8-2021

Bayesian Framework for Causal Inference in Complex Data Structure

Li He
Clemson University, listats2021@gmail.com

Bayesian Framework for Causal Inference in Complex Data Structure

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Statistics

---

by
Li He
August 2021

Accepted by:
Dr. William C Bridges Jr, Committee Chair
Dr. Yu-bo Wang, (Co Advisor)
Dr. Patrick Gerard
Dr. Qiong Zhang

Abstract

Causal inference analysis is one of the most significant and well researched topics in the analysis of observational studies. It addresses the challenge of estimating the relationship between the treatment of interest and the outcome variable (i.e., the treatment effect) in the presence of background covariates. This study aims to determine how causal inference in observational studies can be extended to cover data sets where post treatment variables exist, cluster level covariates exist, or the number of observations in the treatment groups is very unbalanced. More specifically, this study can be summarized into the following two sub-topics.

1. The first part of this study focuses on estimating the treatment effect when both the post-treatment variable and cluster level variables exist in the data sets. In previous literature, one popular method, called, the principal stratification method, can properly handle the problem with a post-treatment variable to reach an unbiased treatment effect estimate. The traditional method to consider the cluster level effect is to include the cluster labels as random effects in the model. When either the size of clusters is large, or each cluster only contains small number of observations, we find this method often results in poor estimates of the treatment effect.

2. The second part of this study focuses on estimating the treatment effect estimation when the number of observations is very unbalanced in the two treatment groups (i.e., the treatment or control is a rare event) and background covariates and cluster effects exist in the data sets. Propensity score analysis is the traditional method to adjust for the covariate imbalance between the treated and control groups. The logit or probit links are often used in the propensity score for balancing the background covariates between the treatment groups. When there is severe imbalance between the treatments, we find this method often results in poor estimates of the treatment effect.

To tackle these issues, we propose comprehensive Bayesian frameworks for estimating the treatment effect in the presence of post-treatment variables, cluster and/or imbalances.

For the first problem, a comprehensive framework with post-treatment variable and a clustered structure is addressed. The proposed framework constructs the clustering structure as random effects with a spike and slab prior in a Bayesian hierarchical model. The key idea is to estimate the causal treatment effect with a more parsimonious and less complex model and thereby also reduce the computing complexity. This is especially useful when a large number of clusters have no significant influence on the outcome. Several different data generating scenarios (including combinations of clustering structure and post-treatment variable) are considered through simulation studies. The simulation results suggest that the proposed methodology generates the most consistent estimates. The advantages of the proposed methodology are also demonstrated using two case studies of educational performance and infant birth weight.

For the second problem, we propose a two-step Bayesian framework. The first step is to estimate the propensity score using a proposed generalized skewed link function. The generalized link function is adopted from a skewed link where the parameter associated with skewness follows a Dirac-spike prior and a mixture structure for the error term. As one of the three commonly used sparsity inducing prior, the Dirac-spike prior, also allows us to determine the necessity of skewness. The second step is to estimate the treatment effect considering propensity score as additional latent variable to adjust for covariates imbalance in outcome analysis. The proposed framework includes the clusters as indicator variables in hierarchical models. The normal mixture inverse gamma (NMIG), one type of spike and slab prior, is used to allow for many of the cluster effects to not be significant. The framework can determine the true underlying relation between background covariates and the binary response with least misspecification rate.

These results of empirical simulations and data application case studies show advantages of the proposed methods. Both approaches can result in more parsimonious models as one distinct advantage. Another one is that the Bayesian framework can use computationally efficient Markov Chain Monte Carlo (MCMC) sampling algorithms separately by making use of data augmentation and rewriting technique like the Pólya–Gamma technique for binary regression in the second problem.

Dedication

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my mum, Xianghong Li whose words of encouragement, long-term spiritual support and never giving up on me push me to where I am and what I can achieve now throughout my PhD studies and my life in general.

I also dedicate this dissertation to my many friends and church family when I first came to Clemson who have supported me throughout the process, helped me whenever I needed. I will always appreciate all they have done, especially the Reeves's family for helping me out at the beginning, introducing me to new friends, making me feel at home. So, I could concentrate on research for these five years.

I dedicate this work and give special thanks to my mentors and friends since college Alan Brown, Vera Tanner and Danita Rios for support and encouragement, being there and never leaving my side whenever I felt lost or during difficult time throughout the entire doctorate program. All of them have been my cheerleaders.

Finally, I also dedicate this work to my dog Summer as a sweet and great accompany especially during the last year of the program under the pandemic.

Acknowledgments

I wish to thank my committee members who were more than generous with their expertise and precious time. I am extremely grateful for all their support and guidance. A special thanks to Dr. William Bridges, my committee chair and an amazing professor and mentor for his countless hours of reflecting, reading, encouraging, and most of all patience on my question no matter in professional or personal level throughout the entire process. Another special thanks to Dr. Yu-bo Wang, my co committee advisor for his amazing feedback and support during this process and countless hours of helping me improve the work, as a great mentor and guide for me as well. Thank you, Dr. Patrick Gerard, Dr. Qiong Zhang for agreeing to serve on my committee. Thank you Dr. Zhulin He for the professional advice and time during the last part of my journey.

I am also grateful for the support from Data Support Core of Prisma Health which provides me needed support, networking, and hand-on projects experiences with interdisciplinary researchers since 2018. In particular, I want to thank Dr. Alex Ewing as my supervisor during the entire internship program for his support, mentor and suggestions.

I would like to acknowledge and thank my school division for allowing me to conduct my research and providing any assistance requested. Special thanks go to the staff members for their continued support.

Finally, I want to thank all of those people who, once upon a time, were my professors and who helped me begin this long journey. Their names are too numerous to mention, but many of them built the foundation for me, inspired me to continue learning and sharing with others.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The study focuses on estimating treatment effect in causal inference based on a general framework which considers adjustments both on intermediate/post-treatment variable and unmeasured cluster level covariates, or pre-treatment covariates and unmeasured cluster level covariates. Post-treatment variables are affected by pre-treatment covariates and have impact on primary outcomes; Pre-treatment covariates affect treatment assignment mechanism. Adjustment on post-treatment variables is very essential for lots of statistical analysis, especially causal inference, and they could be difficult to tackle. Adjustment on pre-treatment covariates is also one of the classical statistical methods for estimating treatment effect with observational data. Specifically, it considered and addressed a range of approaches to deal with non-experimental data or other type of data without implemented by randomized assignment rules [79]. The existence of clustering structure also often comes up in observational studies. These clusters could generate biased inference results because of the inherent structure, which means the units within clusters are "closer" to each other than units among different clusters. In many observational studies, the variables that associated with the clusters, denoted as cluster level variables, are hard to measure or not of interests. Therefore, appropriate adjustment for the clusters in these settings is necessary.

In this study, two Bayesian frameworks are proposed for two pieces of problems: (1) a comprehensive framework combining post-treatment variable and clustering with unmeasured cluster level variables, (2) a two-step Bayesian framework including pre-treatment covariates and clustering with unmeasured cluster level variables. All of approach within two frameworks make use of efficient Markov Chain Monte Carlo (MCMC) sampling algorithms. To begin, Chapter 1 introduces

motivation of the framework and gives a brief overview of the work in relevant literature.

1.1 Background

The concept of causal inference is of great interest in many different fields. Most studies in health, social and behavioral sciences try to answer causality rather than association questions [67]. Causal inference can be problematic in studies in which treatments are not randomly assigned (commonly called the golden standard of causation). Studies in which treatments are randomly assigned are often called experiments or randomized trials, and studies in which treatments are not randomly assigned are often called observational studies. In observational studies, the presence of confounding variables can make causal inference a very challenging task. For example, it has been stated that college graduates earn about twice as much as high school graduates. Does it mean that the college education "causes" an earning advantages; or people who already have earning potential advantages choose to go to college? Without randomized treatment assignment, it is very difficult to draw the proper conclusion. In situations like this, statistical methods for "causal inference" (or more precisely, for estimating an "unbiased" average treatment effects by reducing the treatment selection bias in observational studies) are often used. Since randomized treatment assignment is often not feasible in many health, social and behavioral studies, adjustment methods for causal inference have become very popular. In addition, even in the randomized experiments or clinical trials, there might exists post-treatment variables which are affected by treatment assignment and have impact on the primary outcomes. Principal stratification is often used to evaluate treatment effects with underlying strata. Propensity Score analysis is one of the most commonly used methodologies to adjust these non-randomizations to avoid biased treatment effect in an efficient and effective approach. Propensity score analysis aims to accomplish balancing the data when treatment assignment is nonignorable and then further evaluate treatment effect. There are three main treatment effect estimations in causal inference analysis which are under whole population average treatment effect (ATE), average treatment effect within the treated (ATT), or average treatment effect within the control (AUT). In principal stratification setting, the estimators of interest are focused on treatment effect within identified underlying strata. Throughout the entire study, we focus on ATE for both principal stratification and propensity score settings respectively, as the ATE is one of the most common targets in causal inference analysis, and can also be estimated under a set of reasonable

assumptions in observational studies [33].

The remaining sections in this chapter provide a sense of why principal stratification and propensity score analysis are important for fair causal effect estimation by illustrating literature reviews of these two methodologies. When estimating the treatment effect in observational studies with clustering structure, it is critical in addition to considering such structure, especially when the variables associated to clusters are hard to evaluate or not measured. There is little literature discussing approaches to handling large number of clusters, more specifically majority of the clusters are not significant effect on primary outcomes. Therefore, a general framework would be ideal to tackle both non-randomization and unmeasured cluster level variables issues under appropriate situations.

1.2 Principal Stratification for Post-Treatment Variables

In many causal analysis studies, researchers are forced to deal with post-treatment/intermediate variables which are required for solving some scientific problems with reasonable conclusions. For example, considering the relationship between headache and high temperatures, a high temperature could give you a headache, but it is more possible that a high temperature makes you sweat, loses a lot of body moisture and causes dehydration. If you then do not drink enough water to make up for it, you may get a headache. The post-treatment/intermediate variable here which truly contributes to the headache is losing body moisture. Estimation of the treatment effect without considering the post-treatment or intermediate variables could result in the wrong conclusion about the nature of the treatment impact. Principal stratification is a commonly used statistical methodology in causal inference when adjusting the treatment effect results for post-treatment variables, where standard methodologies for the treatment estimates in causal inference could not be used. The idea of principal stratification was proposed by Frangakis and Rubin back to 2002 in [21]. They aimed to adjust these post-treatment variables where estimates under standard methodologies are not causal effects. A general framework was proposed for estimating treatment effects by adjusting for post-treatment variables which produced principal effects based on principal stratification. Under their framework, principal stratification is a classification of subjects defined by combination of the post-treatment variable under each possible treatment assignment. The basic idea is to compute causal effects within identified underlying strata. Following the notation in [21], for i^{th} subject among

total sample size of N, let Z_i be the binary treatment assignment, where 0 denotes control group and 1 denotes treatment group. And a post-treatment variable $S_i(z)$ is also measured in addition to the primary outcome $Y_i(z)$ associated with the assigned treatment z, as described in Figure 1.1. Assume $S_i(z)$ is binary for simplicity to illustrate here, and it could be extend to multiple levels and even to continuous case [82]. Specifically, $S_i(1) = 1$ denote the observation follows post-treatment behavior in treated group; $S_i(1) = 0$ denote the observation does not follow post-treatment behavior in treated group; $S_i(0) = 1$ denote the observation follows post-treatment behavior in control group; and $S_i(0) = 0$ denote the observation does not follow post-treatment behavior in control group. There are four possible strata, denoted by G_i, based on the combination of potential post-treatment

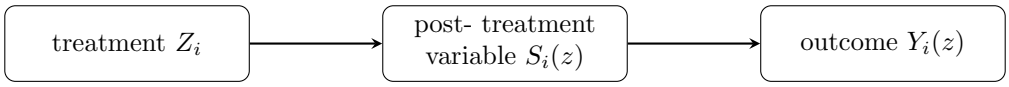


Figure 1.1: Relation of variables under principal stratification setting. Arrows represent effect relation between variables.

values $(S_i(0), S_i(1)) \in \{(1,1),(1,0),(0,1),(0,0)\}$, which is shown in Table 1.1. For $G_i = 11$ principal stratum, all subjects in this principal stratum follow post-treatment variable behavior no matter which treatment they were assigned ($S_i(0) = S_i(1) = 1$). For $G_i = 00$ principal stratum which is the opposite of $G_i = 11$ stratum, it consists of subjects who never follow post-treatment variable behavior regardless of treatment they were assigned ($S_i(0) = S_i(1) = 0$). $G_i = 10$ principal stratum consists of all subjects who follow post-treatment variable behavior only if when they were assigned to control group ($S_i(0) = 1, S_i(1) = 0$). The fourth principal stratum $G_i = 01$ consists of all subjects who follow post-treatment variable behavior only if when they were assigned to treatment group ($S_i(0) = 0, S_i(1) = 1$). The following assumptions are usually made in principal stratification setting:(1) Stable unit treatment value assumption (SUTVA) like in propensity score analysis; (2) Monotonicity, i.e. $S_i(1) \geq S_i(0)$ for any i; (3) Exclusion restriction: there is no direct effect of treatment on primary outcome, implying if $S_i(1) = S_i(0)$, then $Y_i(1) = Y_i(0)$; and finally (4) strong ignorability of assignment. A good example to illustrate principal stratification is the following. It is typically of interest to evaluate quality of patients' life in cancer study comparing treatment and the placebo groups. A direct comparison results in a paradoxical conclusion, namely that patients in placebo groups have a better quality of life than those who were in treatment group, even though there is typically evidence that the treatment should have a positive effect on quality of life. The paradoxical result is due to the effects of post-treatment variables such as patients death. Principal

stratification methodology aims to get rid of the effects from post-treatment variables and compare treatments fairly. Therefore, principal stratification plays an important role here to focus on patients with same post-treatment values, i.e., those who survived regardless of which treatment they were taken. Note that this is similar to adjustment on pre-treatment covariates using propensity score.

Table 1.1: Principal strata based on combination of potential values of $(S_i(0), S_i(1))$. $O(z, s)$ is set of subjects with assigned treatment z and observed post-treatment value s.

