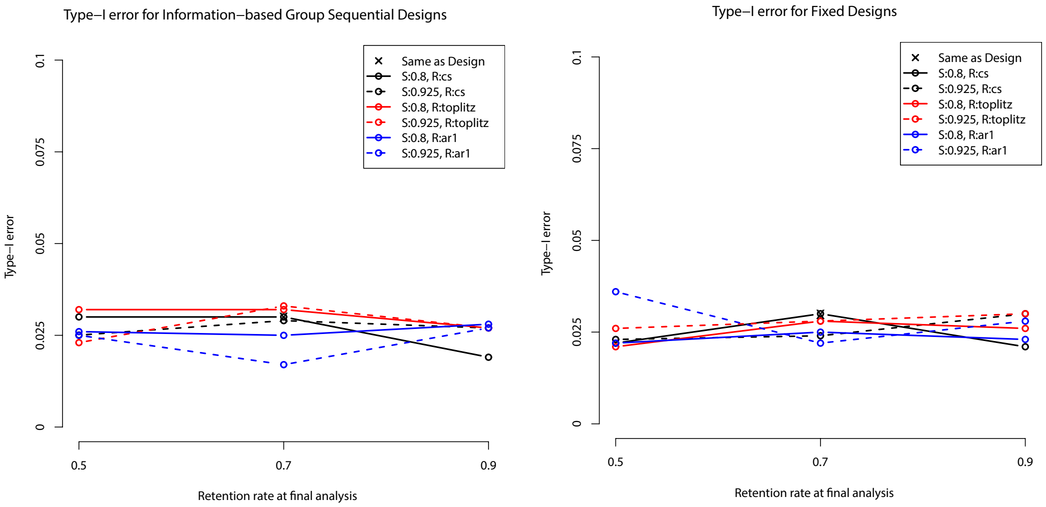

Figure 4.4: Type-I Error Using Our Re-estimation Method (Left) v.s. Fixed Design (Right)

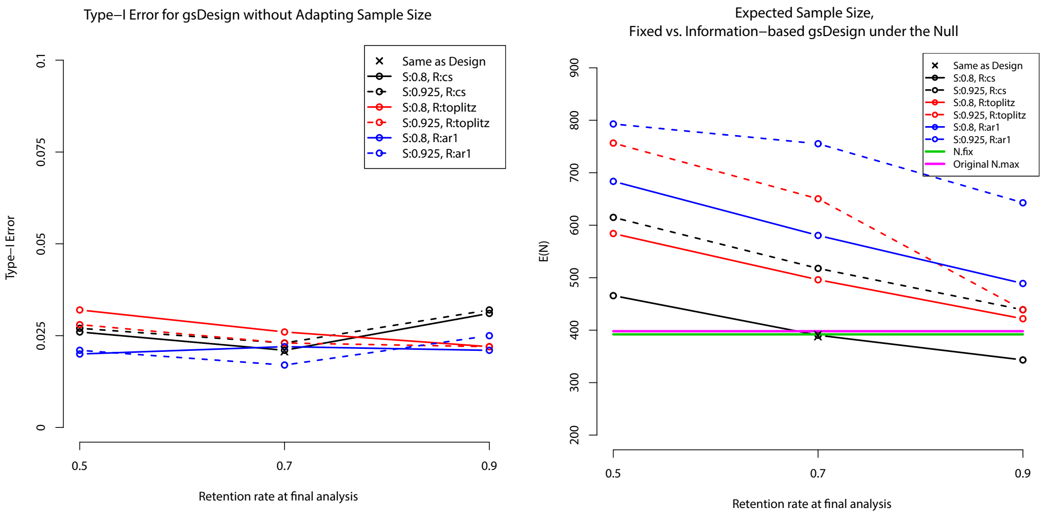

Figure 4.5: Left: Type-I Error Using Group-sequential Design without Re-estimation; Right: Expected Sample Size Using Group Sequential Design with Sample Size Re-estimation v.s. Fixed Design

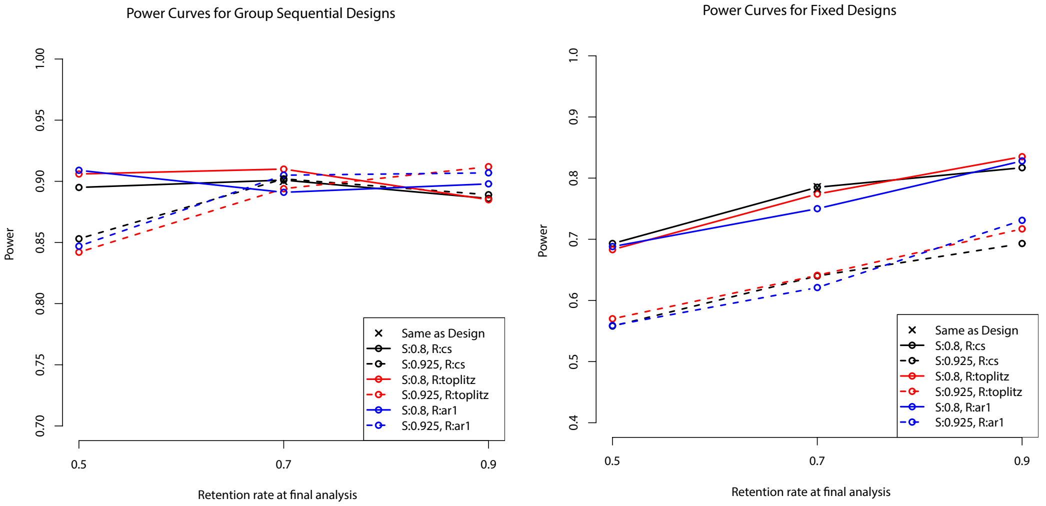

Figure 4.6: Simulations Results Based on the Last Time Point Only Using the Method by Mehta and Tsiatis (2001)

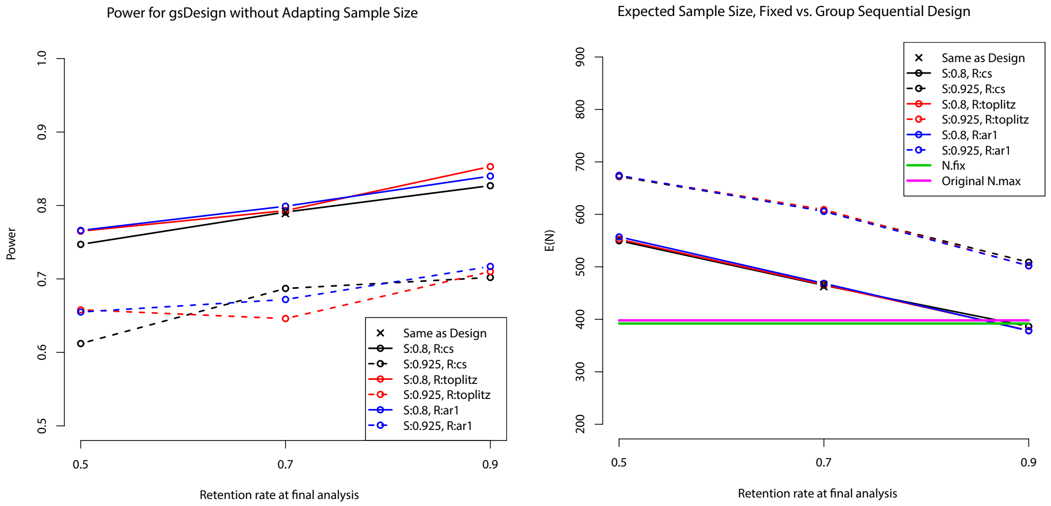

Figure 4.7: Simulation Results Based on the Last Time Point Only Using the Method by Mehta and Tsiatis (2001) (continued)

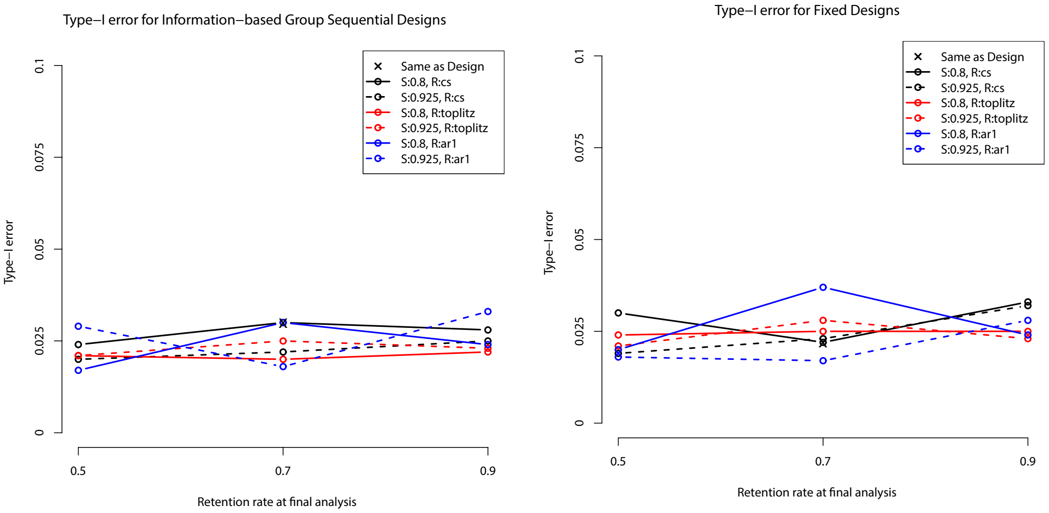

Figure 4.8: Simulation Results Based on the Last Time Point Only Using the Method by Mehta and Tsiatis (2001) (continued)