		Z_i=1		
		$S_i(1) = 1$	$S_i(1) = 0$	
$Z_i = 0$	$S_i(0) = 1$	G_i=11	G_i=10	$O(0, 1)$
	$S_i(0) = 0$	G_i=01	G_i=00	$O(0, 0)$
		$O(1, 1)$	$O(1, 0)$	

Frangakis and Rubin in [21] defined the treatment effects within each principal stratum as principal effects. One of the key features they claimed is that principal effects are always causal effects. They also showed the link between principal causal effects and previous common methodologies used to adjust for post-treatment variables. Researchers have started to apply principal stratification in wide range of fields. For example, Zhang and Rubin in [97] addressed the topic of "truncated by death" problem which often arises in medicine, economics, education, and many other fields. They formulated the "truncated by death" issue as principal stratification approach and derived large sample bounds for causal effects within each principal stratum. As it becomes more popular to address causal inference, there are tutorials designed to show the conceptual basis of this framework and clarify its uses and limitations. Similarly, as when adjusting for pre-treatment covariates using propensity score, principal scores are proposed where the post-treatment variable is not randomized. In Ding's paper [17], they defined principal scores as the conditional probabilities of the underlying principal stratum given covariates and weighted samples. They claimed that the proposed framework principal stratification yields robust causal inference without dependence on outcome model distribution assumptions.

Using Bayesian inference to update the conditional probabilities with more historical information also performs well in principal stratification setting. In Gallop [24] they focused on mediation analysis using principal stratification, where the mediator factors are not randomized as they are in standard mediation analysis. They also assessed the sensitivity of results to heterogeneity of the variance among each principal stratum when using Bayesian estimation procedures. A benefit of the principal stratification model is that parametric assumptions about the structural re-

lationship between the treatment assignment and the mediator are not needed. Given the conjugate prior distribution of parameters, the posterior samples can be easily generated using Gibbs sampler, especially for the probability of each unit assigned to possible principal strata. Specifically, using the Dirichlet distribution as conjugate prior for the multinomial distribution. Similar structure has been developed in many different fields. Ricciardi and Mattei [70] proposed a framework relying on principal stratification. They focused on assessing the treatment effect for longitudinal treatments under difference assumptions. Our proposed framework follows their main structure, and also adopt the Bayesian approach for inference.

There is little literature discussing the issue of principal stratification with clustering. A more generalized framework would be beneficial for both large and small number of clusters and cluster sizes.

1.3 Propensity Score Adjustment for Pre-Treatment Covariates

Propensity score analysis (PSA) is one of the most popular statistical methods of causal inference in observational studies. The propensity score works as a balancing score to adjust for the effect of the pre-treatment covariates in the treatment and control groups. PSA and several other causal inference methods have been shown to perform well and produce unbiased estimates of the average treatment effects. Following the notation in Rosenbaum and Rubin [74], let $i = 1, 2, \cdots, N$ be individual unit and N is the total sample size. The quantity to estimate is the average treatment effect, defined as $E[Y_i(1) - Y_i(0)]$, where 0 and 1 denotes two treatment assignments, $Y_i(0)$ is the potential response if i^{th} unit receives treatment 0, and $Y_i(1)$ is the potential response if i^{th} unit receives treatment 1; $E(.)$ denotes the expectation in the population. Let $Z_i = 1$ if i^{th} unit is actually assigned to treatment 1, and $Z_i = 0$ if i^{th} unit is actually assigned to treatment 0. Let $\boldsymbol{X}_i$ be a vector of pre-treatment covariates for i^{th} unit. All these $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_N^T)^T$ are measured before treatment assignment like age, gender, salary, education level, etc.

As introduced in [74], propensity score is a balancing score which is defined as the conditional probability of assigning to treated group given the covariates. It is denoted as

$$e(\boldsymbol{X}_i) = p(Z_i = 1|\boldsymbol{X}_i),$$

The actual propensity score function is almost always unknown in practice but can be estimated by the observed data. A common choice for this estimation is the logistic model. Under the Neyman-Rubin counterfactual framework, there is at most one of potential outcomes that can be observed in observational studies, i.e., $Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. Therefore, it is important to make the stable unit treatment assumption (**SUTVA**) that the potential outcomes will be same no matter which mechanism is used to assign the treatment to each unit and no matter what treatment the other units receive. Generally, treatment assignment is strongly ignorable given a set of covariates v if $\boldsymbol{(Y(1), Y(0)) \perp Z|v,\ 0 < p(Z = 1|v) < 1}$, for all v, where $\boldsymbol{Y}(.) = (Y_1(.), Y_2(.), \ldots, Y_N(.))^T$, $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_N)^T$, the superscript T denotes the transposition of a vector. For simplicity, if it holds when $\boldsymbol{v = X}$, where $\boldsymbol{X} = (\boldsymbol{X}_1^T, \boldsymbol{X}_2^T, \ldots, \boldsymbol{X}_N^T)^T$, we say the treatment assignment is strongly ignorable.

Basically, there are four main methods for using the propensity score to estimate the average treatment effect: matching, stratification, inverse probability of treatment weighting, and covariates adjustment regression. Czajka [15] demonstrated a general matching methodology which could be applied to a wide range of problems. Heckman [29] examined a two-stage methodology that improved on the classical matching method and improved bias. Lunceford [53] reviewed existing approaches and presented comparisons that provide guidance for practical use. Austin [7] conducted an empirical experiment and Monte Carlo simulations to evaluate the performance of several matching methods in medical literature. In subsequent paper, Austin [8] used time-to-event outcome to compare matching, stratification, inverse probability of treatment weighting [73] and covariates adjustment on performance of propensity score for estimating average treatment effect. There are different versions of practical guide books that provide a list of matching methods, like in [30]. Thoemmes [89] conducted a systematic literature review in social science to illustrate some common misconceptions of propensity score methods.

Moving beyond matching, more and more advanced approaches such as inverse probability of treatment weighting and covariates adjustment using the propensity score have been used in more complicated data and modeling setting. Woo [94] evaluated the use of generalized additive models (GAMs) for propensity score estimation with traditional logit models. They found that GAMs can improve the overall covariances balance. The methodology has been proposed for more general cases. Arpino [6] presented the multilevel models for estimation of the propensity score with hierarchical structure and unobserved cluster level variables. Li [44] showed that exploiting

the multilevel structure can greatly reduce bias. Thoemmes [90] proposed several modeling choices (using fixed and random effects) to extend propensity score analysis to cluster level data. The applied example showed some practical limitations when cluster sizes are small. Li [46] proposed an augmented version of a matching weight estimator to obtain a double robust property if either the propensity score model or the outcome model is correct. Papadogeorgou [63] developed a new method termed as Distance Adjusted Propensity Score Matching (DAPSm) on spatially-indexed data. Instead of binary outcome only, there are also approaches proposed to deal with continuous, health-related life outcome [14][95].

Since machine learning techniques have expanded rapidly, they also play a role in propensity score analysis. Some machine learning techniques like classification trees, neural network, recursive partitioning have been discussed to improve the propensity score estimations and weighting [38][41][51][83]. Some more general and practical questions have been asked and explored. Zhao [99] proposed a random forest based matching method on the propensity score. It also illustrated that this proposed method can efficiently deal with missing data in covariates. Austin [9] examined the propensity score model under complex survey data. Partially observed covariates are also an issue that need to be addressed in propensity score analysis [43]. Research has also been conducted on merging observational data sets with experimental data using propensity score methods [75] and machine learning techniques.

Bayesian inference for propensity score analysis was discussed by McCandless, Gustafson, and Austin [58] to explicitly acknowledge the value of Bayesian methods in calibrating uncertainties in propensity scores and attempting to develop a "practical" Bayesian procedure for that purpose. Traditionally, there are two steps propensity score analysis, estimation the propensity score first, and implementing outcome analysis second. However, the second step usually ignores the uncertainty involved in the propensity score estimation. Therefore, researchers have proposed propensity score regression based on a Bayesian framework for fitting the treatment and outcome model simultaneously. Given prior distributions, $\beta \sim N(0, \sigma_\beta^2)$, $\gamma \sim N(0, \sigma_\gamma^2)$, $\boldsymbol{\zeta} \sim MVN(0, \sigma_{\boldsymbol{\zeta}}^2)$, where the treatment assignment model is logistic regression $\widehat{e(\boldsymbol{X}_i)} = p(\widehat{Z_i = 1}|\boldsymbol{X}_i) = \frac{1}{1+\exp(-\boldsymbol{X}_i^T\boldsymbol{\zeta})}$ and the outcome model is simple linear regression considering the estimated propensity as a additional covariate, $Y_i = \beta \times Z_i + \gamma \times \widehat{e(\boldsymbol{X}_i)} + N(0,1)$. Then the joint distribution of the data and the parameters is written as $p(\beta, \boldsymbol{\zeta}, \gamma \mid Z_i, \boldsymbol{X}_i) \propto \prod_{i=1}^n p(Y_i \mid Z_i, \boldsymbol{X}_i, \beta, \boldsymbol{\zeta}, \gamma) p(Z_i \mid \boldsymbol{X}_i, \boldsymbol{\zeta}) p(\beta) p(\boldsymbol{\zeta}) p(\gamma)$. Posterior sampling from the posterior density $p(\beta, \boldsymbol{\zeta}, \gamma \mid Z_i, \boldsymbol{X}_i)$ using Markov Chain Monte Carlo (MCMC)

Gibbs sampling or Metropolis-Hastings algorithm to update successively from $p(\gamma \mid \beta, \boldsymbol{\zeta}, \boldsymbol{Z}, \boldsymbol{X})$, $p(\beta, \boldsymbol{\zeta} \mid \gamma, \boldsymbol{Z}, \boldsymbol{X})$ and $p(\boldsymbol{\zeta} \mid \beta, \gamma, \boldsymbol{Z}, \boldsymbol{X})$. Similarly for propensity score matching, a semi-parametric Bayesian framework is needed because of non-parametric matching step.

Joint Bayesian methods [3, 104] have been recently introduced to estimate causal effects with a pre-specified $\alpha = \alpha_0$ ([58, 103]) to model both the propensity score and the outcome in one step, and then use the uncertainty in causal inference. The traditional sequential estimation does not attempt to recover the whole relationship between pre-treatment covariates and the outcome. Without specific parametric modeling of this relationship, the joint Bayesian propensity score estimation framework considers the presence of feedback on the propensity score from the outcome information. This means that the outcome model formulation affects the estimation of parameters of the propensity score. Specifically, the propensity score parameters appear as terms of both parts the likelihood which allows for the issue of model feedback. They use the Metropolis-Hasting MCMC algorithm to generate posterior samples and thoroughly evaluate the model feedback under different relationships and showed that the model feedback can produce poor estimates of the treatment causal effects under certain conditions.

1.4 Causal Inference with Clustering Structure

Causal inference in observational studies can become very challenging with issues like missing values, sampling error, unknown probability distributions, etc. One commonly encountered issue is related to the complicated data set design, more specifically, a clustering structure whereby the units within a cluster are more "similar" than units from different clusters. However, in practice, the variables that actually define the clusters (often denoted as cluster level variables) are often not observed, either because the variables are not of the interest in the original study design, or the information about the variables is difficult to observe or measure. Estimating the causal effect adjusting for the cluster level variables is an issue since they are often not found in observed data itself, and even if they are, the actual relationship between unmeasured cluster level covariates and primary outcome is usually unknown [66]. Adjustments for the simple cluster indicators in these settings can be more problematic as discussed below.

For example, given the natural groupings of students within classrooms and schools in educational settings, the effect of classrooms (as clusters) impacts the educational performance of

individual students. Obviously, the cluster level variables to consider are those associated with the teacher of the class. Teachers have unique teaching styles and engagement skills, which are typically measured as a combination of instruction type, classes content, learning support, experience, educational background, etc. [19, 27, 32]. Certain teaching styles can lead to certain students acquiring more knowledge. However, many of these cluster level variables are not observed, or hard to measure in the data collection process. Another issue is that even though some of these cluster level variables are observed, the true nature of the relationship between these variables and the outcome is often difficult to model.

Even though there are issues associated with clusters, they should be carefully considered because the variability within and among clusters can be associated with the outcome; and if they are ignored, the treatment effect estimation might be biased. Arpino and Cannas [5] focused on the implementation of propensity score matching for clustered data. They investigated two methods of exploiting the clustered data structure, one is the estimation of the propensity score model (including fixed or random effects) and the other is the implementation of the matching algorithm. They concluded that both approaches successfully reduced the bias due to the omission of cluster level variables. The combined within-cluster and between cluster matching approach most performed the best in both the large and small clusters. The inverse probability weighting is an important tool in the construction of survey weights which is widely used to correct for the potential biasing impact of correlated nonresponse within clusters as discussed by Skinner and Arrigo [85]. Their proposed weighted estimator based upon conditional logistic regression can avoid the bias which arises when traditional inverse probability weights are considered. The extension of their estimator to observational studies with clustered treatment assignment are also described. Liao et al. [49] used cluster analysis in large healthcare claims databases to help reveal hidden structures and "clusters" especially when the distribution of the expenditure data was severely skewed. More recent works also addressed the importance of clustering structure in causal inference and proposed new robust methods for treatment effect estimation. Yang [96] proposed a calibration technique and provided a consistent propensity score weighting estimator of the ATE when the propensity score and outcome follow generalized linear mixed effects models, which does not require specification of functional forms of the propensity score and outcome models. He [28] proposed an inverse conditional probability weighting method by making use of sufficient statistics, which are robust both to correlation between the unmeasured cluster level covariates and unit level covariates and the correlation between the

unmeasured cluster level covariates and the outcome.

1.5 Dissertation Organization

The dissertation is organized as follows. Chapter 2 addresses the problem of estimating the treatment effect when both the post-treatment variable and clusters exist in an observational study. Chapter 3 focuses on the problem of estimating the treatment effect when the number of observations is very unbalanced between the two treatment groups and clusters exist in an observational study. Both chapters carefully derive methodologies based on Bayesian conditional posterior distributions, evaluate the methodologies using empirical simulation studies, and then implement in the methods using real world data applications. Chapter 4 summarizes the novel contributions of this study, provides overall conclusions based on the proposed framework of this study and some future research directions.