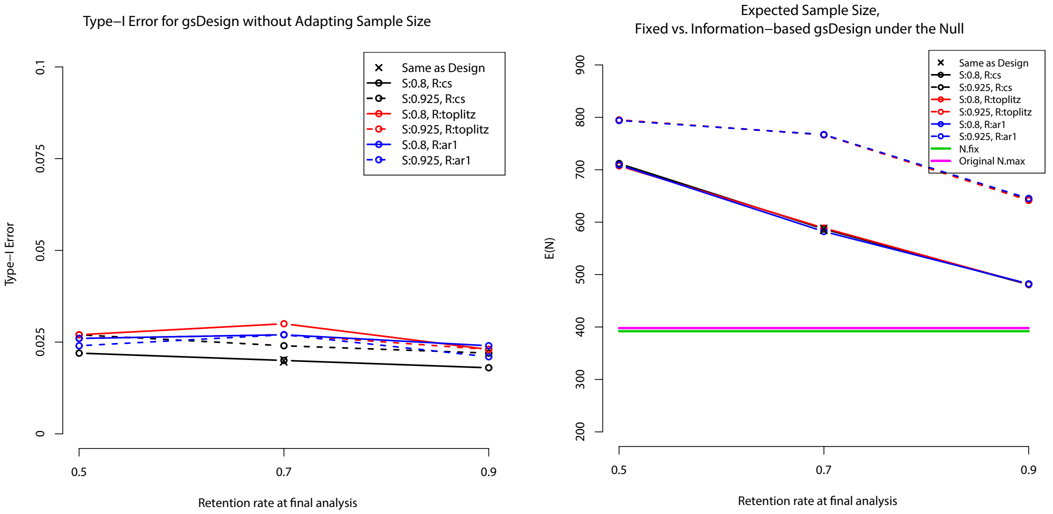

Figure 4.9: Simulation Results Based on the Last Time Point Only Using the Method by Mehta and Tsiatis (2001) (continued)

Table 4.1: Simulation Error

	S.E. of Power			**S.E. of Type-I Error**		
	Retention Rate at Final Interim			Retention Rate at Final Interim		
	0.5	0.7	0.9	0.5	0.7	0.9
R=cs, S=0.8	0.010	0.009	0.009	0.005	0.005	0.004
R=cs, S=0.925	0.009	0.009	0.009	0.005	0.005	0.005
R= toplitz, S=0.8	0.009	0.009	0.009	0.006	0.006	0.005
R=toplitz, S=0.925	0.009	0.008	0.009	0.005	0.006	0.005
R=ar1, S=0.8	0.010	0.009	0.009	0.005	0.005	0.005
R=ar1, S=0.925	0.011	0.009	0.009	0.005	0.004	0.005

Table 4.2: Power Based on Varying Correlation Coefficients

	Compound Symmetry		Toplitz		AR(1)	
	0.3	0.8	0.3	0.8	0.3	0.8
Group Sequential Design	0.9	0.93	0.92	0.91	0.91	0.9
Fixed Design	0.8	1	0.8	0.94	0.77	0.84
gsDesign without Adapting Sample Size	0.82	0.98	0.83	0.95	0.81	0.87

Table 4.3: Type I Error Based on Varying Correlation Coefficients

	Compound Symmetry		Toplitz		AR(1)	
	0.3	0.8	0.3	0.8	0.3	0.8
Group Sequential Design	0.025	0.039	0.023	0.029	0.028	0.017
Fixed Design	0.021	0.033	0.023	0.027	0.02	0.02
gsDesign without Adapting Sample Size	0.025	0.024	0.026	0.041	0.02	0.03

Table 4.4: Longitudinal Data with Four visits

subject	treatment	enrollment time	week	y	flag
1	1	0.0045	0	3.93	0
1	1	0.0045	1	2.95	0
1	1	0.0045	2	3.60	0
1	1	0.0045	3	2.24	0
1	1	0.0045	4	2.17	0
2	2	0.0079	0	3.53	0
2	2	0.0079	1	3.83	0
2	2	0.0079	2	2.63	0
2	2	0.0079	3	1.76	0
3	1	0.0927	0	2.01	1
3	1	0.0927	1	3.22	1
⋮	⋮	⋮	⋮	⋮	⋮

Table 4.5: Interim Analysis Results

alpha	0.025
power	0.9
Planned.timing1	0.5
Planned.timing2	1
delta	0.25
First Interim Analysis	
ifStopTrial	Continue to the next interim
ifNextFinal	TRUE
update.act.t1	0.514
update.act.t2	1
orig.Nmax	466
new.Nmax	411
current.info	0.514
Final Interim Analysis	
ifStopTrial	Stop with sig. Efficacy
ifNextFinal	FALSE
update.act.t1	0.514
update.act.t2	1
orig.Nmax	466
new.Nmax	411
current.info	1.01

APPENDIX I

List of Significant Associations - Birth Defects Study

1. Any heart defects

   Risk effects: BMI 30+, mother black, father black, Small for gestational age, parity, Type 1 diabetes - diagnosis at any time, Type 2 diabetes - diagnosis at any time, high blood pressure, Antihypertensive Exposure b1-p3, Anti-infective Exposure b1-p3, SSRI Exposure b1-p3, Sulfamethoxazole Exposure b1-p3, Trimethoprim Exposure b1-p3, Mother Birth In USA, Father Birth In USA, Pregnant High Blood Pressure, Had Epilepsy, Prenatal Vitamins-3toDOIB, HPV B1-P3, Mother Health Problem Or Birth Defects, Father Health Problem Or Birth Defects, Health Problems To Related Family Members, Arkansas center, Texas center, gestational age<32 weeks, gestational age 32-36 weeks, Birth Weight < 1500 grams, Birth Weight 1500-2500 grams, mother Office and Administrative Support Occupations, father Installation, Maintenance, and Repair Occupations, father Military Specific Occupations, xylene B1-P3, stoddard B1-P3, any aromatic class B1-P3, any solvent B1-P3

   Protective effects: mother's age < 18, mother's age 18-24, father's age 18-24, baby's gender, father other race, mother income <10,000, California center, New Jersey center, New York Center, North Carolina center, Fetal Death >= 20 Wks (Stillbirth), Language of interview Spanish, Interpregnancy Interval 18-23, (father Life, Physical, and Social Science Occupations)

2. Conotruncal defects

   Risk effects: mother's age 30-34, baby's gender, mother other race, father other race, mother income 50,000+, Small for gestational age, abortion, Health Problems To Related Family Members, California center, Massachusetts center, mother Architecture and Engineering Occupations, (mother Life, Physical, and Social Science Occupations), mother Building and Grounds Cleaning and Maintenance Occupations, father Management Occupations, benzene B1-P3

Protective effects: mother's age 18-24, father's age 18-24, mother smoking, mother income <10,000, Mother Birth In USA, Arkansas center, Utah center, gestational age 32-36 weeks, father Sales and Related Occupations, father Production Occupations

3. Left ventricular outflow tract obstruction
   Risk effects: mother's age 30-34, baby's gender, parity, preganancy nausea, Doxylamine Exposure b1-p3, Meclizine Exposure b1-p3, Opoids Exposure b1-p3, Sulfamethoxazole Exposure b1-p3, Trimethoprim Exposure b1-p3, Thyroid/Antithyroid Exposure b1-p3, Antipyretic Exposure b1-p3, Mother Birth In USA, Father Birth In USA, Thyroid Disease B1-P3, HIV/AIDS, Mother Health Problem Or Birth Defects, Health Problems To Related Family Members, Iowa center, Utah center, Total Caffeine<10, mother Legal Occupations, father Management Occupations, father Business and Financial Operations Occupations, father Sales and Related Occupations, father Installation, Maintenance, and Repair Occupations, benzene B1-P3

   Protective effects: mother's age $<$ 18, mother's age 18-24, father's age 18-24, mother black, mother hispanic, father black, father hispanic, mother income <10,000, Household Smoking - B1P3, New Jersey center, gestational age<32 weeks, Birth Weight $<$ 1500 grams, Total Caffeine 200-300, Interpregnancy Interval NA

4. Right ventricular outflow tract obstruction

   Risk effects: BMI 25-30, BMI 30+, mother black, father black, parity, stillbirths, DrinkSoftDrinks, Antihypertensive Exposure b1-p3, Opoids Exposure b1-p3, Paxil Exposure b1-p3, Promethazine Exposure b1-p3 , SSRI Exposure b1-p3, Mother Birth In USA, Household Smoking - B1P3, Pregnant High Blood Pressure, HIV/AIDS, HPV B1-P3, Mother Health Problem Or Birth Defects, Father Health Problem Or Birth Defects, Health Problems To Related Family Members, Arkansas center, Iowa center, Calculated birth date in summer, gestational age<32 weeks, Birth Weight < 1500 grams, mother Office and Administrative Support Occupations, (mother Installation, Maintenance, and Repair Occupations), mother Military Specific Occupations, father Architecture and Engineering Occupations, xylene B1-P3, carbon tetrachloride B1-P3, perchloroethylene B1-P3, trichloroethane B1-P3, trichloroethylene B1-P3, stoddard B1-P3, any aromatic class B1-P3

   Protective effects: baby's gender, mother hispanic, father hispanic , California center, North Carolina center, Language of interview Spanish

5. Ventricular septal defect perimembranous

   Risk effects: mother's age 35+, father's age 35+, mother black, father black, tubal pregnancy, Pelvic inflamatory disease B1P3, Type 2 diabetes - diagnosis at any time, Mother Fertility procedures, used mother fertility procedure, Antihypertensive Exposure b1-p3, (Sexually Transmitted Disease B1-P3 - Includes pelvic inflamatory disease, but not HIV/AIDS or HPV HIV/AIDS HPV B1-P3), pelvic inflamatory disease B1P3, Father Health Problem Or Birth Defects, Arkansas center, Birth Weight 1500-2500 grams, mother Management Occupations, mother Computer and Mathematical Occupations, (mother Life, Physical, and Social Science Occupations), mother Legal Occupations, mother Sales and Related Occupations, mother Production Occupations, mother Military Specific Occupations, (father Life, Physical, and Social Science Occupations), father Healthcare Practitioners and Technical Occupations, father Sales and Related Occupations, benzene B1-P3, chloroform B1-P3, any aromatic class B1-P3

Protective effects: baby's gender, father hispanic, California center , North Carolina center

6. Atrial septal defect (ASD) secundum or ASD not otherwise specified

   Risk effects: mother's age 18-24, BMI 30+, mother low education, father low education, mother smoking, mother hispanic, father black, father hispanic, mother income <10,000, stillbirths, molar pregnancy, Kidney/Bladder/UTI B1P3, Type 1 diabetes - diagnosis at any time, Type 2 diabetes - diagnosis at any time, Gestational diabetes - diagnosis before or during index pregnancy or unknown date, Gestational diabetes - diagnosis during index pregnancy, high blood pressure, Antihypertensive Exposure b1-p3, Anti-infective Exposure b1-p3, Promethazine Exposure b1-p3, Sulfamethoxazole Exposure b1-p3, Trimethoprim Exposure b1-p3, Mother Birth In USA, Father Birth In USA, Household Smoking - B1P3, Pregnant High Blood Pressure, HPV B1-P3, Kidney/Bladder/UTI B1P3, Mother Health Problem Or Birth Defects, Arkansas center, Texas center, gestational age<32 weeks, gestational age 32-36 weeks, Birth Weight < 1500 grams, Birth Weight 1500-2500 grams, (mother Installation, Maintenance, and Repair Occupations), father Production Occupations,father Military Specific Occupations