Chapter 2

Bayesian Framework for Causal Inference with Principal Stratification and Clusters

2.1 Abstract

In observational studies, principal stratification is a well-established method in causal analysis to adjust the treatment effect estimation for post-treatment variables. However, this inference could be challenging when the data have a clustering structure, which is pervasive in observational studies. Adding to the issues is the fact that often the variables associated with the clusters are only recorded as the cluster label due to a budget constraint or measuring difficulties. Furthermore, the true nature of the relationship between these cluster level variables and the outcome may be unclear. Although accommodating this clustering structure via random effects based on the cluster label can address the bias issues, estimating the model is inevitably tedious and overfitting can occur with principal stratification and clustering. In this article, we propose a comprehensive framework for estimating a treatment effect when both post-treatment variable and clustering exist in a data set. Specifically, following the idea of principal stratification, we define the clustering structure as random effects with a spike and slab prior in a Bayesian hierarchical model. As a result, a parsimonious model which only contains clusters with significant effects on the outcome can be obtained without

much computational cost. We demonstrate the desirable features of the proposed method with two real data sets, one about academic performance and the other about infant birth weight. To further examine the empirical performance of the proposed method, simulations with data generating mechanisms similar to our data applications, and other four hypothetical data sets are conducted.

2.2 Introduction

In observational studies, causal analysis is an effective way to adjust for the influence of covariate imbalance on the estimated treatment effect. For example, the propensity score [74] is a common method to balance the systematic differences between control and treated subjects, and make the treatment comparison more meaningful. When causal analysis is further complicated by the existence of post-treatment or intermediate variables that can have an impact on the outcome of interest, then principal stratification [21] serves as a useful tool to adjust the estimated treatment effect [22, 36, 48, 76, 80, 97, 98]. Essentially, it identifies the underlying strata of each observation based on the potential values of a post-treatment variable and computes the treatment effects only within the strata, where the pairs of post-treatment values under two arms can be controlled at certain levels. This approach can also address selection bias assuming the latent ignorability assumption of treatment assignment is satisfied [70].

As part of this research, we investigate how particular types of classrooms (single-gender or mixed-gender) can affect the educational performance of students, where the class size is a post-treatment variable reflecting the educational resource allocated to each individual. In another application, we investigate the effect of maternal exposure to any risk factors relevant to hypertension on birth weight. Based on the recent research by Terade et al. [88] that suggests an association between birth weight and induction of labor, we consider whether labor is induced as the post-treatment variable.

Another common challenge in causal analysis is the omnipresence of clusters of observations. The variables that define the clusters (collectively denoted as cluster level variables) only contain labeling information due to measuring difficulties or beyond the research scope. Or even if they are fully observed, the relationship between these variables and the outcome may remain unclear. In such cases, the clustering structure should be incorporated into the models as random effects to yield unbiased treatment effect estimation [5, 85, 90, 96]. In this work, driven by our data applications,

we focus on the issues in causal analysis when both post-treatment and cluster level variables are present.

In the education study, we consider "class" as the clusters, in that the teacher and the other students in the class have an impact on educational performance. Often the cluster effects are recorded as a "cluster label" instead of all the variables actually defining the clusters. The impact of the teacher assigned to each class is actually an integration of the teacher's educational background, unique experiences, teaching style, and engagement skills etc. There are also additional variables that define the impact of the other students. All these variables are typically not measured due to budget constraints, or they are beyond the research scope. Therefore, the label is used in analysis. Furthermore, in the birth weight study, even if the maternal factors (age and body mass index (BMI)) are measured, the true biological mechanism between these factors and the outcome is difficult to model. Consequently, the clusters formed by maternal age and BMI are used to replace the traditional analysis approaches, where age and BMI are treated as covariates.

Though the aforementioned random effect model can be used to incorporate the principal stratification and the clustering structure, this type of model is usually tedious, and overfitting can occur because of the introduced latent strata. To incorporate the clusters with significant effects and achieve model parsimony, we follow the Bayesian framework of principal stratification with random effects. We propose a novel approach of using a spike and slab prior for the hierarchical structure of random effects. With such a sparsity inducing prior, the selection of cluster effects can be conducted in a single Markov chain Monte Carlo (MCMC) sample. Specifically, based on the posterior inclusion probabilities (PIP), a final model with only significant cluster effects can be determined. To facilitate the posterior computations, an efficient MCMC algorithm is also developed via a series of data augmentations (DA).

The rest of this chapter is organized as follows. Section 2.3 introduces the proposed method along with the assumptions and prior settings. Section 2.4 provides the detailed steps of data augmentation and posterior sampling algorithms. In Section 2.5, six simulation studies are conducted to compare the empirical performance with three other competing methods. In Section 2.6, we apply the proposed method to two examples: academic performance of two middle schools in the upstate of South Carolina and infant birth weight in US territories.

2.3 Methodology

2.3.1 Notations and Assumptions

Assuming the presence of clusters, we let Z_{ij} denote the binary treatment assignment ($Z_{ij} = 1$ for treated and $Z_{ij} = 0$ for control) of the j^{th} subject in the i^{th} cluster. $\boldsymbol{X}_{ij}$ and U_i are the corresponding p-dimensional vectors of the pre-treatment covariates and cluster level variable, respectively, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$, and $N = \sum_i^m n_i$ is the total sample size. Assuming the presence of post-treatment variable, let $S_{ij}(z)$ and $Y_{ij}(z)$ separately denote the binary post-treatment variable and the continuous outcome of interest associated with the assigned treatment z, respectively. Under the Rubin causal model [31] and principal stratification framework [21], let $(S_{ij}(0), S_{ij}(1))$ and $(Y_{ij}(0), Y_{ij}(1))$ represent for the two sets of potential outcomes. The average treatment effect (ATE) on the primary outcome is the focal causal effect estimand defined as

$$\text{ATE}_{z,z'} = \text{E}[Y_{ij}(z) - Y_{ij}(z')], \quad \text{for } z \neq z'. \tag{2.1}$$

Although (2.1) is intuitive and well interpreted, it is impossible to observe both $Y_{ij}(z)$ and $Y_{ij}(z')$ at the same time since each subject is only exposed to a single treatment. Additionally, as shown in Figure 2.1, the treatment effect on the outcome is intermediated through the post-treatment variable, and the cluster level variable can affect (or relate to) the outcome and the other variables in Figure 2.1. Therefore, the observed Y_{ij}'s require some adjustments before being used to evaluate the causal effect. To this end, we present the necessary assumptions below for this counterfactual framework, and extend the principal stratification approach [21] to accommodate data with the clusters.

Similar to other causal inferences approaches, we require the stable unit treatment value assumption (SUTVA) and weaker latent ignorability (LI) of treatment assignment to make (2.1) estimable (i.e., no interference among units under one well-defined treatment and control). Notationally, these assumptions can be expressed as

$$(Y_{ij}(0), Y_{ij}(1)) \perp Z_{ij} \mid S_{ij}(0), S_{ij}(1), X_{ij}, U_i,$$
$$(S_{ij}(0), S_{ij}(1)) \perp Z_{ij} \mid X_{ij}, U_i,$$

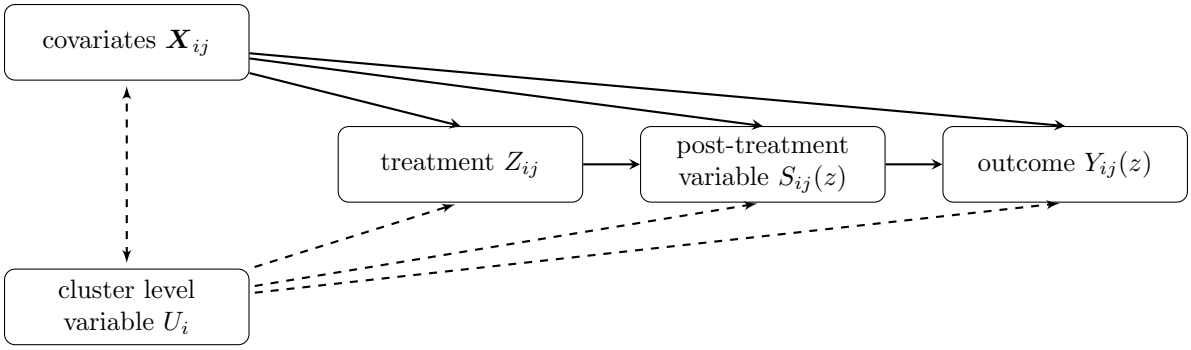


Figure 2.1: Graphical representation of the possible relationships among the observed variables and the cluster level variable. Solid lines between variables denote observed relationships, while dashed lines between variables denote unobserved possible relationships.

for all i and j, respectively. Note that the proposed LI is claimed as the weaker condition due to $(Y_{ij}(0), Y_{ij}(1)) \perp Z_{ij} \mid S_{ij}(0), S_{ij}(1), X_{ij} \Rightarrow (Y_{ij}(0), Y_{ij}(1)) \perp Z_{ij} \mid S_{ij}(0), S_{ij}(1), X_{ij}, U_i$ and $(S_{ij}(0), S_{ij}(1)) \perp Z_{ij} \mid X_{ij} \Rightarrow (S_{ij}(0), S_{ij}(1)) \perp Z_{ij} \mid X_{ij}, U_i$, but not vice versa. Now, following these two assumptions, the joint likelihood function of potential outcomes is given by

$$
\begin{aligned}
& f(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{S}(1), \boldsymbol{S}(0) \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\eta}) \\
=& f(\boldsymbol{S}(1), \boldsymbol{S}(0) \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\eta}) f(\boldsymbol{Y}(1), \boldsymbol{Y}(0) \mid \boldsymbol{S}(1), \boldsymbol{S}(0), \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\eta}) \\
=& \prod_{i,j} f(S_{ij}(1), S_{ij}(0) \mid Z_{ij}, \boldsymbol{X}_{ij}, U_i, \boldsymbol{\eta}) f(Y_{ij}(1), Y_{ij}(0) \mid S_{ij}(1), S_{ij}(0), Z_{ij}, \boldsymbol{X}_{ij}, U_i, \boldsymbol{\eta}) \qquad \text{(by SUTVA)} \\
=& \prod_{i,j} f(S_{ij}(1), S_{ij}(0) \mid \boldsymbol{X}_{ij}, U_i, \boldsymbol{\eta}) f(Y_{ij}(1), Y_{ij}(0) \mid S_{ij}(1), S_{ij}(0), \boldsymbol{X}_{ij}, U_i, \boldsymbol{\eta}), \qquad \text{(by LI)}
\end{aligned}
$$

where $\boldsymbol{Y}(.) = (Y_{11}(.), Y_{12}(.), \ldots, Y_{mn_m}(.))^T$, $\boldsymbol{S}(.) = (S_{11}(.), S_{12}(.), \ldots, S_{mn_m}(.))^T$, $\boldsymbol{Z} = (Z_{11}, Z_{12}, \ldots, Z_{mn_m})^T$, $\boldsymbol{X} = (\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{mn_m})^T$, $\boldsymbol{U} = (U_1, U_2, \ldots, U_m)^T$, the superscript T denotes the transposition of a vector, and $\boldsymbol{\eta}$ are all parameters for this conditional structure. Proceeding in this fashion, let $Y_{ij}^{obs} = Z_{ij}Y_{ij}(1) + (1 - Z_{ij})Y_{ij}(0)$ denote the observed outcome of the j^{th} subject in the i^{th} cluster, and $Y_{ij}^{mis} = (1 - Z_{ij})Y_{ij}(1) + Z_{ij}Y_{ij}(0)$ denote the missing outcome of the j^{th} subject in the i^{th} cluster. Similarly, $S_{ij}^{obs} = Z_{ij}S_{ij}(1) + (1 - Z_{ij})S_{ij}(0)$ denotes observed post-treatment value, and $S_{ij}^{mis} = (1 - Z_{ij})S_{ij}(1) + Z_{ij}S_{ij}(0)$ denotes missing post-treatment value. Then, the posterior distribution of $\boldsymbol{\eta}$ given the observed data likelihood is proportional to

$$\pi(\boldsymbol{\eta} \mid \boldsymbol{Y}^{obs}, \boldsymbol{S}^{obs}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{U})$$

$$\propto \pi(\boldsymbol{\eta}) \times \int_{i,j} \prod_{i,j} f(S_{ij}(1), S_{ij}(0) \mid \boldsymbol{X}_{ij}, U_i, \boldsymbol{\eta}) f(Y_{ij}(1), Y_{ij}(0) \mid S_{ij}(1), S_{ij}(0), \boldsymbol{X}_{ij}, U_i, \boldsymbol{\eta}) dY_{ij}^{mis} dS_{ij}^{mis}, \tag{2.2}$$

where $\boldsymbol{Y}^{obs} = (Y_{11}^{obs}, Y_{12}^{obs}, \ldots, Y_{mn_m}^{obs})^T$, $\boldsymbol{S}^{obs} = (S_{11}^{obs}, S_{12}^{obs}, \ldots, S_{mn_m}^{obs})^T$, and $\pi(\boldsymbol{\eta})$ is the prior distribution of $\boldsymbol{\eta}$. A few comments are warranted. First, with the LI assumption, both conditional structures are free from the treatment assignment [21, 57, 70]. As a result, the correction of selection bias is not necessary. Secondly, since a series of integration over S_{ij}^{mis}, Y_{ij}^{mis} in (2.2) is usually analytically intractable [22, 36], the data augmentation strategy [87] is suggested to tackle this computational difficulty.

2.3.2 Model and Likelihood Function

Following the aforementioned data augmentation strategy, and based on the binary post-treatment variable, propose the principal stratification approach [21] to address the estimation problem in (2.2). Specifically, they define the principal strata $G_{ij} = g, g \in \{1, 2, 3, 4\}$ separately for the combinations of $(S_{ij}(0), S_{ij}(1)) \in \{(1,1), (1,0), (0,1), (0,0)\}$, and let $O(z, s)$ denote the set of subjects with the assigned treatment z and observed post-treatment value s. Accordingly, each set consists of two out of four latent strata as illustrated in Table 2.1 (for example, $O(1,1)$ is a mixed group of strata $g = 1$ and $g = 3$), and (2.2) can be re-written as

$$\begin{aligned}
&\pi(\boldsymbol{\eta} \mid \boldsymbol{Y}^{obs}, \boldsymbol{S}^{obs}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{U}) \propto \pi(\boldsymbol{\eta}) \times \\
&\prod_{i,j \in O(1,1)} \left\{ p_{ij}(1) f_{ij,1}^{1}(Y_{ij}^{obs}) + p_{ij}(3) f_{ij,3}^{1}(Y_{ij}^{obs}) \right\} \times \prod_{i,j \in O(1,0)} \left\{ p_{ij}(2) f_{ij,2}^{1}(Y_{ij}^{obs}) + p_{ij}(4) f_{ij,4}^{1}(Y_{ij}^{obs}) \right\} \times \\
&\prod_{i,j \in O(0,1)} \left\{ p_{ij}(1) f_{ij,1}^{0}(Y_{ij}^{obs}) + p_{ij}(2) f_{ij,2}^{0}(Y_{ij}^{obs}) \right\} \times \prod_{i,j \in O(0,0)} \left\{ p_{ij}(3) f_{ij,3}^{0}(Y_{ij}^{obs}) + p_{ij}(4) f_{ij,4}^{0}(Y_{ij}^{obs}) \right\},
\end{aligned} \tag{2.3}$$

where $p_{ij}(g)$ denotes the probability that the subject is from the g^{th} stratum, and $f_{ij,g}^{z}(Y_{ij}^{obs})$ is the conditional probability density function of the primary potential outcome. To further model the four principal strata and consider the existence of cluster level variable such that the effect $h_g(U_i)$ on each stratum is unobserved (where $h_g(.)$ is an intractable function due to either unavailable U_i or unclear relationship between U_i and the g^{th} stratum), we incorporate the cluster-specific intercept terms

Table 2.1: Principal strata g based on combination of potential values of post-treatment variable $(S_{ij}(0), S_{ij}(1))$, where $g = 1$ is (1,1), $g = 2$ is (1,0), $g = 3$ is (0,1) and $g = 4$ is (0,0). $O(z, s)$ is set of subjects with assigned treatment z and observed post-treatment value s.

		Z_{ij}=1		
		$S_{ij}(1) = 1$	$S_{ij}(1) = 0$	
$Z_{ij} = 0$	$S_{ij}(0) = 1$	g=1	g=2	$O(0, 1)$
	$S_{ij}(0) = 0$	g=3	g=4	$O(0, 0)$
		$O(1, 1)$	$O(1, 0)$	

$\zeta_{i,1}$, $\zeta_{i,2}$, and $\zeta_{i,3}$ into a series of underlying structures [70] to quantify these effects. Specifically, we have

$$
\begin{aligned}
&G_{ij} = 1, \quad \text{if} \quad G^*_{ij,1} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \zeta_{i,1} + \epsilon_{ij,1} \leq 0,\\
&G_{ij} = 2, \quad \text{if} \quad G^*_{ij,1} > 0 \quad \text{and} \quad G^*_{ij,2} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_2 + \zeta_{i,2} + \epsilon_{ij,2} \leq 0,\\
&G_{ij} = 3, \quad \text{if} \quad G^*_{ij,1} > 0 \quad \text{and} \quad G^*_{ij,2} > 0 \quad \text{and} \quad G^*_{ij,3} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_3 + \zeta_{i,3} + \epsilon_{ij,3} \leq 0,\\
&G_{ij} = 4, \quad \text{otherwise},
\end{aligned}
$$

where $G^*_{ij,1}$, $G^*_{ij,2}$, and $G^*_{ij,3}$ are the set of latent variables jointly determining the stratum of each subject, and $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$ are the corresponding p-dimensional vectors of regression coefficients while $\epsilon_{ij,1}$, $\epsilon_{ij,2}$, and $\epsilon_{ij,3}$ are the error terms independently following $N(0, 1)$. Under such a setting, $p_{ij}(g)$ is given by

$$
\begin{aligned}
p_{ij}(1) &= 1 - \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \zeta_{i,1}),\\
p_{ij}(2) &= \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \zeta_{i,1}) \left[1 - \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_2 + \zeta_{i,2})\right],\\
p_{ij}(3) &= \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \zeta_{i,1})\Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_2 + \zeta_{i,2}) \left[1 - \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_3 + \zeta_{i,3})\right],\\
p_{ij}(4) &= \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \zeta_{i,1})\Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_2 + \zeta_{i,2})\Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_3 + \zeta_{i,3}),
\end{aligned}
\tag{2.4}
$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. Similarly, we consider $\xi_{i,g}$ for the unmeasured-cluster effect $q_g(U_i)$ on each stratified outcome of interest as follows

$$
Y_{ij}(z) \mid g, \boldsymbol{X}_{ij}, U_i \sim N(\boldsymbol{X}_{ij}^T\boldsymbol{\alpha}_g + \gamma_g z + \xi_{i,g}, \sigma_g^2), \tag{2.5}
$$

where $\boldsymbol{\alpha}_g$ and γ_g are the the regression coefficients, and σ_g^2 is the corresponding variance. We then can estimate (2.1) via

$$\widehat{\text{ATE}}_{z,z'} = \frac{1}{N}\sum_i \sum_j \sum_g \hat{\gamma}_g (z - z')\hat{p}_{ij}(g), \quad \text{for } z \neq z', \tag{2.6}$$

where $\hat{\gamma}_g$ and $\hat{p}_{ij}(g)$ are the posterior means of γ_g and the posterior predictive distribution $p_{ij}(g)$, respectively.

2.3.3 Prior Specifications

Traditionally, $\zeta_{i,l}$ for $l = 1, 2, 3$, and $\xi_{i,g}$ for $g = 1, 2, 3, 4$ in (2.4) and (2.5) are treated as random effects (equivalent to the Gaussian priors in the Bayesian framework) to avoid the so-called Neyman-Scott incidental parameter problem [44] when the number of clusters (m) is relatively large compared to the cluster size (n_i). Nevertheless, the model complexity derived from the increment of m remains unmitigated. Additionally, motivated by the assumption that all except few cluster effects are ignorable, we adopt the sparsity-inducing prior to efficiently filter out ignorable ones and reach model parsimony. Specifically, instead of the Gaussian prior, we consider the spike and slab prior [59] for each $\zeta_{i,l}$ and $\xi_{i,g}$ as follows

$$\begin{aligned}
&\zeta_{i,l} \mid v_{i,l} \sim (1 - v_{i,l})N(0, \delta\sigma_\zeta^2) + v_{i,l}N(0, \sigma_\zeta^2), \qquad \text{for } l = 1, 2, 3,\\
&v_{i,l} \mid p_{v_{i,l}} \sim \text{Ber}(p_{v_{i,l}}),\\
&p_{v_{i,l}} \sim \text{Beta}(a_v, b_v),\\
&\sigma_\zeta^2 \sim \text{InvGamma}(a_\zeta, b_\zeta),\\
&\xi_{i,g} \mid w_{i,g} \sim (1 - w_{i,g})N(0, \delta\sigma_\xi^2) + w_{i,g}N(0, \sigma_\xi^2), \qquad \text{for } g = 1, 2, 3, 4,\\
&w_{i,g} \mid p_{w_{i,g}} \sim \text{Ber}(p_{w_{i,g}}),\\
&p_{w_{i,g}} \sim \text{Beta}(a_w, b_w),\\
&\sigma_\xi^2 \sim \text{InvGamma}(a_\xi, b_\xi),
\end{aligned} \tag{2.7}$$

where $v_{i,l}$ (or $w_{i,g}$) is an indicator variable with the value 1 denoting a nonignorable cluster effect from the slab distribution $N(0, \sigma_\zeta^2)$ (or $\xi_{i,g} \sim N(0, \sigma_\xi^2)$ when $w_{i,g} = 1$) while 0 stands for an ignorable one from the spike component $N(0, \delta\sigma_\zeta^2)$ (or $\xi_{i,g} \sim N(0, \delta\sigma_\xi^2)$ when $w_{i,g} = 0$), $p_{v_{i,l}}$ and $p_{w_{i,g}}$ are the

prior inclusion probabilities, $a_v, b_v, a_\zeta, b_\zeta, a_w, b_w, a_\xi$, and b_ξ are pre-specified hyperparameters, and δ is a pre-specified small value. Here, we follow the suggestion in [81] and set δ as 0.00025 for both simulation studies and data applications. Note that considering no common intercept term in the proposed model, we select the normal mixture inverse-gamma (NMIG) prior [34] in (2.7). This is because compared with a Dirac delta function [56, 91] that may yield exact zero-valued intercepts, NMIG is a relatively smoothing setting and could better accommodate these cluster-and-intercept mixed terms. Besides, unlike the stochastic search variable selection (SSVS) proposed by George et al. [25], it does not require prior knowledge to specify σ_ζ^2 and σ_ξ^2.

It also worth pointing out that although the spike and slab prior can efficiently explore the entire cluster sample space, and select those nonignorable cluster effects simultaneously in the estimation procedure, the ATE calculation still requires an additional MCMC sample from the final chosen model to evaluate. Therefore, we adopt the two-step procedure to attain the ATE estimation: first, based on the average of binary switch $p_{v_{i,l}}$ (or $p_{w_{i,g}}$) in the MCMC chain (i.e., the posterior inclusion probabilities (PIP)), we select all effects with PIPs at least 0.5 to include in the final model. Accordingly, with the assigned fix values of $v_{i,l}$ and $w_{i,g}$ in each iteration, the same sampling scheme is used to generate another MCMC sample for (2.6).

For other parameters in the proposed model, we consider the following conjugate priors

$$
\begin{aligned}
&\boldsymbol{\beta}_l \mid \sigma_{\beta_l}^2 \sim N(\mathbf{0}, \sigma_{\beta_l}^2 I_p), \quad \text{for } l = 1, 2, 3, \\
&\sigma_{\beta_l}^2 \sim \text{InvGamma}(a_\beta, b_\beta), \\
&\boldsymbol{\alpha}_g \mid \sigma_{\alpha_g}^2 \sim N(\mathbf{0}, \sigma_{\alpha_g}^2 I_p), \quad \text{for } g = 1, 2, 3, 4, \\
&\sigma_{\alpha_g}^2 \sim \text{InvGamma}(a_\alpha, b_\alpha), \\
&\gamma_g \sim N(0, \sigma_{\gamma_g}^2), \\
&\sigma_{\gamma_g}^2 \sim \text{InvGamma}(a_\gamma, b_\gamma), \\
&\sigma_g^2 \sim \text{InvGamma}(a_0, b_0),
\end{aligned}
$$

where I_p is a $p \times p$ identity matrix, and $a_\beta, b_\beta, a_\alpha, b_\alpha, a_\gamma, b_\gamma, a_0, b_0$ are pre-specified hyperparameters.

2.4 Data Augmentation and Bayesian Inference

A series of data augmentation (DA) is performed to facilitate posterior computation. First, the DA algorithm in [87] is applied to the latent strata to bypass the mixture structures in (2.3). Specifically, let "| ." denote "given all the other parameters and observed data". We have G_{ij} updated in each iteration via the following manner: for $i, j \in O(1,1)$,

$$\begin{cases} P(G_{ij} = 1 \mid .) = \frac{p_{ij}(1) f^1_{ij,1}(Y^{obs}_{ij})}{p_{ij}(1) f^1_{ij,1}(Y^{obs}_{ij}) + p_{ij}(3) f^1_{ij,3}(Y^{obs}_{ij})}, \\ P(G_{ij} = 3 \mid .) = 1 - P(G_{ij} = 1 \mid .); \end{cases}$$

for $i, j \in O(1,0)$,

$$\begin{cases} P(G_{ij} = 2 \mid .) = \frac{p_{ij}(2) f^1_{ij,2}(Y^{obs}_{ij})}{p_{ij}(2) f^1_{ij,2}(Y^{obs}_{ij}) + p_{ij}(4) f^1_{ij,4}(Y^{obs}_{ij})}, \\ P(G_{ij} = 4 \mid .) = 1 - P(G_{ij} = 2 \mid .); \end{cases}$$

for $i, j \in O(0,1)$,

$$\begin{cases} P(G_{ij} = 1 \mid .) = \frac{p_{ij}(1) f^0_{ij,1}(Y^{obs}_{ij})}{p_{ij}(2) f^0_{ij,2}(Y^{obs}_{ij}) + p_{ij}(1) f^0_{ij,1}(Y^{obs}_{ij})}, \\ P(G_{ij} = 2 \mid .) = 1 - P(G_{ij} = 1 \mid .); \end{cases}$$

for $i, j \in O(0,0)$,

$$\begin{cases} P(G_{ij} = 3 \mid .) = \frac{p_{ij}(3) f^0_{ij,3}(Y^{obs}_{ij})}{p_{ij}(3) f^0_{ij,3}(Y^{obs}_{ij}) + p_{ij}(4) f^0_{ij,4}(Y^{obs}_{ij})}, \\ P(G_{ij} = 4 \mid .) = 1 - P(G_{ij} = 3 \mid .). \end{cases}$$

Once the stratum memberships are determined, all full conditionals of parameters related to (2.5) have conjugate forms under our prior settings. To attain the MCMC sample of other parameters associated with (2.4), another DA technique, as shown in Section 2.2, is implemented, where $G^*_{ij,l}$'s are introduced to rewrite the probit models as the underlying normal structures [2]. Proceeding in this fashion, all remaining parameters including the latent variables can be sampled by the Gibbs sampler. We provide the detailed steps of posterior computation in Appendix A.

2.5 Simulations

2.5.1 Simulation Settings

To evaluate the performance of the proposed method, we conduct the simulation studies under the following six scenarios: (i) $m = 50$, $n_i = 2$; (ii) $m = 250$, $n_i = 2$; (iii) $m = 500$, $n_i = 2$; (iv) $m = 50$, n_i uniformly ranges from 6 to 15; (v) $m = 32$, n_i ranges from 8 to 24; and (vi) $m = 99$, n_i ranges from 1 to 32. The first three scenarios focus on small sample size problems with different numbers of cluster. Scenario (iv) is designed for a direct comparison to scenario (i) to evaluate the impact of sample size per cluster. The last two scenarios are designed to mimicking the applications discussed in Section 5.