   Protective effects: mother's age 30-34, father's age 35+ , baby's gender, drinking, drinking but not binge, mother income 50,000+, abortion, California center, Iowa center, Massachusetts center, New Jersey center, New York Center, mother Food Preparation and Serving Related Occupations, father Management Occupations, father Education, Training, and Library Occupations, father Sales and Related

Occupations

7. Entricular and atrial septal defect

   Risk effects: mother's age 30-34, father's age 35+, BMI 30+, binge drinking, Small for gestational age, molar pregnancy, Pelvic inflamatory disease B1P3, Type 1 diabetes - diagnosis at any time, Type 2 diabetes - diagnosis at any time, high blood pressure, Mother Fertility procedures, used mother fertility procedure, Antihypertensive Exposure b1-p3, Meclizine Exposure b1-p3, Trimethoprim Exposure b1-p3, HIV/AIDS, pelvic inflamatory disease B1P3, Fever B1P3, Massachusetts center, Texas center, Induced Abortion, Language of interview Spanish, gestational age 32-36 weeks, Birth Weight 1500-2500 grams, Total Caffeine 10-100, (mother Life, Physical, and Social Science Occupations), mother Community and Social Services Occupations, mother Healthcare Practitioners and Technical Occupations, mother Protective Service Occupations , mother Construction and Extraction Occupations, (mother Installation, Maintenance, and Repair Occupations), mother Military Specific Occupations , (father Education, Training, and Library Occupations), (father Installation, Maintenance, and Repair Occupations), father Military Specific Occupations, benzene B1-P3, carbon tetrachloride B1-P3, perchloroethylene B1-P3, trichloroethylene B1-P3

   Protective effects: baby's gender , New Jersey center New York Center

8. Neural tube defects

   Risk effects: BMI 30+, mother hispanic, father hispanic, mother income <10,000, Small for gestational age, parity, Anticonvulsants Exposure b1-p3, Doxylamine Exposure b1-p3, Meclizine Exposure b1-p3, Fever B1P3, Health Problems To Related Family Members, California center, Iowa center, Fetal Death >= 20 Wks (Stillbirth), Induced Abortion, Language of interview Spanish, gestational age<32 weeks, gestational age 32-36 weeks, Birth Weight < 1500 grams, Birth Weight 1500-2500 grams, mother Military Specific Occupations, father Computer and Mathematical Occupations, father Community and Social Services Occupations, father Healthcare Support Occupations, father Personal Care and Service Occupations, (father Farming, Fishing, and Forestry Occupations), methylene chloride B1-P3, trichloroethylene B1-P3

Protective effects: BMI < 18.5, mother smoking, baby's gender, drinking, drinking but not binge, mother income 50,000+, Mother Birth In USA, Father Birth In USA, Pregnant High Blood Pressure, Massachusetts center, New Jersey center, Birth Weight >4000 grams, Interpregnancy Interval NA, mother Management Occupations, mother Education, Training, and Library Occupations, father Management Occupations, father Business and Financial Operations Occupations, father Protective Service Occupations

9. Anotia/microtia

   Risk effects: mother's age 18-24, father low education, mother hispanic, mother other race, father hispanic, mother income <10,000, molar pregnancy, miscarriage, Type 1 diabetes - diagnosis at any time, Meclizine Exposure b1-p3, HIV/AIDS, HPV B1-P3, California center, New Jersey center, Texas center, Induced Abortion, Language of interview Spanish, mother Building and Grounds Cleaning and Maintenance Occupations, (mother Farming, Fishing, and Forestry Occupations), mother Military Specific Occupations, father Building and Grounds Cleaning and Maintenance Occupations, (father Farming, Fishing, and Forestry Occupations)

   Protective effects: mother black, father black, mother income 50,000+, Antipyretic Exposure b1-p3, Mother Birth In USA, Father Birth In USA, Arkansas center,Iowa center, Massachusetts center, mother Business and Financial Operations Occupations

10. Cleft palate

    Risk effects: father's age 35+, mother smoking, drinking, drinking but not binge, mother other race, Pelvic inflamatory disease B1P3, Gestational diabetes - diagnosis before or during index pregnancy or unknown date, Gestational diabetes - diagnosis during index pregnancy, used fertility Meds Procedure, use fertility meds, used mother fertility procedure, Acetaminophen Exposure b1-p3, Anticonvulsants Exposure b1-p3, Meclizine Exposure b1-p3, NSAIDS Exposure b1-p3, Antipyretic Exposure b1-p3, Household Smoking - B1P3, Had Epilepsy, HIV/AIDS, pelvic inflamatory disease B1P3, Mother Health Problem Or Birth Defects, Health Problems To Related Family Members, Massachusetts center, Calculated date of conception in winter, mother Community and Social Services Occupations, mother Legal Occupations, (mother Arts, Design, Entertainment, Sports, and Media Occupations), mother Building and Grounds Cleaning and Maintenance Occupations, (mother Installation, Maintenance, and Repair Occupations), mother Military Specific Occupations, carbon tetrachloride B1-P3, stoddard B1-P3

    Protective effects: mother's age < 18, baby's gender, mother black, mother hispanic, father black, father hispanic, mother income <10,000, Folate, DFE >=683.316

11. Cleft lip with cleft palate

    Risk effects: mother's age 18-24, father's age 18-24, BMI < 18.5, mother low education, father low education, mother smoking, baby's gender, mother hispanic, father hispanic, mother income <10,000, DrinkSoftDrinks, Cold Meds Exposure b1-p3, NSAIDS Exposure b1-p3, (Sexually Transmitted Disease B1-P3 - Includes pelvic inflamatory disease, but not HIV/AIDS or HPV), Health Problems To Related Family Members, California center, Calculated birth date in spring, Fetal

Death >= 20 Wks (Stillbirth), Language of interview Spanish, mother Personal Care and Service Occupations, (father Farming, Fishing, and Forestry Occupations), father Construction and Extraction Occupations

Protective effects: mother's age 30-34, father's age 30-34, mother black, father black, mother income 50,000+, Acetaminophen Exposure b1-p3, Thyroid/Antithyroid Exposure b1-p3, Mother Birth In USA, Prenatal Vitamins-3toDOIB, Massachusetts center, New Jersey center, Calculated birth date in winter, gestational age 32-36 weeks, Birth Weight 1500-2500 grams, father Personal Care and Service Occupations

12. Cleft lip without cleft palate

    Risk effects: mother smoking, baby's gender, molar pregnancy, NSAIDS Exposure b1-p3, Mother Birth In USA, Father Birth In USA, Pregnant High Blood Pressure, Mom Substance Abuse B3-DOIB, HIV/AIDS, Father Health Problem Or Birth Defects, Health Problems To Related Family Members, California center, Iowa center, mother Computer and Mathematical Occupations, mother Protective Service Occupations, mother Construction and Extraction Occupations, (mother Installation, Maintenance, and Repair Occupations), father Personal Care and Service Occupations, chloroform B1-P3

    Protective effects: mother black, mother hispanic, father black, father hispanic, Prenatal Vitamins-3toDOIB, Language of interview Spanish, gestational age<32 weeks, gestational age 32-36 weeks, Birth Weight < 1500 grams, Birth Weight 1500-2500 grams, (father Farming, Fishing, and Forestry Occupations)

13. Esophageal atresia

    Risk effects: mother's age 35+, father's age 35+, drinking, drinking but not binge, mother income 50,000+, Small for gestational age, used fertility Meds

Procedure, Mother Fertility procedures, use fertility meds, used mother fertility procedure, Antihypertensive Exposure b1-p3, Sulfamethoxazole Exposure b1-p3, Trimethoprim Exposure b1-p3, gestational age<32 weeks, gestational age 32-36 weeks, Birth Weight 1500-2500 grams, Interpregnancy Interval NA, mother Management Occupations, mother Business and Financial Operations Occupations, (mother Education, Training, and Library Occupations), (mother Installation, Maintenance, and Repair Occupations), mother Military Specific Occupations, father Legal Occupations, father Healthcare Practitioners and Technical Occupations

Protective effects: mother's age 18-24, mother low education, father low education, baby's gender, mother black, father black, father hispanic, parity, preganancy nausea, Mother Health Problem Or Birth Defects, Birth Weight >4000 grams, Interpregnancy Interval 12-17, father Transportation and Material Moving Occupations

14. Anorectal atresia/stenosis

    Risk effects: mother hispanic, father hispanic, Small for gestational age, Type 1 diabetes - diagnosis at any time, Type 2 diabetes - diagnosis at any time, used fertility Meds Procedure, use fertility meds, used mother fertility procedure, (Sexually Transmitted Disease B1-P3 - Includes pelvic inflamatory disease, but not HIV/AIDS or HPV), HIV/AIDS, pelvic inflamatory disease B1P3, New Jersey center, Fetal Death >= 20 Wks (Stillbirth) , Birth Weight 1500-2500 grams, Total Caffeine 300+, 320.264<= Folate, DFE <472.63, mother Architecture and Engineering Occupations, father Healthcare Practitioners and Technical Occupations

    Protective effects: Cold Meds Exposure b1-p3, Prenatal Vitamins-3toDOIB, Mom Substance Abuse B3-DOIB, Iowa center, mother Personal Care and Service Occupations