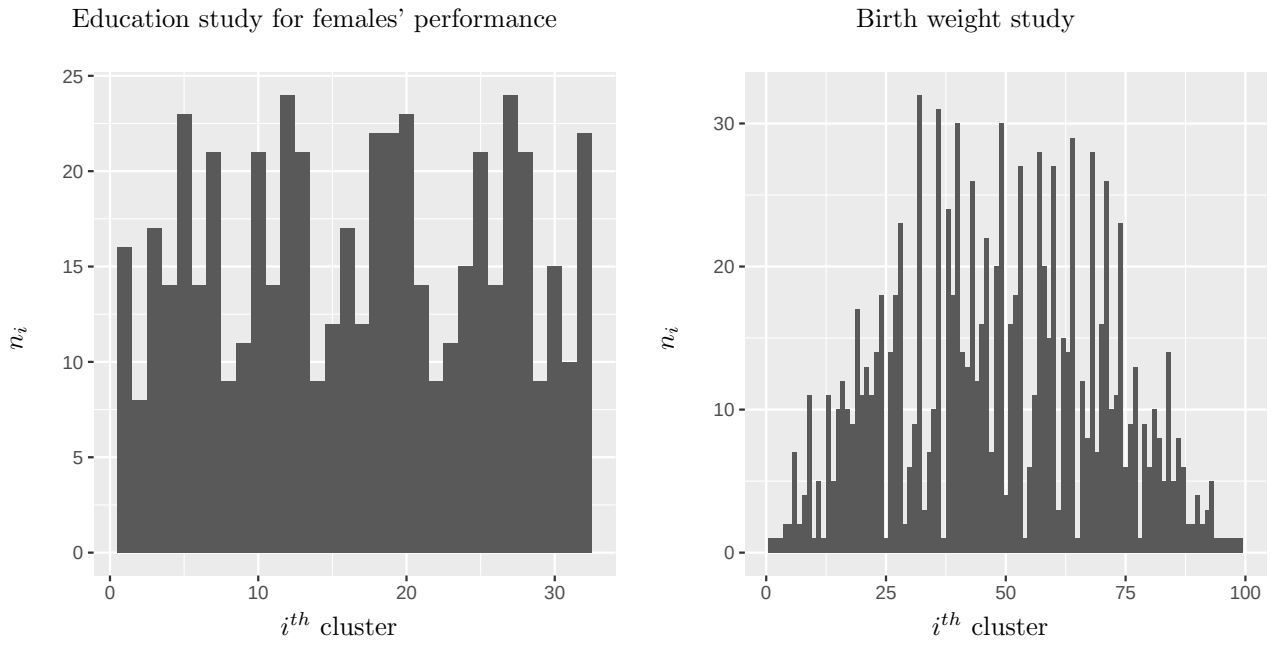
In (i)-(iv), we include two pre-treatment covariates $\boldsymbol{X}_{ij}$ that are generated from the multivariate normal distribution with mean vector $(0, 2)^T$ and covariance matrix $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, and let the treatment assignment Z_{ij} follow a Bernoulli distribution with probability of success equal to 0.5. To allow for cluster effects, we have 60% and 40% of U_i generated from $N(0, 0.1^2)$ and $N(3, 1^2)$, respectively, and assume $h_l(U_i) = \beta_{l,u}U_i$ and $q_g(U_i) = \alpha_{g,u}U_i$, where the values of $\beta_{l,u}$ and $\alpha_{g,u}$ are provided along with other parameters in Appendix B. Accordingly, the principal stratum and outcomes ($Y_{ij}(0)$ or $Y_{ij}(1)$) of each subject can be generated from (4) and (5), respectively. The pairs of outcomes are further used to attain the approximated ATE under each setting via $\widetilde{\text{ATE}}_{1,0} = \sum_i \sum_j (Y_{ij}(1) - Y_{ij}(0))/N$.

In scenarios (v) and (vi), we follow the data structures of two applications (females' scores in math from education study and birth weight study) in Section 5, respectively. In particular, we consider the same numbers of covariates (2 and 4) and clusters (32 and 99), and use the parameter estimates from the applications to generate the data, where the n_i's are fixed at the same levels as the original data sets (see Figure 2.2). Also, note that in each replicate, covariates are generated from the distributions based on the original samples. For example, a categorical covariate is created from a multinomial distribution with event probabilities equal to the sample proportions.

2.5.2 Simulation Results

For each simulated data set, we set the hyper-parameters as $a_v = a_w = 5$, $b_v = b_w = 50$, and $a_\zeta = b_\zeta = a_\xi = b_\xi = a_\beta = b_\beta = a_\alpha = b_\alpha = a_\gamma = b_\gamma = a_0 = b_0 = 0.001$, and run a single MCMC chain

Figure 2.2: Histograms of sample size per cluster for the two data applications. For the education study, sample size per class is between 8 and 24, and for the birth weight study, clusters based on age and BMI have sample sizes between 1 and 32.



of 11,000 iterations with the first 1,000 as burn-ins. Note that the chosen values of a_v, a_w, b_v, and b_w reflect our prior belief that only few cluster effects are nonignorable. Based on 100 replications, we report the estimated ATE's in Figures 2.3 and 2.4, and summarize the root mean square biases (rMSB) in Table 2.2, where rMSB is defined as $\sqrt{\sum_{t=1}^{100}(\widehat{\text{ATE}}_{1,0,t} - \widetilde{\text{ATE}}_{1,0,t})^2/100}$, and $\widehat{\text{ATE}}_{1,0,t}$ and $\widetilde{\text{ATE}}_{1,0,t}$ are the estimated and approximated ATE of the t^{th} replicate, respectively. Besides the proposed method (which is labeled as "nmig"), there are three competing models considered in the simulations: (1) the naive approach (labeled as "naive") that omits all cluster effects; (2) the traditional random effect model (labeled as "random") that uses the normal priors for the cluster effects; and (3) the regularized random effect model (labeled as "lasso") that uses the Laplace priors for the cluster effects. Note that the regularized random effect model follows the conditional Laplace setting in [64] on each $\zeta_{i,l}$ and $\xi_{i,g}$. For more details may refer to Appendix C.

In Figures 2.3 and 2.4, boxplots of estimated ATE ($\widehat{\text{ATE}}_{1,0}$) values for all four methods are shown for the different scenarios. A few comments are warranted. First, the true ATE is analytically intractable due to the complicated data generating mechanisms. Therefore, the aforementioned $\widetilde{\text{ATE}}_{1,0,t}$ serves as the target value for comparison to $\widehat{\text{ATE}}_{1,0,t}$ in each replicate. Since these values

vary across replicates, we use the red solid and dashed lines to represent the average values and middle-90% intervals, of $\widetilde{\text{ATE}}_{1,0}$, respectively. Second, to match the real data analysis in Section 5, we also report the ATE estimates only based on certain strata in Figure 2.2. In those figures, the ATE calculation becomes

$$\widehat{\text{ATE}}_{1,0} = \frac{1}{N}\sum_i \sum_j \sum_{g=1,4} \frac{\hat{\gamma}_g \hat{p}_{ij}(g)}{\hat{p}_{ij}(1) + \hat{p}_{ij}(4)} \tag{2.8}$$

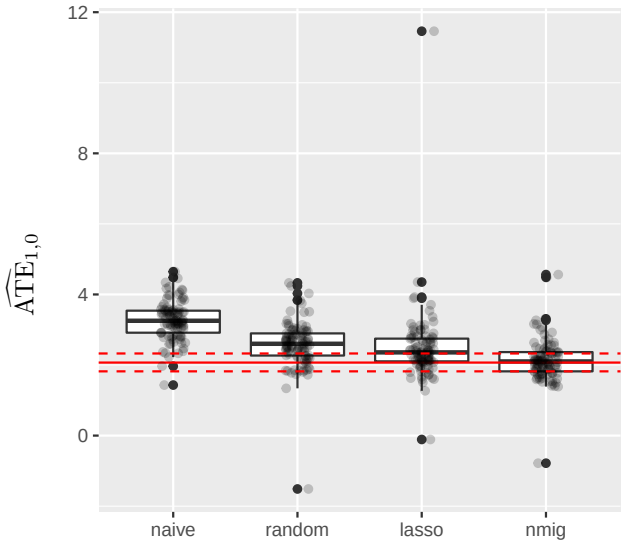
for (v), and $\widehat{\text{ATE}}_{1,0} = \hat{\gamma}_4$ for (vi). Third, the universally inferior performance of the naive approach implies the need to incorporate the cluster effects into the model. Even though in scenario (vi) the naive approach considers the two variables that form the clusters as covariates, the results still have the most variation in $\widehat{\text{ATE}}_{1,0}$, suggesting that linear additive terms in a traditional model is not appropriate in this study with clustering structure. Last, our proposed method outperforms the others both in the hypothetical and the real-data-simulated scenarios, in the sense that most of its interquartile ranges for estimated $\widehat{\text{ATE}}_{1,0}$ of nmig's are within the red dashed lines. This advantage becomes more pronounced when the number of clusters is large or sample size is small. This suggests better capability of the proposed method in addressing the extreme cases. The above findings are also echoed by Tables 2.2 and 2.3, where the proposed approach yields the smallest rMSB's in the six scenarios.

Table 2.2: Simulation results in terms of root mean squared bias (rMSB) for different scenarios based on 100 replications. The estimation methods are: naive method without clusters (denoted as "naive"), random effect for clusters (denoted as "random"), the lasso prior for clusters (denoted as "lasso") and the spike and slab prior for clusters (denoted as "nmig").

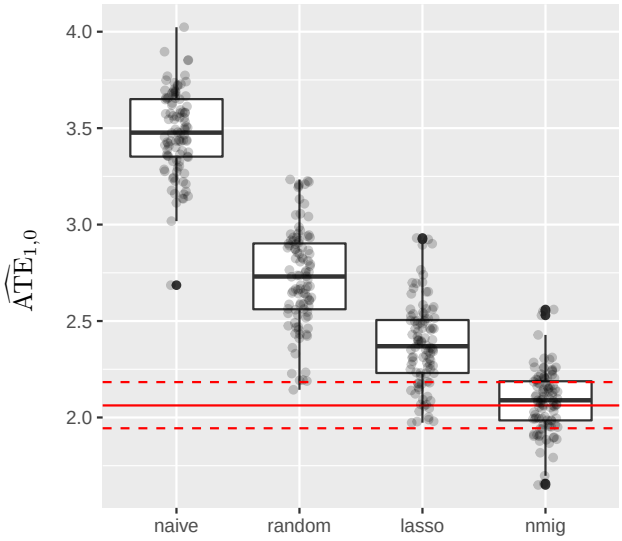
Scenario	(i)	(ii)	(iii)	(iv)
Average N	100	500	1,000	524
Average $\widetilde{\text{ATE}}_{1,0}$	2.070	2.062	2.103	2.083
SD of $\widetilde{\text{ATE}}_{1,0}$	0.154	0.078	0.055	0.074
	rMSB			
naive	1.318	1.433	1.458	1.475
random	0.843	0.708	0.751	0.095
lasso	1.128	0.374	0.383	0.089
nmig	0.603	0.146	0.117	0.095

Figure 2.3: Boxplots of simulation results for different scenarios based on 100 replications, where the x-axis denotes the different estimation methods: naive method without clusters (denoted as "naive"), random effect for clusters (denoted as "random"), the lasso prior for clusters (denoted as "lasso") and the spike and slab prior for clusters (denoted as "nmig"). The y-axis is $\widehat{\text{ATE}}_{1,0}$. The solid line is the average $\widehat{\text{ATE}}_{1,0}$, and the two dashed lines are 5^{th} and 95^{th} quantiles of $\widehat{\text{ATE}}_{1,0}$.

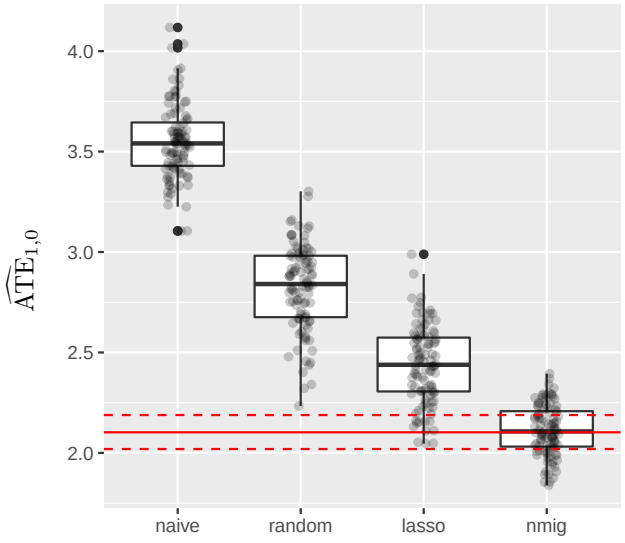
(i) $m = 50$, size=2



(ii) $m = 250$, size=2



(iii) $m = 500$, size=2



(iv) $m = 50$, size=6 to 15

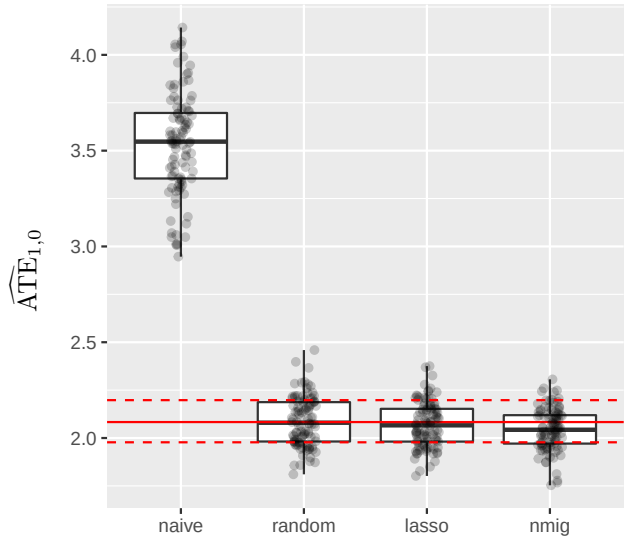
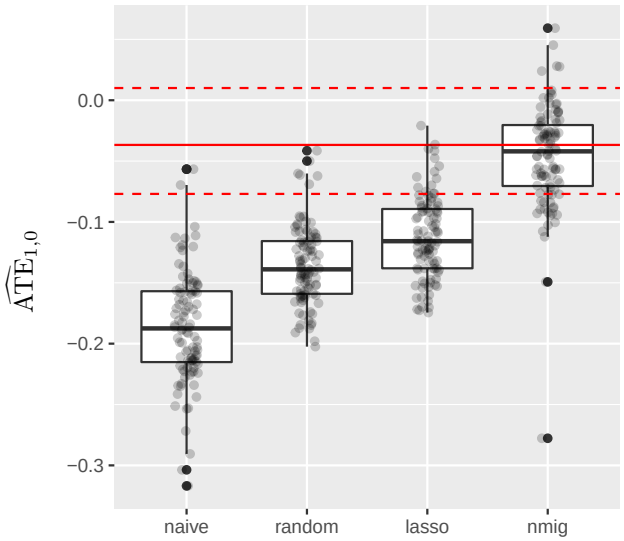
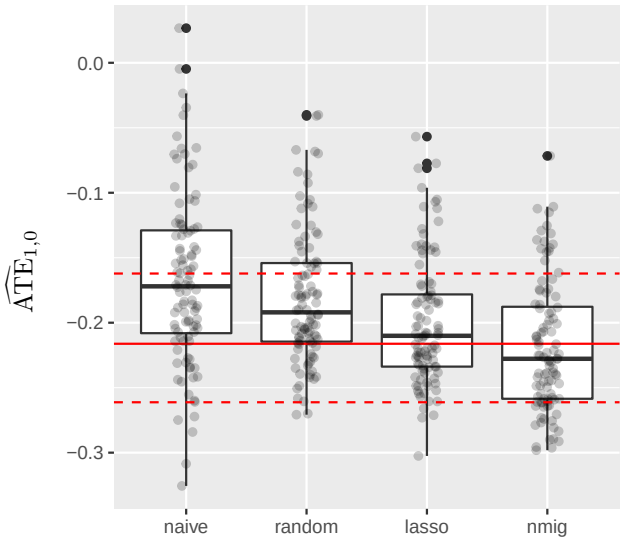


Figure 2.4: Boxplots of simulation results for the two scenario mimicking the two applications based on 100 replications, where the x-axis denotes the different estimation methods: naive method without clusters (denoted as "naive"), random effect for clusters (denoted as "random"), the lasso prior for clusters (denoted as "lasso") and the spike and slab prior for clusters (denoted as "nmig"). The y-axis is $\widehat{\text{ATE}}_{1,0}$. The solid line is the average $\widehat{\text{ATE}}_{1,0}$, two dashed lines are 5^{th} and 95^{th} quantiles of $\widehat{\widetilde{\text{ATE}}}_{1,0}$.

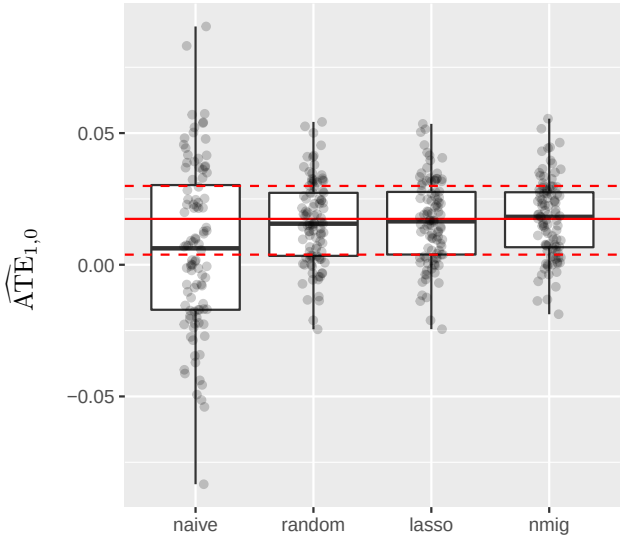
(v) $m = 32$, size=8 to 24, for all strata



(v) $m = 32$, size=8 to 24, for $g = 1$ & $g = 4$



(vi) $m = 99$, size=1 to 32, for $g = 1$ & $g = 4$



(vi) $m = 99$, size=1 to 32, for only $g = 4$

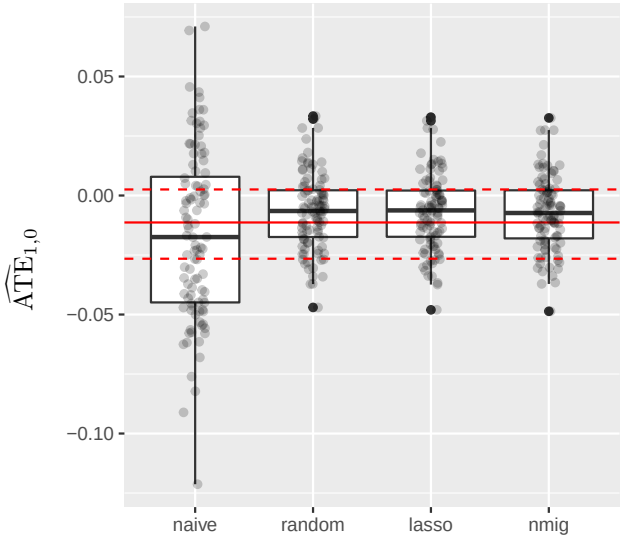


Table 2.3: Simulation results in terms of root of mean squared bias (rMSB) for the two scenarios mimicking the two applications based on 100 replications. The estimation methods are: naive method without clusters (denoted as "naive"), random effect for clusters (denoted as "random"), the lasso prior for clusters (denoted as "lasso") and the spike and slab prior for clusters (denoted as "nmig").

Scenario	(v)		(vi)	
Average N	515		1068	
	all strata	$g = 1$ & $g = 4$	$g = 1$ & $g = 4$	$g = 4$ only
Average $\widetilde{\text{ATE}}_{1,0}$	-0.037	-0.216	0.017	-0.011
SD of $\widetilde{\text{ATE}}_{1,0}$	0.026	0.036	0.009	0.009
	rMSB			
naive	0.155	0.080	0.033	0.036
random	0.104	0.058	0.013	0.014
lasso	0.083	0.046	0.013	0.014
nmig	0.045	0.044	0.012	0.013

where naive in scenario (vi) considers the two variables that form the clusters as covariates

2.6 Applications

The two data sets are used to illustrate the empirical advantages of the proposed method. One data set involves the educational research on the impacts of classroom types ($Z_{ij} = 1$ for single-gender and $Z_{ij} = 0$ for mixed-gender) on academic performance. The other data set is a birth weight study where pregnancies complicated by gestational diabetes mellitus (GDM) are also impacted by hypertension associated risk factors ($Z_{ij} = 1$ for the impacted group and $Z_{ij} = 0$ for the non-impacted group). In both applications, due to the natural existence of post-treatment variable and clusters, our proposed method could serve as a suitable tool to evaluate the causal effects. In particular, considering that resource allocation potentially affects student academic achievements [86], and that birth outcomes vary between natural birth and labor induction, we choose the class size (class size < 26 or ≥ 26) and induction of labor (yes or no) as the dichotomous post-treatment variables in the educational and birth weight data sets, respectively. Also, due to the difficulty in quantifying all the environmental factors (such as teacher's teaching ability) affecting student outcomes in each class, "class" defines clusters and is incorporated into the model in the educational data set to reflect the heterogeneity of classes under the same classroom type. In the birth weight data set, although both maternal age and pre-pregnancy BMI are measured, their effects could be more than just linear additive terms to the birth outcomes. To this end, we relax this linear assumption and consider the combined groups of integer age and pre-pregnancy BMI as the clusters, where following World Health Organization (WHO) guidelines, pre-pregnancy BMI is categorized into four

groups: underweight ($< 18.5kg/m^2$), normal ($18.5 - 24.9kg/m^2$), overweight ($25.0 - 29.9kg/m^2$), and obese ($\geq 30.0kg/m^2$) [61]. Other background information and settings are provided later in the discussion of each individual data set.

2.6.1 Educational Study

Research into the classroom setting that is most beneficial to learning is a primary focus of modern education research. The impact of gender class make-up is an important component of this research. For example, Elam [18] and Ferrara [19] suggested single-gender classroom could create an environment with less social stress and peer pressure. Parker and Rennie [65] also claimed that this type of classroom could avoid the sexual harassment and misbehavior which inhibit girls' learning greatly. On the other hand, Lee et al. [42] found that male students in the co-educational environment could have better academic performance than those in the single-gender environment while no differences were detected in female students. In this analysis, we attempt to utilize the data collected from two middle schools in the upstate of South Carolina in 2012 (from an NSF funded study by some of the coauthors) to provide some causal insights. These schools contain single-gender and mixed-gender classes (the treatment). The 2012 standardized test scores in log scale are the outcomes of interest. Ethnicity and 2011 standardized test scores as baseline covariates, the classroom size (post-treatment variable) and class (clusters) are taken into account in the data set.

The data set contains 46 classes, and is composed of 1,057 complete cases (515 female and 542 male students) out of 1,173, where the incomplete data is around 9.9%. To avoid the potential interaction effects between students' gender and treatment as discussed in the previous studies, we conduct separate analysis for females and males. Therefore, for females the treatments are female-only (treated) and females in mixed-gender (control) classes, and for males the treatments are male-only (treated) and males in mixed-gender (control) classes. We present the number of class in each combined group of Z_{ij} and $S_{ij}(z)$ in Table 2.4. Note the tendency to have a larger class size in the mixed-gender group. This provides motivation to consider the class size as the post-treatment variable and calculate the ATE via (2.8), where the strata $g = 1$ and 4 both hold $S_{ij}(0) = S_{ij}(1)$, thus avoiding the confounding effect of class size. Note that although n_i's in Fig 2 for education study are all smaller than 26 for the female student comparisons, the status of $S_{ij}(z)$ is determined by the original class size, which includes the students of the opposite gender or with incomplete

data, rather than the cluster size n_i.

Table 2.4: Classes categorization by sizes based on raw data set for three classroom types, each class is considered as a large class if the size is at least 26, otherwise as small class, where 26 is the average number of students in each class across all classes in original data set. There are 32 and 31 classes for female and male students performance comparisons respectively.

	$S_{ij}(z) = 0$ (class size < 26)	$S_{ij}(z) = 1$ (class size ≥ 26)
$Z_{ij} = 0$ (mixed-gender)	3	14
$Z_{ij} = 1$ (female-only)	11	4
$Z_{ij} = 1$ (male-only)	9	5

Adopting the hyper-parameter settings in Section 2.5.2, we generate an MCMC sample of 11,000 iterations with a 1,000 burn-ins, and report the posterior means and 95% highest posterior density (HPD) intervals of ATE in Table 2.5, where the 95% HPD intervals not containing zero are highlighted as the significant findings. The table contains the ATE's for the female treatment and the male treatment. It also shows the ATE's for different academic subjects ("ELASS" is the student score on the English Language Arts section of the exam, "MATH" is the score on the math section, "SCI" is the score on the science section, and "SOC" is the score on the Social Studies section). A few comments are warranted. First, the proposed approach yields the more stable posterior predictive distributions of ATE than "random" and "lasso". While the naive approach has the smallest standard deviations (SD) in "MATH", "SCI", and "SOC", it has extremely large SD's in "ELASS" suggesting the method fails in certain circumstances. This is in addition to the potential for the naive approach to be way off target as demonstrated in Section 4. Second, it seems that the single-gender class treatment is not beneficial to students' learning since there are six out of the eight posterior means have negative values. The two significant findings in female "MATH" and male "ELASS" suggest that the mixed-gender class may improve the female students' math scores and male students' language abilities, but more research is certainly needed.

2.6.2 Birth Weight Study

Numerous studies have shown the links between low birth weight and later-life health conditions such as obesity [84], type 2 diabetes [26, 47, 71], mental deficiencies [54], etc. Therefore, many studies have been devoted to identifying the factors causing this low birth weight. For example, Makgoba et al. [55] assessed the effect of GDM on birth weight, and for this particular group, i.e., children born from pregnancies complicated by GDM. Zhu et al. [101], Zhu et al. [102] and Wang et

Table 2.5: Summary results of the estimated ATE (treated - control) for educational performance with four different approaches to adjusting for post-treatment variables and clustering. Estimate (mean), standard deviation (SD) and 95% posterior credible intervals are based on 10,000 Monte Carlo samples after 1,000 burn-in periods. The analysis focuses on principal strata 1 and 4. Note that the outcome academic subject score are in log scale.

		F only vs F mixed				M only vs M mixed			
Subject	**Approach**	**mean**	**SD**	**95% HPD Int.**		**mean**	**SD**	**95% HPD Int.**	
				lower	**upper**			**lower**	**upper**
MATH	naive	-0.002	0.010	-0.022	0.017	0.028	0.014	0.002	0.055
	random	-0.113	0.136	-0.390	0.147	-0.089	0.420	-0.712	0.514
	lasso	0.020	0.103	-0.178	0.226	0.190	0.304	-0.486	0.730
	nmig	-0.218	0.050	-0.318	-0.123	0.048	0.063	-0.075	0.170
ELASS	naive	-7.797	3.832	-15.246	0.025	-1.724	6.667	-21.683	8.097
	random	0.104	0.139	-0.159	0.385	0.174	0.420	-0.369	1.359
	lasso	0.108	0.157	-0.187	0.399	0.090	0.121	-0.152	0.325
	nmig	-0.031	0.095	-0.239	0.139	-0.127	0.064	-0.258	-0.004
SCI	naive	0.018	0.019	-0.020	0.052	0.019	0.011	-0.000	0.042
	random	0.099	0.147	-0.204	0.375	-0.285	1.513	-5.195	1.264
	lasso	0.042	0.128	-0.214	0.303	0.006	0.122	-0.248	0.238
	nmig	-0.015	0.089	-0.195	0.155	0.068	0.067	-0.069	0.193
SOC	naive	0.016	0.017	-0.018	0.047	0.003	0.036	-0.066	0.068
	random	-0.142	0.121	-0.378	0.098	0.039	0.106	-0.168	0.239
	lasso	-0.058	0.111	-0.276	0.158	-0.005	0.087	-0.183	0.158
	nmig	-0.147	0.091	-0.329	0.028	-0.004	0.090	-0.181	0.172

al. [93] further investigated the impacts of maternal dietary behaviors on children's body mass index from birth through childhood. Additionally, Ananth et al. [4] found that hypertensive disorders in pregnancy could more likely trigger low birth weight, preterm births and small-for-gestational-age births. Motivated by these works, we attempt to evaluate the causal effect of hypertensive disorders on birth weight of a singleton pregnancy with the presence of gestational diabetes, where the birth weight in log scale is the outcome of interest. Gestational age, maternal education, ethnicity, and marital status are baseline covariates. Maternal exposure to any risk factors relevant to hypertension in terms of maternal smoking, hypertension, or eclampsia are recorded (as the treatment, specifically, $Z_{ij} = 1$ for exposed treatment). Induction of labor (post-treatment variable) and age and pre-pregnancy BMI groups (clusters) are in the data set.

The data set is collected from the National Vital Statistics System of the National Center for Health Statistics [20] and includes $N = 1{,}068$ women-offspring pairs with GDM during 2015 calendar year in US territories. Similar to the first application, the ATE calculation is only based on certain strata to yield a fair comparison. Specifically, under the same hyper-parameter settings as Section 2.5.2 and based on an MCMC sample of 11,000 iterations with a 1,000 burn-ins, we report the

results of stratum $g = 4$; i.e., the posterior summaries of γ_4, in Table 2.6. This is because there are various medical reasons (such as postterm pregnancy, growth problems of the baby, little amniotic fluid, and a uterine infection etc.) for inducing labor, and these reasons could dramatically affect the birth outcomes [88]. From Table 2.6, although all four methods results in insignificant findings, the naive approach (that treats age and BMI as covariates instead of clusters) yields a positive effect of the treatment on birth weight and the largest HPD interval. This suggests potential issues in using traditional modeling to adjust for the post-treatment and cluster level variables. On the other hand, consistent with the simulation scenarios (iv) and (vi), the similar results of random, lasso, and nmig methodologies are expected when most of clusters have medium sample sizes n_i's.