15. Hypospadias second/third degree

    Risk effects: mother's age 30-34, mother's age 35+, father's age 30-34, father's age 35+, baby's gender, drinking, drinking but not binge, mother income 50,000+, Small for gestational age, miscarriage, high blood pressure, used fertility Meds Procedure, Mother Fertility procedures, use fertility meds, used mother fertility procedure, Mother Birth In USA, Father Birth In USA, Pregnant High Blood Pressure, HIV/AIDS, Father Health Problem Or Birth Defects, Massachusetts center, New Jersey center, gestational age<32 weeks, gestational age 32-36 weeks, Birth Weight < 1500 grams, Birth Weight 1500-2500 grams, Birth Weight >4000 grams, Interpregnancy Interval NA, mother Management Occupations, mother Business and Financial Operations Occupations, mother Computer and Mathematical Occupations, mother Architecture and Engineering Occupations, (mother Life, Physical, and Social Science Occupations), (mother Arts, Design, Entertainment, Sports, and Media Occupations), mother Healthcare Practitioners and Technical Occupations, mother Protective Service Occupations, father Business and Financial Operations Occupations , father Computer and Mathematical Occupations, father Architecture and Engineering Occupations, (father Arts, Design, Entertainment, Sports, and Media Occupations), father Healthcare Practitioners and Technical Occupations, father Office and Administrative Support Occupations, carbon tetrachloride B1-P3

    Protective effects: mother's age < 18, mother's age 18-24, father's age 18-24, mother low education, father low education, mother smoking, binge drinking, mother hispanic, father hispanic, mother income <10,000, parity , preganancy nausea, DrinkSoftDrinks, Household Smoking - B1P3, Dad Substance Abuse B3-DOIB, Fever B1P3, California center, Iowa center, Texas center, Language of interview Spanish, Total Caffeine 10-100, (Folate, DFE)>=683.316, (mother Farming, Fishing, and Forestry Occupations), (father Farming, Fishing, and Forestry Occupations), father Construction and Extraction Occupations, father Production Occupations

16. Limb deficiency

    Risk effects: mother hispanic, father hispanic, Small for gestational age, Type 2 diabetes - diagnosis at any time, DrinkSoftDrinks, Anticonvulsants Exposure b1-p3, Anti-infective Exposure b1-p3, Opoids Exposure b1-p3, Sulfamethoxazole Exposure b1-p3, Trimethoprim Exposure b1-p3, HIV/AIDS, Fever B1P3, California center, gestational age<32 weeks, Birth Weight < 1500 grams, Birth Weight 1500-2500 grams, mother Healthcare Support Occupations, (mother Farming, Fishing, and Forestry Occupations), mother Military Specific Occupations, father Protective Service Occupations, father Personal Care and Service Occupations, chloroform B1-P3, methylene chloride B1-P3, trichloroethane B1-P3, stoddard B1-P3, chlorinated class B1-P3, any solvent B1-P3

    Protective effects: Arkansas center, Birth Weight >4000 grams

17. Craniosynostosis

    Risk effects: mother's age 35+, father's age 35+, baby's gender, mother income 50,000+, parity, miscarriage, Pelvic inflamatory disease B1P3, Meclizine Exposure b1-p3, Paxil Exposure b1-p3, SSRI Exposure b1-p3, Thyroid/Antithyroid Exposure b1-p3, Mother Birth In USA, Father Birth In USA, Thyroid Disease B1-P3, HIV/AIDS, pelvic inflamatory disease B1P3, Arkansas center, Massachusetts center, North Carolina center, Utah center, Calculated birth date in spring, Induced Abortion, Birth Weight >4000 grams, Interpregnancy Interval <12, Interpregnancy Interval 18-23, mother Military Specific Occupations, father Management Occupations, father Healthcare Support Occupations, father Protective Service Occupations, father Military Specific Occupations, carbon tetrachloride B1-P3

Protective effects: mother's age 18-24, father's age 18-24, mother low education, father low education, mother black, mother hispanic, father black, father other race, mother income <10,000, Small for gestational age, FA Supplement Use Flag (B3-P1), Dad Substance Abuse B3-DOIB, California center, New Jersey center, gestational age 32-36 weeks, Birth Weight 1500-2500 grams, Interpregnancy Interval NA, father Production Occupations

18. Diaphragmatic hernia

    Risk effects: Small for gestational age, abortion, Kidney/Bladder/UTI B1P3, Meclizine Exposure b1-p3, Sulfamethoxazole Exposure b1-p3, Trimethoprim Exposure b1-p3, HIV/AIDS, Kidney/Bladder/UTI B1P3, Language of interview Spanish, mother Business and Financial Operations Occupations, mother Community and Social Services Occupations, mother Healthcare Practitioners and Technical Occupations, mother Protective Service Occupations, mother Military Specific Occupations, father Computer and Mathematical Occupations , father Community and Social Services Occupations, (father Installation, Maintenance, and Repair Occupations), benzene B1-P3 Protective effects: Mother Birth In USA, Mother Birth In USA, Father Health Problem Or Birth Defects

19. Gastroschisis

    Risk effects: mother's age < 18, mother's age 18-24, father's age <18, father's age 18-24, BMI < 18.5, mother low education, father low education, mother smoking, binge drinking, mother hispanic, mother other race, father hispanic, mother income <10,000, Small for gestational age, oral contraceptive use B1P3, Kidney/Bladder/UTI B1P3, DrinkSoftDrinks, NSAIDS Exposure b1-p3, Household Smoking - B1P3, Mom Substance Abuse B3-DOIB, Dad Substance Abuse B3-DOIB, Mom Substance Abuse B1-P3, (Sexually Transmitted Disease B1-P3 - Includes pelvic inflamatory disease, but not HIV/AIDS or HPV), HIV/AIDS, Kidney/Bladder/UTI B1P3, California center, Fetal Death >= 20 Wks (Stillbirth) , gestational age<32 weeks, gestational age 32-36 weeks, Birth Weight 1500-2500 grams, Interpregnancy Interval NA, (Folate, DFE) >=683.316, mother Food Preparation and Serving Related Occupations, mother Personal Care and Service Occupations, mother Sales and Related Occupations, father Food Preparation and Serving Related Occupations, father Building and Grounds Cleaning and Maintenance Occupations, father Construction and Extraction Occupations

Protective effects: mother's age 30-34, mother's age 35+, father's age 30-34, father's age 35+, BMI 25-30, BMI 30+, drinking but not binge, mother income 50,000+, parity, miscarriage, Type 2 diabetes - diagnosis at any time, Gestational diabetes - diagnosis before or during index pregnancy or unknown date, Gestational diabetes - diagnosis during index pregnancy, high blood pressure, used fertility Meds Procedure, use fertility meds, used mother fertility procedure, Acetaminophen Exposure b1-p3, Antipyretic Exposure b1-p3, Pregnant High Blood Pressure, Massachusetts center, Birth Weight >4000 grams, Interpregnancy Interval 18-23, 320.264<= (Folate, DFE) <472.63, mother Management Occupations, mother Business and Financial Operations Occupations , (mother Education, Training, and Library Occupations), mother Healthcare Practitioners and Technical Occupations, father Management Occupations, father Business and Financial

Operations Occupations, father Computer and Mathematical Occupations, father Architecture and Engineering Occupations, (father Arts, Design, Entertainment, Sports, and Media Occupations), father Protective Service Occupations, any solvent B1-P3

APPENDIX II

Computation Code for Chapter 2

```
% -- to find dependence in splice-junction data -- %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Goal: Sparse PARAFAC Tensor factorization model
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%N= 3175; p= 60; n = 2*p; indn = randsample(N,n);
%save('D:\Dissertation Research\30-Dunson&Xing Extension\DataApplication\
splice1_indn_2p','indn');

clear;clc;close all;

tic;
% -- define global parameters -- %
d = 4; k = 10;                  % d =  # categories, k = # factors
nrun = 25000; burn = 10000; thin = 5;
eff_samp1 = (nrun - burn)/thin;
prop_vars = 1; prop_obs = 1;    % prop. of vars/obs to use for fitting

% -- load data -- %
%load('D:\Dissertation Research\30-Dunson&Xing Extension\
DataApplication\splicedata.mat');
load('/home/users/jingzhou/Dissertation/30-DunsonXing_Extension/
DataApplication/splicedata.mat');
[N,p] = size(x); n = 2*p;
%load('D:\Dissertation Research\30-Dunson&Xing Extension\
DataApplication\splice1_indn.mat');
load('/home/users/jingzhou/Dissertation/30-DunsonXing_Extension/
DataApplication/splice1_indn_2p.mat');
Yn = x(indn,:);
rows = repmat(1:p,n,1); rows = rows(:); cols = Yn(:);
lin_idx = sub2ind([p,d],rows,cols);



% --  empirical marginal probabilities -- %
lam0 = repmat([1/d 1/d 1/d 1/d],p,1);    %p*d


% -- initialize parameters -- %
alpha = 1;               % dp hyperparameters for stick breaking weights
aal = 1; bal = 1;        % gamma hyperparameters for alpha
a = ones(d,1);           % dirichlet hyperparameter for \lambda_h^{(j)}
Lambda = zeros(k,p,d);
% loadings for preds-- initialize with sample cell frequencies
for h = 1:k
    for j = 1:p
    Lambda(h,j,:) = [sum(Yn(:,j)==1)/n sum(Yn(:,j)==2)/n ...
    sum(Yn(:,j)==3)/n sum(Yn(:,j)==4)/n];
    end
end
```

```
mat = zeros(p,d);   % for updating Lambda
mat0 = zeros(p,d);  %for updating from the null group

gamm=20;   %prior pi_h~Beta(1,gamm)
pi=betarnd(1,gamm,[k,1]);
indw0=zeros(k,p);   % for prob of degerating distribution

nus = betarnd(1,alpha,[k-1,1]); nu = zeros(k,1);
nu(1:k-1) = nus.*cumprod([1;1-nus(1:k-2)]); nu(k) = 1 - sum(nu(1:k-1));
z = zeros(n,1); zupdateprob = zeros(n,k); zcts = zeros(k,1);


% -- define output files -- %
alphaout = zeros(nrun,1); nuout = zeros(nrun,k);
nu_max = zeros(nrun,1); piout= zeros(nrun,k);
nu_comps_95 = zeros(nrun,1); %nu_comps_99 = zeros(nrun,1);

NMIout = zeros(eff_samp1,p*(p-1)/2); crvout = zeros(eff_samp1,p*(p-1)/2);
MI_postmean = zeros(p,p); NMI_postmean = zeros(p,p); crv_postmean = zeros(p,p);
NMI_post_lq = zeros(p,p); NMI_post_uq = zeros(p,p);
crv_post_lq = zeros(p,p); crv_post_uq = zeros(p,p);


    % -- start Gibbs sampler -- %
    for i = 1:nrun
```

```
% -- update z -- %
for h = 1:k
    Lambdah = reshape(Lambda(h,:,:),p,d);
    tmpmatL = reshape(Lambdah(lin_idx),[n,p]);
    zupdateprob(:,h) = nu(h) * prod(tmpmatL,2);
end
zupdateprob1 = bsxfun(@times,zupdateprob,1./(sum(zupdateprob,2)));
mat1 = [zeros(n,1) cumsum(zupdateprob1,2)];
rr = unifrnd(0,1,[n,1]);
for l = 1:k
    ind = rr > mat1(:,l) & rr <= mat1(:,l+1); z(ind) = l;
end


% -- update lambda -- %
for h = 1:k
    zh = (z==h); zcts(h) = sum(zh);
    for c = 1:d
        mat(:,c) = (a(c) + sum(bsxfun(@times,(Yn==c),zh)))’;
        mat0(:,c) = (sum(bsxfun(@times,(Yn==c),zh)))’;
    end
    Lamh1 = gamrnd(mat,1); Lamh = bsxfun(@times,Lamh1,1./sum(Lamh1,2));
    Lambda(h,:,:) = Lamh;

    tmpw0h=(1-pi(h))*prod(lam0.ˆmat0,2);    %p*1
    tmpw1h=pi(h)*gamma(sum(a))/prod(gamma(a))*prod(gamma(mat),2)./
    gamma(sum(a)+zcts(h));  %p*1
    w0h=tmpw0h./(tmpw0h+tmpw1h);
    indw0(h,:)=(rand(1,p)<w0h');
    w0row =find(indw0(h,:));
    if length(w0row)>0
        Lambda(h,w0row,:)=lam0(w0row,:);  %d=4
    end
end


% -- update pi -- %
pi = betarnd(1 + p- sum(indw0,2),gamm + sum(indw0,2));

% -- update nu -- %
for h = 1:k-1
    nus(h) = betarnd(1 + zcts(h),alpha + sum(zcts(h+1:k)));
    nu(1:k-1) = nus.*cumprod([1;1-nus(1:k-2)]);
    nu(k) = 1-sum(nu(1:k-1));
end


% -- update alpha-- %
nuss = 1 - nus(1:k-1); nuss(nuss < 1e-6) = 1e-6;
alpha = gamrnd(aal + k - 1, 1/(bal - sum(log(nuss))));
```

```
% -- first write files to be stored across each iteration -- %
nuout(i,:) = nu';nu_ord = sort(nu,'descend'); nu_max(i) = nu_ord(1);
nu_comps_95(i,:) = sum((cumsum(nu_ord) >=0.95) == 0) + 1;
%nu_comps_99(i,:) = sum((cumsum(nu_ord) >=0.99) == 0) + 1;
alphaout(i) = alpha;
piout(i,:)=pi';

% -- positional dependence -- %
if (mod(i,thin) == 0 && i > burn)
    % -- between variables (all with same d_j)-- %
    ct_loop = 0;
    for j1 = 1:p-1
        Lamj1 = reshape(Lambda(:,j1,:),k,d);
        pj1 = sum(bsxfun(@times,Lamj1,nu))'; Ij1 = - sum(pj1.*log(pj1));
        for j2 = j1+1:p
            ct_loop = ct_loop + 1;
            Lamj2 = reshape(Lambda(:,j2,:),k,d);
            pj2 = sum(bsxfun(@times,Lamj2,nu))';
            Ij2 = - sum(pj2.*log(pj2));
            pj1j2 = bsxfun(@times,Lamj1,sqrt(nu))'*
            bsxfun(@times,Lamj2,sqrt(nu));

            tmp_MI = sum(sum(pj1j2.*log(pj1j2./(pj1*pj2'))));
            tmp_NMI = tmp_MI/sqrt(Ij1*Ij2);
            crv = ((pj1j2 - pj1*pj2').^2)./(pj1*pj2');
```

```
                    tmp_crv = sqrt(sum(sum(crv/(d-1))));

                    if mod(i-burn,thin) == 0
                        NMIout((i-burn)/thin,ct_loop) = tmp_NMI;
                        crvout((i-burn)/thin,ct_loop) = tmp_crv;
                    end
                end
            end
        end

        %if mod(i,100) == 0, disp(i); end
    end

    % -- post processing -- %
    nu_comps_md_95 = mode(nu_comps_95(burn+1:end));
    nu_comps_lq_95 = quantile(nu_comps_95(burn+1:end),0.025);
    nu_comps_uq_95 = quantile(nu_comps_95(burn+1:end),0.975);

% -- for positional dependence -- %
nmi_lq = quantile(NMIout,0.025); nmi_uq = quantile(NMIout,0.975);
nmi_postmean = mean(NMIout);
crv_lq = quantile(crvout,0.025); crv_uq = quantile(crvout,0.975);
crv_postmean1 = mean(crvout);
ct_loop = 0;
for j1 = 1:p-1
    for j2 = j1+1:p
```

```
        ct_loop = ct_loop + 1;
        NMI_post_lq(j1,j2) = nmi_lq(ct_loop);
        NMI_post_lq(j2,j1) = NMI_post_lq(j1,j2);
        NMI_post_uq(j1,j2) = nmi_uq(ct_loop);
        NMI_post_uq(j2,j1) = NMI_post_uq(j1,j2);
        NMI_postmean(j1,j2) = nmi_postmean(ct_loop);
        NMI_postmean(j2,j1) = NMI_postmean(j1,j2);
        crv_post_lq(j1,j2) = crv_lq(ct_loop);
        crv_post_lq(j2,j1) = crv_post_lq(j1,j2);
        crv_post_uq(j1,j2) = crv_uq(ct_loop);
        crv_post_uq(j2,j1) = crv_post_uq(j1,j2);
        crv_postmean(j1,j2) = crv_postmean1(ct_loop);
        crv_postmean(j2,j1) = crv_postmean(j1,j2);
    end
end
toc;


save(’/home/users/jingzhou/Dissertation/30-DunsonXing_Extension/DataApplication/
Results&plots/splice4c_K25000_dxAnir’,’crv_post_lq’,’crv_post_uq’,...
’crv_postmean’,’alphaout’,’nuout’,’piout’,’nu_max’, ’NMI_postmean’,...
’NMI_post_lq’,’NMI_post_uq’,’nu_comps_95’,’nu_comps_lq_95’,...
’nu_comps_uq_95’,’n’,’p’,’N’);
```

APPENDIX III

Computation Code for Chapter 3

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Goal: NBDPS analysis using Case-control sp-PARAFAC model
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


clear;clc;close all;


g = dlmread(’gg.txt’);  %index for birth defects outcome
%g = 4;


tic;
% -- define global parameters -- %
d = 2; k = 10;             % d =  # categories, k = # factors
nrun = 20000; burn = 5000; thin = 5;
%nrun = 4; burn = 1; thin =1;
eff_samp1 = (nrun - burn)/thin;


% -- load data -- %
%data1 = dlmread(’/Users/Amani/Dropbox/Dissertation Research/
30-Dunson&XingExtension/NBDPS2013/dataOct22/dataPart1.txt’);
%data2 = dlmread(’/Users/Amani/Dropbox/Dissertation Research/
30-Dunson&XingExtension/NBDPS2013/dataOct22/dataPart2.txt’);
data1 = dlmread(’/netscr/jingzhou/NBDPS2013/dataPart1.txt’);
```

```
data2 = dlmread('/netscr/jingzhou/NBDPS2013/dataPart2.txt');
data = [data1 data2];
[n,q] = size(data); p = q-54;

Xn = data(:,55:end)+1;  Yn = data(:,g);
n0 = sum(Yn==0); n1 = sum(Yn==1);
rowsn0 = repmat(1:p,n0,1); rowsn0 = rowsn0(:);
rowsn1 = repmat(1:p,n1,1); rowsn1 = rowsn1(:);
Xn0 = Xn(Yn==0,:); colsn0 = Xn0(:); lin_idx0 = sub2ind([p,d],rowsn0,colsn0);
Xn1 = Xn(Yn==1,:); colsn1 = Xn1(:); lin_idx1 = sub2ind([p,d],rowsn1,colsn1);

% --  discrete uniform probabilities -- %
lam0 = repmat([1/d 1/d],p,1);    %p*d, d=2

% -- initialize parameters -- %
alpha = 1;              % dp hyperparameters for stick breaking weights
aal = 1; bal = 1;       % gamma hyperparameters for alpha
a = ones(d,1);          % dirichlet hyperparameter for \lambda_h^{(j)}
Lambda0 = zeros(k,p,d);
% loadings for preds--initialize with sample cell frequencies
Lambda1 = zeros(k,p,d);
% loadings for preds--initialize with sample cell frequencies

for h = 1:k
    for j = 1:p
        Lambda0(h,j,:) = [sum(Xn(:,j)==1)/n sum(Xn(:,j)==2)/n];  %d=2
        Lambda1(h,j,:) = [sum(Xn(:,j)==1)/n sum(Xn(:,j)==2)/n];
    end
end

gamm=20;   %prior pi_h~Beta(1,gamm)
pi0=betarnd(1,gamm,[k,1]); pi1=betarnd(1,gamm,[k,1]);

nus0 = betarnd(1,alpha,[k-1,1]); nu0 = zeros(k,1);
nu0(1:k-1) = nus0.*cumprod([1;1-nus0(1:k-2)]); nu0(k) = 1 - sum(nu0(1:k-1));
nus1 = betarnd(1,alpha,[k-1,1]); nu1 = zeros(k,1);
nu1(1:k-1) = nus1.*cumprod([1;1-nus1(1:k-2)]); nu1(k) = 1 - sum(nu1(1:k-1));
z0 = zeros(n0,1); zupdateprob0 = zeros(n0,k); zcts0 = zeros(k,1);
z1 = zeros(n1,1); zupdateprob1 = zeros(n1,k); zcts1 = zeros(k,1);

%matrix to be used in posterior
mat0a = zeros(p,d);     %for updating from Dirichlet for Lambda0
mat0null = zeros(p,d);  %for updating from the null group for Lambda0
indu0d0= zeros(k,p);      %indicator for Lambda0 being assigned to baseline

mat1a = zeros(p,d);     %for updating from non-null group for Lambda1
mat1null = zeros(p,d);  %for updating from the null group for Lambda1
indu0d1= zeros(k,p);      %indicator for Lambda1 being assigned to baseline

% -- define output files -- %
alphaout = zeros(nrun,1); nu0out = zeros(nrun,k);
nu1out = zeros(nrun,k); nu0_max = zeros(nrun,1);
```