Table 2.6: Summary Results of the estimated ATE (treated - control) for birth weight with four different approaches to adjusting for post-treatment variables and clusters. Estimate (mean), standard deviation (SD) and 95% posterior credible intervals are based on 10,000 Monte Carlo samples after 1,000 burn-in periods. The analysis focuses on principal strata 4. Note the birth weight is in log-scale, and naive* approach includes age and BMI as covariates in the model.

Cluster	**Approaches**	**mean**	**SD**	**95% HPD Int.**	
				Lower	**Upper**
-	naive*	0.050	0.049	-0.043	0.149
Age & BMI	random	-0.013	0.017	-0.048	0.022
	lasso	-0.013	0.017	-0.046	0.023
	nmig	-0.011	0.017	-0.046	0.021

Chapter 3

Bayesian Estimation of Causal Effects using A Generalized Skewed Link Function for Observational Studies with Clustering and Unequal Sample Sizes

3.1 Abstract

Causal inference is one of the most significant and well researched topics in the analysis of observational studies. The propensity score is the most commonly used method to address the challenge of estimating causal effect of the treatment of interest on the outcome in the presence of unbalanced background covariates. Using the standard symmetric link function (logit or probit link) for the propensity score often results in poor estimates of the treatment effect when the number of

observations in treated group is much different than number of observations in control group (i.e. the sample sizes are unbalanced as well as the covariates). Furthermore, it is common in observational studies for the data to have a clustering structure. Random effects can be used to accommodate such structure, but inference for such a model is often difficult. In this work, a two-step Bayesian framework is proposed. The first step is to estimate the propensity score using a proposed generalized skewed link function associated with a Dirac delta function to adjust for the sample size imbalance. The second step is to estimate the propensity score as an additional latent variable to adjust for the covariate imbalance. We model the clustering structure as random effects with a spike and slab prior in a Bayesian hierarchical model. This allows many cluster effects to not be included in the final model. Through simulation data sets, and a data application about lipid profile outcomes, the proposed method is shown to perform better than commonly used approaches for determining the true underlying relationship between the binary treatment assignments and the outcome.

3.2 Introduction

Propensity score analysis [74] is one of the most commonly used technique to adjust for covariates imbalance between treated and control groups in observational studies. The standard link functions (like logit or probit links) are often used in modeling the relationship between covariates and treatment response. Both logit and probit links are symmetric links. They assume that the probability of a binary response approaches 0 or 1 at the same rate [13, 16]. Symmetric links are not always appropriate for modeling especially when the sample sizes are very different between the treatment and control groups (i.e., the number of observations are unbalanced between the treatment and control groups). This is often the case in public health science, social science and other fields. King and Zeng [39] showed that when dealing with rate event data (like infections by uncommon diseases or bankruptcy), modeling with the logit link has some limitations and could dramatically underestimate the probabilities of events. Chen et al. [13], Liu [52], Wang et al.[92], Jiang et al. [35], and Prasetyo et al. [69] have proposed a variety of generalized links/asymmetric links as alternatives to logit and probit links. These researchers discussed the importance of selecting the appropriate link function, and indicated that the the estimated probabilities can have significant bias, and the inference is suspect if a symmetric link function is used where an asymmetric link function is more appropriate.

One of the commonly used asymmetric links is the complementary loglog (Cloglog) link function. The Cloglog link has negative skewness, which means it approaches 0 fairly slowly but approaches 1 quite sharply and not vice versa [35]. It may generate biased inference when probabilities approach 0 and 1 symmetrically, or have positive skewness. Therefore, the Cloglog link is not a flexible choice for modeling a binary response. A more generalized and flexible link function allowing for varying amounts of skewness would be much better. Chen et al. [13] proposed skewed links using a latent variable approach but the approach did not give a principle for selection of the appropriate link function in different situations.

In addition to the unbalanced number of observations, the existence of clustering of observations also often occurs in observational studies. Estimating the treatment effect without considering these clusters could generate biased results. The variables that define the clusters (collectively denoted as cluster level variables) should be incorporated into the causal analysis to yield unbiased treatment effect estimates [5, 85, 90, 96]. Often only the cluster labels are available in the data set since the actual cluster level variables are typically not measured due to budget constraints or they are not part of the study objectives. Also, the true relationship between the cluster level variables and the outcome can be intractable. Therefore, the label is often used for appropriate adjustment in analysis. In the lipid profile study, we use "age groups" as the clusters, considering that the true relationship between age associated factors and the outcome (HDL cholesterol/triglyceride and/or LDL cholesterol) is very complex.

In this work, we propose a two-step Bayesian framework for estimating the treatment effect in observational data sets where both cluster level variables and unbalanced sample sizes between treatment and control groups are present. The first step is to use a new form of the generalized skewed link in the propensity score. The new link form is adopted from the skewed link in [13], where the parameter associated with the latent variable for skewness level is a Dirac delta function and a normal mixture structure for the error term. The advantage of the Dirac delta function is that it allows us to determine the appropriate level of skewness. The second step is to develop a regression model for the outcome that includes the treatment variable, the cluster effects and the propensity score as an additional latent variable to be estimated in a Bayesian framework.

This modeling and estimation approach differs from the most commonly used propensity score approaches to estimate the treatment effect [74]. In these approaches the propensity score is first estimated using the covariates (but not the outcome) and then matching, stratification, or

inverse weighting are used for treatment effect estimation. However, this new proposed approach offers some potential advantages over the common methods, and is also consistent with some previous research. While matching and stratification [58, 3] can be considered semi-parametric or non-parametric Bayesian approaches, the proposed modeling approach allows for fully Bayesian estimation. The inverse weighting could be influenced by extreme probabilities [10], especially in unbalanced cases, and the proposed modeling approach incorporates the skewed link to correct for this. In [23] and [100], it has been indicated that covariate adjustment in an outcome model provides lower bias in the context of rare binary outcomes and removes more imbalanced than stratification. Robins in [72], showed that when the treatment effect is homogeneous (in the sense that it is the same at all levels of propensity scores), a simple regression adjustment is robust against misspecification of the outcome model. The proposed modeling approach modeling approach allows the clusters to be addressed using a spike and slab prior. With such a sparsity inducing prior, it can be very beneficial when many of the cluster effects are not significant, resulting in a more parsimonious model. Another advantage is the proposed modeling approach leads to the development of a computationally efficient Markov Chain Monte Carlo (MCMC) sampling algorithm by making use of data augmentation [2] and the Pólya-Gamma technique [68] for logistic regression. Finally, the proposed modeling approach can be used to determine the true underlying relationship between the background covariates and binary response with least misspecification rate.

The rest of this chapter is organized as follows: Section 3.3 introduces the proposed generalized skewed link method along with the assumptions and prior settings. Section 3.4 provides Bayesian estimation and inference details with the data augmentation techniques and the posterior sampling algorithms. Section 3.5 compares the proposed generalized skewed link function and the logit and probit links using simulated datasets under various scenarios. Section 3.6 applies the proposed modeling approach to the lipid profile study from the 2013-2014 National Health and Nutrition Examination Survey (NHANES).

3.3 Methodology

3.3.1 Notation and Modeling

Let Z_{ij} denote the binary treatment assignment ($Z_{ij} = 1$ for treated and $Z_{ij} = 0$ for control) of the j^{th} subject in i^{th} cluster. $\boldsymbol{X}_{ij}$ denote the corresponding p-dimensional vector of pre-treatment

covariates and U_i denote the cluster level variable, where $i = 1, 2, \ldots m$, and $j = 1, 2, \ldots n_i$. The total sample size is $N = \sum_i^m n_i$. Also let $Y_{ij}(z)$ denote the continuous outcome given treatment assignment z. Under the counterfactual framework [31], $(Y_{ij}(0), Y_{ij}(1))$ represents a pair of potential primary outcomes, where at most one of the potential values could be observed. We focus on the population average treatment effect (ATE) of the primary outcome as the causal effect estimand,

$$\text{ATE}_{z,z'} = E[Y_{ij}(z) - Y_{ij}(z')], \quad \text{for } z \neq z' \tag{3.1}$$

Following the definition proposed by Rosenbaum and Rubin [74], let ps_{ij} denote a propensity score of the j^{th} subject in i^{th} cluster, which is the probability of the individual being assigned to the treated group, conditioning on the pre-treatment covariates $\boldsymbol{X}_{ij}$ and the cluster level variable U_i, notationally,

$$ps_{ij} = P(Z_{ij} = 1 \mid \boldsymbol{X}_{ij}, U_i) \tag{3.2}$$

The propensity score serves as a balancing score with the requirement of the stable unit treatment value assumption (SUTVA) (i.e., no interference among units under treated and control groups) and strong ignorability in the treatment assignment [74, 77, 78]. This can be expressed as

$$(Y_{ij}(0), Y_{ij}(1)) \perp Z_{ij} \mid \boldsymbol{X}_{ij}, U_i \Rightarrow (Y_{ij}(0), Y_{ij}(1)) \perp Z_{ij} \mid ps_{ij}$$

Therefore, the joint likelihood function of observed data is given by

$$\begin{aligned}
f(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{U}) &= \prod_{i,j} f(Y_{ij}(1), Y_{ij}(0), Z_{ij} \mid \boldsymbol{X}_{ij}, U_i) && \text{(by SUTVA)} \\
&= \prod_{i,j} f(Y_{ij}(1), Y_{ij}(0) \mid Z_{ij}, \boldsymbol{X}_{ij}, U_i) f(Z_{ij} \mid \boldsymbol{X}_{ij}, U_i) \\
&= \prod_{i,j} f(Y_{ij}(1), Y_{ij}(0) \mid Z_{ij}, ps_{ij}) ps_{ij}, && \text{(by strong ignorability)}
\end{aligned}$$

where $\boldsymbol{Y}(.) = (Y_{11}(.), Y_{12}(.), \ldots, Y_{mn_m}(.))^T$, $\boldsymbol{Z} = (Z_{11}, Z_{12}, \ldots, Z_{mn_m})^T$, $\boldsymbol{X} = (\boldsymbol{X}_{11}^T, \boldsymbol{X}_{12}^T, \ldots, \boldsymbol{X}_{mn_m}^T)^T$, $\boldsymbol{U} = (U_1, U_2, \ldots, U_m)^T$, the superscript T denotes the transposition of a vector. For our proposed two-step Bayesian propensity score approach, based on insights from [58], we model the treatment assignment and outcome models respectively. First, to further model the treatment assignment while considering the existence of a cluster effect $h(U_i)$, where $h(.)$ is an unknown function (due to

either unavailable U_i or complex relationship between U_i and treatment assignment), we quantify the cluster effect in terms of $\zeta_{i,1}$ into treatment assignment model. Specifically, we have

$$ps_{ij} = P(Z_{ij} = 1 \mid \boldsymbol{X}_{ij}, U_i) = f^*(\boldsymbol{X}_{ij}, \zeta_{i,1} \mid \boldsymbol{\beta}) \tag{3.3}$$

where f^* is some known analytic link function of covariates $\boldsymbol{X}_{ij}$ with corresponding parameters $\boldsymbol{\beta}$, and $\zeta_{i,1}$ for cluster effect. The standard choice of f^* is logit or probit link function expressed as

$$ps_{ij}^{logit} = \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1})}{1 + \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1})} \tag{3.4}$$

$$ps_{ij}^{probit} = \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1}) \tag{3.5}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. By using a latent variable approach for skewness embed into the standard link functions [2, 13], we propose a generalized skewed link function for f^* as follows. The binary treatment assignment Z_{ij} is given by

$$Z_{ij} = \begin{cases} 0 & \text{if} \quad w_{ij} < 0 \\ 1 & \text{if} \quad w_{ij} \geq 0 \end{cases}$$

where

$$w_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij} + \delta_2 e_{1,ij} + (1 - \delta_2)e_{2,ij}, \qquad t_{ij} \sim G \qquad e_{1,ij} \sim F_1, \qquad e_{2,ij} \sim F_2$$

t_{ij} , $e_{1,ij}$ and $e_{2,ij}$ are independent, G is the cdf of a known skewed distribution, and F_1 and F_2 are the cdf's of two known symmetric distributions. The parameter δ_1 is associated with the latent variable t_{ij} which controls the skewness, and the parameter δ_2 determines the choice of link function, that is, if $\delta_2 = 1$, the symmetric link with cdf F_1 is selected, otherwise $\delta_2 = 0$, the symmetric link with cdf F_2 is selected. In order to ensure the identifiability of model parameters with a finite third moment of the distribution of w_{ij}, throughout this article, we assume G is the cdf of the half-standard normal distribution with density $g(t) = \frac{2}{\sqrt{2\pi}}e^{-(t^2/2)}$, $t > 0$, which is tractable analytically and can be implemented easily from a computational perspective [13]. Therefore, we now propose

that the propensity score be defined as,

$$ps_{ij} = P(Z_{ij} = 1 \mid \boldsymbol{X}_{ij}, U_i) = \int_{-\infty}^{\infty} \Big(\delta_2 F_1(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) + (1-\delta_2) F_2(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})\Big)\, g(t_{ij}) dt_{ij} \tag{3.6}$$

For simplicity and ease of implementation, we use the two most links $F_1 = \exp(s)/(1+\exp(s))$ and $F_2 = \Phi(s)$ to derive the proposed generalized skewed logit and generalized skewed probit models, that is,

$$ps_{ij}^{skewl} = \int_{-\infty}^{\infty} \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} g(t_{ij}) dt_{ij} \tag{3.7}$$

$$ps_{ij}^{skewp} = \int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) g(t_{ij}) dt_{ij} \tag{3.8}$$

Then the complete likelihood for treatment assignment is given by,

$$\begin{aligned} L(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{t} \mid \boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1, \delta_2) = \prod_{i,j} &\Big[\delta_2 F_1(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) + (1-\delta_2) F_2(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})\Big]^{Z_{ij}} \\ &\times \Big[1 - \Big(\delta_2 F_1(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) + (1-\delta_2) F_2(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})\Big)\Big]^{1-Z_{ij}} g(t_{ij}) \end{aligned} \tag{3.9}$$

where $\boldsymbol{t} = (t_{11}, t_{12}, \ldots, t_{mn_m})'$, $\boldsymbol{\zeta}_1 = (\zeta_{1,1}, \zeta_{2,1}, \ldots, \zeta_{m,1})'$. Then the joint posterior distribution of $(\boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1, \delta_2)$ is given by

$$\begin{aligned} \pi(\boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1, \delta_2 \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{t}) \propto \pi(\boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1, \delta_2) \prod_{i,j} &\Big[\delta_2 F_1(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) + (1-\delta_2) F_2(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})\Big]^{Z_{ij}} \\ &\times \Big[1 - \Big(\delta_2 F_1(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) + (1-\delta_2) F_2(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})\Big)\Big]^{1-Z_{ij}} g(t_{ij}) \end{aligned} \tag{3.10}$$

Similarly, we define $\zeta_{i,2}$ to quantify the cluster effect $q(U_i)$ in the outcome model, and propose the following regression model of the outcome with a propensity score adjustment [62],

$$Y_{ij}(z) \mid ps_{ij}, U_i \sim N(\alpha z + \gamma ps_{ij} + \zeta_{i,2}, \sigma_y^2) \tag{3.11}$$

where α and γ are the the regression coefficients, and σ_y^2 is the outcome variance. Then we can estimate (3.1) by

$$\widehat{\text{ATE}}_{z,z'} = \hat{\alpha} \tag{3.12}$$

where $\hat{\alpha}$ is the posterior means of α in outcome analysis. under the above model setting with complete parameters including $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, $\boldsymbol{\zeta}_1 = (\zeta_{1,1}, \zeta_{2,1}, \ldots, \zeta_{m,1})$, δ_1, δ_2, $\boldsymbol{t}$, α, γ, $\boldsymbol{\zeta}_2 = (\zeta_{1,2}, \zeta_{2,2}, \ldots, \zeta_{m,2})$, σ_y^2.