```
pi0out= zeros(nrun,k); pi1out= zeros(nrun,k);
nu_comps_95 = zeros(nrun,1); %nu_comps_99 = zeros(nrun,1);

OR = zeros(eff_samp1,p); mainEff = zeros(eff_samp1,p);

% -- start Gibbs sampler -- %
for i = 1:nrun

    % -- update z0 and z1: done -- %
    for h = 1:k
        Lambda0h = reshape(Lambda0(h,:,:),p,d);
        Lambda1h = reshape(Lambda1(h,:,:),p,d);
        tmpmat0Ld0 = reshape(Lambda0h(lin_idx0),[n0,p]);
        tmpmat1Ld1 = reshape(Lambda1h(lin_idx1),[n1,p]);
        zupdateprob0(:,h) = nu0(h) * prod(tmpmat0Ld0,2);
        zupdateprob1(:,h) = nu1(h) * prod(tmpmat1Ld1,2);
    end
    zupdateprob0a = bsxfun(@times,zupdateprob0,1./(sum(zupdateprob0,2)));
    zupdateprob1a = bsxfun(@times,zupdateprob1,1./(sum(zupdateprob1,2)));
    matz0 = [zeros(n0,1) cumsum(zupdateprob0a,2)];
    matz1 = [zeros(n1,1) cumsum(zupdateprob1a,2)];
    rr0 = unifrnd(0,1,[n0,1]); rr1 = unifrnd(0,1,[n1,1]);
    for l = 1:k
        ind0 = rr0 > matz0(:,l) & rr0 <= matz0(:,l+1);
        z0(ind0) = l;    %size: n0*1
        ind1 = rr1 > matz1(:,l) & rr1 <= matz1(:,l+1);
        z1(ind1) = 1;    %size: n1*1
    end



    % -- update lambda0 and lambda1: done -- %
    for h = 1:k
        zh0 = (z0==h); zcts0(h) = sum(zh0);
        zh1 = (z1==h); zcts1(h) = sum(zh1);
        for c = 1:d
            term0 = sum(bsxfun(@times,(Xn0==c),zh0));
            term1 = sum(bsxfun(@times,(Xn1==c),zh1));
            mat0a(:,c) = (a(c) + term0)’; mat0null(:,c) = (term0)’;     %p*d
            mat1a(:,c) = (a(c) + term1)’; mat1null(:,c) = (term1)’;     %p*d
        end
        tmpLam0h = gamrnd(mat0a,1);
        Lam0h = bsxfun(@times,tmpLam0h,1./sum(tmpLam0h,2));
        Lambda0(h,:,:) = Lam0h;
        tmpLam1h = gamrnd(mat1a,1);
        Lam1h = bsxfun(@times,tmpLam1h,1./sum(tmpLam1h,2));
        Lambda1(h,:,:) = Lam1h;

        tmpu0hd0=(1-pi0(h))*prod(lam0.^mat0null,2);    %p*1
        tmpu1hd0=pi0(h)*gamma(sum(a))/prod(gamma(a))*prod(gamma(mat0a),2)./
        gamma(sum(mat0a,2));  %p*1
        u0hd0=tmpu0hd0./(tmpu0hd0+tmpu1hd0);
        indu0d0(h,:)=(rand(1,p)<u0hd0’);
        u0d0row =find(indu0d0(h,:));
        if length(u0d0row)>0
            Lambda0(h,u0d0row,:)=lam0(u0d0row,:);
        end

        tmpu0hd1=(1-pi1(h))*prod(lam0.^mat1null,2);    %p*1
        tmpu1hd1=pi1(h)*gamma(sum(a))/prod(gamma(a))*prod(gamma(mat1a),2)./
        gamma(sum(mat1a,2));  %p*1
        u0hd1=tmpu0hd1./(tmpu0hd1+tmpu1hd1);
        indu0d1(h,:)=(rand(1,p)<u0hd1’);
        u0d1row =find(indu0d1(h,:));
        if length(u0d1row)>0
            Lambda1(h,u0d1row,:)=lam0(u0d1row,:);
        end
    end



    % -- update pi: done -- %
    pi0 = betarnd(1 + p- sum(indu0d0,2),gamm + sum(indu0d0,2));
    pi1 = betarnd(1 + p- sum(indu0d1,2),gamm + sum(indu0d1,2));

    % -- update nu0 and nu1: done -- %
    for h = 1:k-1
        nus0(h) = betarnd(1 + zcts0(h),alpha + sum(zcts0(h+1:k)));
        nu0(1:k-1) = nus0.*cumprod([1;1-nus0(1:k-2)]);
        nu0(k) = 1-sum(nu0(1:k-1));
        nus1(h) = betarnd(1 + zcts1(h),alpha + sum(zcts1(h+1:k)));
        nu1(1:k-1) = nus1.*cumprod([1;1-nus1(1:k-2)]);
        nu1(k) = 1-sum(nu1(1:k-1));
    end


    % -- update alpha: done-- %
    nuss = 1 - nus0(1:k-1); nuss(nuss < 1e-6) = 1e-6;
    nuss2 = 1 - nus1(1:k-1); nuss2(nuss2 < 1e-6) = 1e-6;
    alpha = gamrnd( aal + 2*k - 1, 1/(bal - sum(log(nuss))-sum(log(nuss2))) );



    % -- first write files to be stored across each iteration -- %
    nu0out(i,:) = nu0';nu0_ord = sort(nu0,'descend'); nu0_max(i) = nu0_ord(1);
    nu1out(i,:) = nu1';
    nu_comps_95(i,:) = sum((cumsum(nu0_ord) >=0.95) == 0) + 1;
    %nu_comps_99(i,:) = sum((cumsum(nu_ord) >=0.99) == 0) + 1;
    alphaout(i) = alpha;
    pi0out(i,:)=pi0'; pi1out(i,:)=pi1';

    % -- marginal dependence -- %
    if (mod(i,thin) == 0 && i > burn)
        pr1=sum( bsxfun(@times,nu1,Lambda1(:,:,2)),1 );
        %check dimension compatable  %size: 1*p
```

```
        pr2=sum( bsxfun(@times,nu1,Lambda1(:,:,1)),1 );
        pr3=sum( bsxfun(@times,nu0,Lambda0(:,:,2)),1 );
        pr4=sum( bsxfun(@times,nu0,Lambda0(:,:,1)),1 );
        OR((i-burn)/thin,:)= pr1.*pr4./pr2./pr3;
        mainEff((i-burn)/thin,:)= log(pr1.*pr4./pr2./pr3);

    end

    %if mod(i,100) == 0, disp(i); end
end

OR_lq = quantile(OR,0.025); OR_uq = quantile(OR,0.975); OR_postmean = mean(OR);
mainEff_lq = quantile(mainEff,0.025); mainEff_uq = quantile(mainEff,0.975);
mainEff_postmean = mean(mainEff);

OR_postmean_sig = OR_postmean;
OR_postmean_sig(OR_lq<1 & OR_uq>1)=1;    %size: 1*p
mainEff_postmean_sig = mainEff_postmean;
mainEff_postmean_sig(mainEff_lq<0 & mainEff_uq>0)=0;    %size: 1*p


toc;


dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/sigBeta.txt’,...
    [g’,mainEff_postmean_sig],’-append’,’coffset’, 1,’delimiter’, ’\t’);
```

```
dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/sigOR.txt’,...
    [g’,OR_postmean_sig],’-append’,’coffset’, 1,’delimiter’, ’\t’);

dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/Beta_mean.txt’,...
    [g’,mainEff_postmean],’-append’,’coffset’, 1,’delimiter’, ’\t’);
dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/Beta_lq.txt’,...
    [g’, mainEff_lq],’-append’,’coffset’, 1,’delimiter’, ’\t’);
dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/Beta_uq.txt’,...
    [g’,mainEff_uq],’-append’,’coffset’, 1,’delimiter’, ’\t’);
dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/OR_mean.txt’,...
    [g’,OR_postmean],’-append’,’coffset’, 1,’delimiter’, ’\t’);
dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/OR_lq.txt’,...
    [g’,OR_lq],’-append’,’coffset’, 1,’delimiter’, ’\t’);
dlmwrite(’/netscr/jingzhou/SPDmodel/NBDPS2013/Results/OR_uq.txt’,...
    [g’,OR_uq],’-append’,’coffset’, 1,’delimiter’, ’\t’);
```

APPENDIX IV

Computation Code for Chapter 4

```
################################################################################
# Data analysis example for SSR-cLDA
################################################################################
library(MASS)
library(gsDesign)
library(nlme)

#Modified spending function for sfLDOF after info.frac 0.85
sfLDOFtrunc <-
  function (alpha, t, param, limit=0.85)
  {
    checkScalar(alpha, "numeric", c(0, Inf), c(FALSE, FALSE))
    checkVector(t, "numeric", c(0, Inf), c(TRUE, FALSE))
    t[t > limit] <- 1
    z <- -qnorm(alpha/2)
    x <- list(name = "Lan-DeMets O'brien-Fleming approximation",
              param = NULL, parname = "none", sf = sfLDOF, spend = 2 *
                (1 - pnorm(z/sqrt(t))), bound = NULL, prob = NULL)
    class(x) <- "spendfn"
    x
  }
```

```
#This function generates Toeplitz correlation structure
toepmat <- function(T=4,rho=rho){ #rho is a vector of length T-1
sigma <- matrix(0,nrow=T,ncol=T)
for (i in 1:(T-1)) {
for (j in 1:(T-i)) {
sigma[i,i+j] <- rho[j]; sigma[i+j,i] <- rho[j]
}
}
diag(sigma) = rep(1,T)
sigma <- sigma
return(sigma)
}
```

```
#This function generates compound symmetry correlation structure
csmat <- function(T=4,rho=rho){ #rho is a numeric constant
sigma <- matrix(0,nrow=T,ncol=T)
for (i in 1:(T-1)) {
for (j in 1:(T-i)) {
sigma[i,i+j] <- rho; sigma[i+j,i] <- rho
}
}
diag(sigma) = rep(1,T)
return(sigma)
}
```

```
#This function generates AR1 correlation structure

ar1mat <- function(T=4,rho=rho){ #rho is a numeric constant

sigma <- matrix(0,nrow=T,ncol=T)

for (i in 1:(T-1)) {

for (j in 1:(T-i)) {

sigma[i,i+j] <- rho^j; sigma[i+j,i] <- rho^j

}

}

diag(sigma) = rep(1,T)

sigma <- sigma

return(sigma)

}



#This function is required by the function Lambda which is

# in turn required by function ss.LDA

Rtt.0.inv <- function(t,T=T,cormat,rbase){

Rtt.0.inv <- matrix(0,nrow=T,ncol=T)

R0t <- rbase[1:t]

Rtt.0 <- cormat[1:t,1:t] - R0t%*%t(R0t)

Rtt.0.inv[1:t,1:t] <- solve(Rtt.0)

Rtt.0.inv <- (Rtt.0.inv+t(Rtt.0.inv))/2

return(Rtt.0.inv)

}


#This function is required by the function ss.LDA
```

```
Lambda <- function(cormat,rbase,pt,T=T) {

L = 0

for (t in 1:T) {L <- L + pt[t]*Rtt.0.inv(t,T=T,cormat,rbase)}

return(L)

}



ss.LDA <- function(

I.max,#Required maximum information

S, #vector of standard deviations of the repeated measures (including baseline)

R, #Correlation matrix of the repeated measures (including baseline)

r1,#Group 1 - Retention rate at each postbaseline timepoint

r2=NULL,  #Group 2 - Retention rate at each postbaseline timepoint

cvec=NULL,#Contrast vector for the hypothesis based on postbaseline means

lambda=1  #Randomization ratio: lambda = n2/n1

)

{

  if (is.null(r2)) r2 <- r1

  p1 <- r1 - c(r1[-1],0)  #Group 1 - Proportion of enrolled subject who

  # dropped out between consecutive postbaseline time points

  # starting from first postbaseline and second

  p2 <- r2 - c(r2[-1],0)  #Similarly for Group 2


  J <- dim(R)[1] - 1     #Number of postbaseline measurements
  post.sig <- S[-1]
  if (is.null(cvec)) cvec <- c(rep(0,J-1),1)
  #This is the contrast vector for the hypothesis of interest, the default is
  #the difference between mu.c and mu.t at the last postbaseline time point


  cormat <- R[2:(J+1),2:(J+1)]
  #Sub-matrix of correlation for postbaseline measures
  rbase <- R[1,2:(J+1)]
  #Correlation of each postbaseline measures with baseline

  Lambda.inv1 <- solve(Lambda(cormat,rbase,pt=p1,T=J))
  Lambda.inv2 <- solve(Lambda(cormat,rbase,pt=p2,T=J))
  n1 <- ceiling(I.max*(t(cvec)%*%diag(post.sig)%*%Lambda.inv1%*%
  diag(post.sig)%*%cvec + (1/lambda)*t(cvec)%*%diag(post.sig)%*%
  Lambda.inv2%*%diag(post.sig)%*%cvec))
  n2 <- ceiling(lambda*n1)
  n.max <-  as.numeric(n1+n2)    #n.max is the calculated total sample size
  return(n.max)
}


#The following function simulate information-based group
# sequential design for continuous longitudinal measurements.
MaxInfoDesign.LDA <- function(
  ## para for gsDesign for calculating Imax and boundaries##
  alpha=.025,  #Target type 1 error (1-sided alpha expected)
  beta=.1,     #Power = 1 - beta
  delta=0.25,  #Treatment difference at last time point
  test.side=1,  #Test sides (1-sided is default)
  timing=c(0.5,1),#Analysis times

  ## para for ss.LDA + Imax calc. from gsDesign ##
  R =csmat(T=5,rho=0.579),#Correlation matrix of the
   #repeated measures (including baseline)
  S =rep(.925,5),#vector of standard deviations of the
   #repeated measures (including baseline)
  r1 =0.9^((1:4)/4),#Group 1 - Retention rate at each postbaseline timepoint;
   #the default is exponential retention rate
  r2 =0.9^((1:4)/4),#Group 2 - Retention rate at each postbaseline timepoint

  ## real data input & results from previous interim looks ##
  longdat,      # data with long and slim format
  update.act.t, # I(t)/Imax from all previous interim looks with
   # future timing as original plan (length same as timing vector)
  interim.no,    # the current interim look number
  new.Nmax=NULL, # new Nmax calculated from last previous interim if any
  ifFinal  #if current analysis is final(T or F)
){

  #Calculation maximum information
  K <- length(timing)
```

```
I.fix <- ((qnorm(1-alpha)+qnorm(1-beta))/delta)^2
 #Calculates maximum information for a fixed design
I.max <- gsDesign(k=K,sfu=sfLDOF,sfl=sfLDOF,timing=timing,
n.fix=I.fix,beta=beta,alpha=alpha,test.type=test.side)$n.I[K]
#Calculates maximum information for group sequential design using gsDesign

#Converting maximum information to sample size
J <- dim(R)[1] - 1 #Number of postbaseline measurements

n.max <- ss.LDA(I.max,S,R,r1)

#Initialize monitoring parameters
rej.ub<- 0
rej.lb <- 0
int.info <- NA
inf.frac <- NA
if (interim.no+1 == K) {ifNextFinal <- 'TRUE'} else {ifNextFinal <- 'FALSE'}

if (is.null(new.Nmax)) {n <- n.max} else {n <- new.Nmax}
no.enrolled <- length(unique(longdat$alloc))
comp.cases.dat <- longdat[which(longdat$flag==0),]
#flag is 1 if the patient is still continuing
#only subjects that have completed the last measurement or
#dropped out of study will be used at any analysis.
n.int <- length(unique(comp.cases.dat$alloc))
#only completed or discontinued patients are included
```

```
#Ensure the model fitted in the next line is the same as
#the analysis model above.
#The first model fits constrained LDA with unstructured var-cov matrix;
#second fit compound symmetry;
#a <- try(fit.gls <- gls(y~factor(week)+trt2wk1+trt2wk2+trt2wk3+trt2wk4,
data=comp.cases.dat,corr=corSymm(form = ~ 1 |alloc),weights =
 varIdent(form = ~ 1 | week)),TRUE)
a <- try(fit.lme <- lme(y ~ as.factor(week) + trt2wk1 + trt2wk2 + trt2wk3 +
trt2wk4,data=comp.cases.dat,random=~1|alloc),TRUE)

if (length(a)==1) { #This catches error from the model fitting
  rej.ub <- NA; rej.lb <- NA
  break
}

#est.effect <- coef(fit.gls)[9]
#Use this set if gls function is used for model fitting
#se <- sqrt(diag(fit.gls$varBeta)[9])

est.effect <- fixed.effects(fit.lme)[9]
#Use this set if lme function is used for model fitting
se <- sqrt(fit.lme$varFix[9,9])

#Calculate new n_max to see whether sample size needs to be changed
r.n <- ceiling(n.int/min(se^(-2)/I.max,1))
if (interim.no < K) { # 2nd & 3rd bullet of adaptation from the slide
```

```
  if (r.n<=no.enrolled || r.n <= ceiling(n*timing[interim.no+1])) {
    n <- max(r.n,no.enrolled)
    ifNextFinal = 'TRUE'
  }}
if (interim.no<K && r.n>n) {  #5th bullet of adaptation from the slide
  n <- min(r.n,2*n.max)
  #The re-estimated sample size has been capped at 2 times original.
}
if (interim.no==K) { #last interim
  #n <- min(r.n,2*n.max)  #although won't collect more data
  n <- no.enrolled
}


#Test whether we can stop early
test.stat <- est.effect/se
p.value <- test.side*(1 - pnorm(test.stat))
#test.side*(1-pt(test.stat,df=df)) #
inf.frac <- min(se^(-2)/I.max,1)
int.info <- se^(-2)/I.max
update.act.t[interim.no] <- inf.frac

if (interim.no<K && ifFinal==F) {
  if (update.act.t[interim.no] == 1) {
  update.act.t <- update.act.t[1:interim.no]}
  #change from last version to make sure gsDesign can work
  else if (update.act.t[interim.no]>=update.act.t[interim.no+1])
  {update.act.t=update.act.t[-(interim.no+1)]}
  b <- try(bdd <- gsDesign(k=length(update.act.t),sfu=sfLDOF,
   sfl=sfLDOF,timing=update.act.t,
  n.fix=I.fix,beta=beta,alpha=alpha,test.type=test.side),TRUE)
}

if (interim.no==K) {
  b0 <- try(bdd0 <- gsDesign(k=K,sfu=sfLDOFtrunc,sfl=sfLDOF,
  timing=update.act.t[-K],n.fix=I.fix,beta=beta,alpha=alpha,
  test.type=test.side),TRUE)
  b <- try(bdd <- gsDesign(k=K,sfu=sfLDOFtrunc,sfl=sfLDOF,n.fix=I.fix,
  n.I=c(b0$n.I[1:(K-1)],update.act.t[interim.no]*b0$n.I[K]),
   beta=beta,alpha=alpha, test.type=test.side),TRUE)
}

else if (ifFinal==T) {
  b0 <- try(bdd0 <- gsDesign(k=interim.no,sfu=sfLDOFtrunc,sfl=sfLDOF,
  timing=update.act.t[1:(interim.no-1)],n.fix=I.fix,beta=beta,
  alpha=alpha,test.type=test.side),TRUE)
  b <- try(bdd <- gsDesign(k=interim.no,sfu=sfLDOFtrunc,sfl=sfLDOF,
  n.fix=I.fix,n.I=c(b0$n.I[1:(interim.no-1)],update.act.t[interim.no]*
  b0$n.I[interim.no]),beta=beta,alpha=alpha,test.type=test.side),TRUE)
}

if (length(b)==1) {
  rej.ub <- NA
  rej.lb <- NA
  break
}
if (!(is.null(bdd$upper$bound))) {
  if (-test.stat>=bdd$upper$bound[interim.no]) {
    rej.ub <- 1
  }
}
if (!(is.null(bdd$lower$bound))) {
  if (-test.stat<=bdd$lower$bound[interim.no]) {
    rej.lb <- 1
  }
}



input.parameters = c(
  alpha=alpha,
  power=1-beta,
  Planned.timing=timing,
  delta=delta
)

if (rej.ub==1) {ifSTOPtrial = "Stop with sig. Efficacy"} else
  if (rej.lb==1) {ifSTOPtrial = "Stop with sig. Futility"}  else
    if (inf.frac>=1) {ifSTOPtrial =
       "Max Information reached but no sig. detected"}  else
        if (interim.no==K) {ifSTOPtrial =
         "Stop for planned K but information not reached"} else
        {ifSTOPtrial = "Continue to the next interim"}
  #   ifSTOPtrial = rej.ub==1||rej.lb==1||inf.frac>=1
  #   Efficacy = (rej.ub==1)
  #   Futility = (rej.lb==1)
  #   MaxInfoReached = (inf.frac>=1)
  return(c(input.parameters, ifSTOPtrial=ifSTOPtrial,
  ifNextFinal=ifNextFinal, update.act.t=round(update.act.t,3),
  orig.Nmax=n.max, New.Nmax=n, current.inf.frac=round(int.info,3)))
}


###### First interim analysis ###
setwd('/Users/Amani/Dropbox/Dissertation Research/
26-SSR-LDA writeup/StatInMed Revision/Code/dataExample')
results = MaxInfoDesign.LDA(
  alpha    =.025,
  beta   =.1,
  delta   =0.25,
  test.side =1,
  timing   =c(0.5,1),
  R  =csmat(T=5,rho=0.579),
  S        =rep(.925,5),
  r1       =0.8^((1:4)/4),
  r2       =0.8^((1:4)/4),

  longdat = adeff,
  update.act.t = timing,
  interim.no =1,
  ifFinal=F,
  new.Nmax =NULL)

as.data.frame(results)


###### Final analysis ###
results2 = MaxInfoDesign.LDA(
  alpha    =.025,
  beta   =.1,
  delta      =0.25,
  test.side =1,
  timing    =c(0.5,1),
  R  =csmat(T=5,rho=0.579),
  S        =rep(.925,5),
  r1       =0.8^((1:4)/4),
  r2       =0.8^((1:4)/4),
  longdat = adeff2,
  update.act.t = c(0.514,1),
  interim.no =2,
  ifFinal=T,
  new.Nmax =411)


as.data.frame(results2)
```

BIBLIOGRAPHY

Agresti, A. (2002). *Categorical data analysis*, volume 359. Wiley-interscience.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59,** 19–35.

Anderson, K. (2014). gsDesign: Group sequential design R package and its GUI. *http://cran.r-project.org/web/packages/gsDesign/index.html* .

Antoniak, C. (1974). Mixtures of dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics* pages 1152–1174.

Armagan, A., Dunson, D., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23,** 119–143.

Ashby, D., Hutton, J. L., and McGee, M. A. (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *The Statistician* pages 385–397.

Ashford, J. and Sowden, R. (1970). Multi-variate probit analysis. *Biometrics* pages 535–546.

Bhattacharya, A. and Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98,** 291–306.

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

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics* **1,** 353–355.

Bro, R. (1997). PARAFAC. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38,** 149–171.

Burington, B. E. and Emerson, S. S. (2003). Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* **59,** 770–777.

Byrne, S. P. and Dawid, A. P. (2013). Retrospective-prospective symmetry in the likelihood and Bayesian analysis of case-control studies. *Biometrika* page ast050.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics* **9,** 717–772.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103,** 1438–1456.

Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Carvalho, C. and Scott, J. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96,** 497–512.

Chartrand, R. (2012). Nonconvex splitting for regularized low-rank plus sparse decomposition. *IEEE Transactions on Signal Processing* **60,** 5810–5819.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85,** 347–361.

Chuang-Stein, C., Anderson, K., Gallo, P., and Collins, S. (2006). Sample size reestimation: a review and recommendations. *Drug Information Journal* **40,** 475–484.

Coffey, C. S. and Kairalla, J. A. (2008). Adaptive clinical trials: Progress and challenges. *Drugs in R&D* **9,** 220–242.

Cohen, J. and Rothblum, U. (1993). Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications* **190,** 149–168.

Coull, B. A., Hobert, J. P., Ryan, L. M., and Holmes, L. B. (2001). Crossed random effect models for multiple outcomes in a study of teratogenesis. *Journal of the American Statistical Association* **96,** 1194–1204.

Dawid, A. and Lauritzen, S. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21,** 1272–1317.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications* **21,** 1253–1278.

Diggle, P. (2002). *Analysis of longitudinal data*, volume 25. Oxford University Press, USA.

Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (moss) algorithm for log-linear models with conjugate priors. *Statistical Methodology* **7,** 240–253.

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

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

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77,** 875–892.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics* pages 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* pages 615–629.

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

Friedlander, M. and Hatz, K. (2005). Computing non-negative tensor factorizations. *Optimization Methods and Software* **23,** 631–647.

Galbraith, S. and Marschner, I. C. (2003). Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine* **22,** 1787–1805.

Gao, P., Liu, L., and Mehta, C. (2013). Exact inference for adaptive group sequential designs. *STATISTICS IN MEDICINE* **32,** 3991–4005.

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2,** 1360–1383.

Ghosh, M. and Chen, M.-H. (2002). Bayesian inference for matched case-control studies. *Sankhyā: The Indian Journal of Statistics, Series B* pages 107–127.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Gregory, D. and Pullman, N. (1983). Semiring rank: Boolean rank and nonnegative rank factorizations. *J. Combin. Inform. System Sci* **8,** 223–233.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association* **106,** 1383–1393.

Harshman, R. (1970). Foundations of the PARAFAC procedure: models and conditions for an " explanatory" multimodal factor analysis.

Harshman, R. and Lundy, M. (1994). PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis* **18,** 39–72.

Jennison, C. and Turnbull, B. (2000). *Group sequential methods with applications to clinical trials*. CRC Press.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20,** 388–400.

Karatzoglou, A., Amatriain, X., Baltrunas, L., and Oliver, N. (2010). Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86. ACM.

Kim, K. (1989). Point estimation following group sequential tests. *Biometrics* pages 613–617.

Kim, Y. and Choi, S. (2007). Nonnegative tucker decomposition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Kittelson, J. M., Sharples, K., and Emerson, S. S. (2005). Group sequential clinical trials for longitudinal data with analyses using summary statistics. *Statistics in Medicine* **24,** 2457–2475.

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

Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM Review* **51,** 455–500.

Kunihama, T. and Dunson, D. B. (2013). Bayesian modeling of temporal dependence in large sparse contingency tables. *Journal of the American Statistical Association* **108,** 1324–1338.

Lan, K. and DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

Lazarsfeld, P. and Henry, N. (1968). *Latent structure analysis.* Houghton, Mifflin.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791.

Liang, K. and Zeger, S. (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhyā: The Indian Journal of Statistics, Series B* pages 134–148.

Lim, L. and Comon, P. (2009). Nonnegative approximations of nonnegative tensors. *Jour. Chemometrics* pages 432–441.

Liu, J., Liu, J., Wonka, P., and Ye, J. (2012). Sparse non-negative tensor factorization using columnwise coordinate descent. *Pattern Recognition* **45,** 649–656.

Lu, K., Luo, X., and Chen, P. (2008). Sample size estimation for repeated measures analysis in randomized clinical trials with missing data. *The International Journal of Biostatistics* **4,** 1–16.

Lu, K., Mehrotra, D., and Liu, G. (2009). Sample size determination for constrained longitudinal data analysis. *Statistics in Medicine* **28,** 679–699.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics* **1,**.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation* **23,** 727–741.

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

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* pages 215–232.

Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine* **7,** 1223–1230.

Mehta, C. and Tsiatis, A. (2001). Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* **35,** 1095–1112.

Müller, P., Parmigiani, G., Schildkraut, J., and Tardella, L. (1999). A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* **55,** 858–866.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84,** 523–537.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics* **22,** 43–65.

Nurminen, M. and Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* pages 67–77.

Ochi, Y. and Prentice, R. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71,** 531–543.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5,** 111–126.

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

Pati, D., Bhattacharya, A., Pillai, N., and Dunson, D. (2013). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *arXiv:1206.3627* .

Polson, N. and Scott, J. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9 (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds.)*, pages 501–538. Oxford University Press, New York.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66,** 403–411.

Scharfstein, D., Tsiatis, A., and Robins, J. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* pages 1342–1350.

Scott, J. and Berger, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38,** 2587–2619.

Seaman, S. R. and Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika* **88,** 1073–1088.

Seaman, S. R. and Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91,** 15–25.

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

Shashua, A. and Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799. ACM.

Shih, W. and Gould, A. (1995). Re-evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. *Statistics in Medicine* **14,** 2239–2248.

Sinha, S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case–control studies with multiple disease states. *Biometrics* **60,** 41–49.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B. K., and Carroll, R. J. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association* **100,** 591–601.

Staicu, A.-M. (2010). On the equivalence of prospective and retrospective likelihood methods in case-control studies. *Biometrika* **97,** 990–996.

Talagrand, M. (1996). A new look at independence. *The Annals of Probability* **24,** 1–34.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* **58,** 267–288.

Tsiatis, A. A. (2006). Information-based monitoring of clinical trials. *Statistics in Medicine* **25,** 3236–3244.

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31,** 279–311.

Wang, H. and Ahuja, N. (2005). Rank-R approximation of tensors using image-as-matrix representation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 346–353. IEEE.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. *Bayesian statistics* **7,** 723–732.

Whittaker, J. (1990). Graphical models in applied multivariate analysis. *Chichester New York et al.: John Wiley & Sons* .

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9,** 65–72.

Yang, Y. and Dunson, D. B. (2013). Bayesian conditional tensor factorizations for high-dimensional classification. *arXiv:1301.4950* .

Yoon, P. W., Rasmussen, S. A., Lynberg, M., Moore, C., Anderka, M., Carmichael, S., Costa, P., Druschel, C., Hobbs, C., Romitti, P., et al. (2001). The national birth defects prevention study. *Public health reports* **116,** 32.

Zelen, M. and Parker, R. (1986). Case-control studies and Bayesian inference. *Statistics in Medicine* **5,** 261–269.

Zhang, T. and Golub, G. (2001). Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications* **23,** 534–550.

Zhang, X., Boscardin, W., and Belin, T. (2008). Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Computational statistics & data analysis* **52,** 3697–3708.

Zhou, J., Bhattacharya, A., Herring, A., and Dunson, D. (2013). Bayesian factorizations of big sparse tensors. *arXiv preprint arXiv:1306.1598* .

Zhu, L., Ni, L., and Yao, B. (2011). Group sequential methods and software applications. *The American Statistician* **65,** 127–135.

Zucker, D. and Denne, J. (2002). Sample–size redetermination for repeated measures studies. *Biometrics* **58,** 548–559.