3.3.2 Prior Setting

The proposed link function for propensity score can be defined in terms of parameters for skewness (measured by δ_1) and for the standard symmetric link function (measured by δ_2). These two parameters define four link functions: 1) symmetric logit link function ($\delta_1 = 0$ and $\delta_2 = 1$), 2) symmetric probit link function ($\delta_1 = 0$ and $\delta_2 = 0$), 3) skewed logit link function (δ_1 is nonzero and $\delta_2 = 1$) and 4) skewed probit link function (δ_1 is nonzero and $\delta_2 = 0$). Therefore, Dirac delta function [56, 91] is adopted for δ_1 as follows

$$\begin{aligned} &\delta_1 \sim (1-\eta)\Delta_0(\delta_1) + \eta N(0, \sigma_{\delta_1}^2), \\ &\eta \mid p^\eta \sim \text{Ber}(p^\eta), \\ &p^\eta \sim \text{Beta}(a, b), \end{aligned} \tag{3.13}$$

where $\sigma_{\delta_1}^2$ is pre-specified hyperparameter, Δ_0 denote a Dirac delta function which can yield values zero and nonzero for binary selection in proposed model, η is an indicator variable with the value 1 considering skewness variable in the proposed link function, while value 0 ignoring skewness effect. p^η is the prior inclusion probability. The update of δ_1 and δ_2 will be discussed in the next section.

Normally, $\zeta_{i,1}$ in (3.6), and $\zeta_{i,2}$ in (3.11) are treated as random effects (equivalent to the Gaussian priors in the Bayesian framework) in the standard approach to handle clusters. However, as the number of clusters (m) gets larger, the model complexity increases monotonously. In addition, motivated by the prior belief that majority of cluster effects are ignorable with few significant effects, we adopt a sparsity-inducing prior to efficiently select significant cluster effects and reach model parsimony. Specifically, instead of the Gaussian prior, we consider spike and slab priors [59] for $\zeta_{i,1}$

and $\zeta_{i,2}$ respectively as follows

$$
\begin{aligned}
&\zeta_{i,1} \mid v_i \sim (1-v_i)N(0, r\sigma_1^2) + v_i N(0, \sigma_1^2),\\
&v_i \mid p_{v_i} \sim \text{Ber}(p_{v_i}),\\
&p_{v_i} \sim \text{Beta}(a_v, b_v),\\
&\sigma_1^2 \sim \text{invGamma}(a_1, b_1),\\
&\zeta_{i,2} \mid v_i^y \sim (1-v_i^y)N(0, r\sigma_2^2) + v_i^y N(0, \sigma_2^2),\\
&v_i^y \mid p_{v_i^y} \sim \text{Ber}(p_{v_i^y}),\\
&p_{v_i^y} \sim \text{Beta}(a_{v^y}, b_{v^y}),\\
&\sigma_2^2 \sim \text{invGamma}(a_2, b_2)
\end{aligned}
\tag{3.14}
$$

where v_i (or v_i^y) is an indicator variable with the value 1 denoting a significant cluster effect from the slab distribution $N(0, \sigma_1^2)$ (or $\zeta_{i,2} \sim N(0, \sigma_2^2)$ when $v_i^y = 1$) while 0 stands for an ignorable cluster effect from the spike component $N(0, r\sigma_1^2)$ (or $\zeta_{i,2} \sim N(0, r\sigma_2^2)$ when $v_i^y = 0$). p_{v_i} and $p_{v_i^y}$ are the prior inclusion probabilities, $a_v, b_v, a_1, b_1, a_{v^y}, b_{v^y}, a_2$, and b_2 are pre-specified hyperparameters, and r is a pre-specified small value. Here, we set r as 0.00025 for both simulation studies and data application following the suggestion by Scheipl [81]. we select the normal mixture inverse-gamma (NMIG) prior [34] in (3.14) under consideration of no common intercept term in the proposed model. This is because unlike the stochastic search variable selection (SSVS) proposed by George and McCulloch [25], NMIG does not require specifying σ_1^2 and σ_2^2 based on prior knowledge. Also, it can better accommodate these cluster-and-intercept mixed terms in a relatively smoothing setting comparing with a Dirac delta function [56, 91] that yields intercepts with value zero.

It also worth noting that the spike and slab prior can efficiently explore the entire cluster effect sample space, and select those nonignorable cluster effects simultaneously in the estimation procedure. The ATE calculation is then based on the selected MCMC sample to evaluate. Therefore, we adopt the two-step procedure to obtain the ATE estimate: first, determine the most appropriate link function among the four possibilities defined previously for propensity score estimation. In the MCMC chain, the posterior samplers associated with the most often occurred link function are used to generate propensity score estimation in (3.6) and second, further generate MCMC samples for the outcome model. With the selected link function, the same sampling scheme is used to generate

another MCMC sample for 3.12.

For the remaining parameters in the two-step proposed methodology, consider the following conjugate priors

$$\begin{aligned}
&\boldsymbol{\beta} \mid \sigma_\beta^2 \sim MVN(\mathbf{0}, \sigma_\beta^2 I_p),\\
&\sigma_\beta^2 \sim \text{invGamma}(a_\beta, b_\beta),\\
&t_{ij} \sim \text{half-Normal}(0, 1),\\
&\alpha \mid \sigma_\alpha^2 \sim N(0, \sigma_\alpha^2),\\
&\sigma_\alpha^2 \sim \text{invGamma}(a_\alpha, b_\alpha),\\
&\gamma \sim N(0, \sigma_\gamma^2),\\
&\sigma_\gamma^2 \sim \text{InvGamma}(a_\gamma, b_\gamma),\\
&\sigma_y^2 \sim \text{invGamma}(a_y, b_y)
\end{aligned}$$

where I_p is a $p \times p$ identity matrix, and $a_\beta, b_\beta, a_\alpha, b_\alpha, a_y, b_y, a_\gamma, b_\gamma$ are pre-specified hyperparameters.

3.4 Data Augmentation and Bayesian Inference

At the beginning of each iteration, there are four possible cases as discussed in previous section considering with the combination of δ_1 and δ_2 values. For simplicity, recall that $F_1 = \exp(s)/(1 + \exp(s))$ and $F_2 = \Phi(s)$ are the logit and probit link respectively. If δ_1 is 0, then (3.6) reduces to the standard symmetric logit/probit link. If δ_1 is nonzero, we consider the skewed logit/probit link discussed by Chen et.al [13]. Let "| ." denote "given all the other parameters and observed data". Specifically, to update δ_2 in each iteration, we draw from a Bernoulli distribution where the probability is calculated as, if $\delta_1 = 0$ at k^{th} MCMC iteration,

$$P(\delta_2^{(k)} = 1 \mid .) = \frac{A}{A + B} = \frac{1}{1 + \exp(\log(B) - \log(A))}$$

where

$$A = \prod_{i,j} \left[F_1 \left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}^{(k)} + \zeta_{i,1}^{(k)} \right) \right]^{Z_{ij}} \left[1 - F_1 \left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}^{(k)} + \zeta_{i,1}^{(k)} \right) \right]^{1 - Z_{ij}}$$

$$B = \prod_{i,j} \left[F_2(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}^{(k)} + \zeta_{i,1}^{(k)}) \right]^{Z_{ij}} \left[1 - F_2 \left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}^{(k)} + \zeta_{i,1}^{(k)} \right) \right]^{1-Z_{ij}}$$

where $\boldsymbol{\beta}^{(k)}$, $\zeta_{i,1}^{(k)}$ are the corresponding k^{th} MCMC samplers. Otherwise, if δ_1 is nonzero, then, we can sample from a Bernoulli distribution with probability

$$P(\delta_2^{(k)} = 1 \mid .) = \frac{A}{A+B} = \frac{1}{1 + \exp\left(\log(B) - \log(A)\right)}$$

where

$$A = \prod_{i,j} (ps_{ij}^{skewl})^{Z_{ij}} (1 - (ps_{ij}^{skewl}))^{1-Z_{ij}}$$

$$B = \prod_{i,j} (ps_{ij}^{skewp})^{Z_{ij}} (1 - (ps_{ij}^{skewp}))^{1-Z_{ij}}$$

After determining the link function, the full conditionals of parameters related to (3.6) can be sampled from the conjugate forms under our prior settings. To obtain the MCMC sample of other parameters associated with (3.6), a series of data augmentations is performed to facilitate posterior computation. First, one of the data augmentations is introducing the Pólya-Gamma variables for Bayesian inference of logistic regression proposed by Polson et. al [68]. This technique avoids the traditional heavy computation of Metropolis-Hasting for logistic regression in Bayesian inference. The main result is representing the binomial likelihood of the log-odds as mixtures of Gaussians with respect to a Pólya-Gamma distribution as follows

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{k\psi} \int_0^{\infty} e^{-w\psi^2/2} p(w) dw, \tag{3.15}$$

where $k = a - b/2$ and $w \sim \text{PG}(b, 0)$, i.e. w is generated from the Pólya-Gamma distribution. When ψ is a linear function of factors, saying $\psi = \boldsymbol{X}^T \boldsymbol{\beta}$, then the integrand of the distribution the Gaussian kernel in $\boldsymbol{\beta}$. Polson et. al [68] also showed that the conditional distribution of $w \mid \psi$ also follows a Pólya-Gamma distribution, which suggests a strategy based on Gibbs sampling for posterior inference of $\boldsymbol{\beta}$. Specifically, in this study, the skewed logit link function considering cluster effects is defined in equation (3.7). To run the Gibbs sampler, we need to sample from the following conditional distribution of $\boldsymbol{\beta}$

$$\pi(\boldsymbol{\beta} \mid .) \propto \pi(\boldsymbol{\beta}) \prod_{i,j} \left[\frac{\exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1 + \exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{Z_{ij}} \left[1 - \frac{\exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1 + \exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{1-Z_{ij}} g(t_{ij})$$

$$
\begin{aligned}
&\propto \pi(\boldsymbol{\beta}) \prod_{i,j} \left[\frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{Z_{ij}} \left[1 - \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{1-Z_{ij}} \\
&\propto \pi(\boldsymbol{\beta}) \prod_{i,j} \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})^{Z_{ij}}}{\left(1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})\right)^1} \\
&\propto \pi(\boldsymbol{\beta}) \exp\left(-1/2(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})^T \boldsymbol{\Omega} (\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})\right)
\end{aligned}
$$

the last line is rewritten by introducing the Pólya-Gamma variables, and $k_{ij} = Z_{ij} - 1/2$, w_{ij} is generated from a Pólya-Gamma distribution, that is, $w_{ij} \mid . \sim \text{PG}(1, \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})$, and $L_{ij} = k_{ij}/w_{ij}$. Reorganizing and collecting the terms with $\boldsymbol{\beta}$, we then have,

$$
\begin{aligned}
&\boldsymbol{\beta} \mid . \sim N(\mu^*, \Sigma^*) \\
&\Sigma^* = \left(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X} + 1/\sigma_\beta^2 \mathbf{I}_p\right)^{-1} \\
&\mu^* = \Sigma^* \left(\boldsymbol{X}^T\boldsymbol{\Omega}(L - C\zeta_1 - \delta_1\boldsymbol{t})\right)
\end{aligned}
$$

where $C = (C_{11}, \ldots, C_{1n_1}, \ldots, C_{m1}, \ldots, C_{mn_m})^T$ is a $N \times m$ matrix indicating which cluster each subject belongs to, $w = \{w_{ij}\}_{i,j}$., $\boldsymbol{\Omega}$ is a diagonal matrix with elements w_{ij}, $k = \{k_{ij}\}_{i,j}$, and $L = \{L_{ij}\}_{i,j}$. Correspondingly, the posterior distribution of δ_1 can be derived when $\eta = 1$ as follows, while the posterior sample $\delta_1 = 0$ when $\eta = 0$.

$$
\begin{aligned}
\pi(\delta_1 \mid .) &\propto \pi(\delta_1) \prod_{i,j} \left[\frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{Z_{ij}} \left[1 - \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{1-Z_{ij}} g(t_{ij}) \\
&\propto \pi(\delta_1) \prod_{i,j} \left[\frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{Z_{ij}} \left[1 - \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})}{1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})} \right]^{1-Z_{ij}} \\
&\propto \pi(\delta_1) \prod_{i,j} \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})^{Z_{ij}}}{\left(1+\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})\right)^1} \\
&\propto \pi(\delta_1) \exp\left(-1/2(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})^T \boldsymbol{\Omega} (\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})\right)
\end{aligned}
$$

the last line is rewritten by introducing the Pólya-Gamma variables, where

$$
\begin{aligned}
&\omega_{ij} \mid . \sim \text{PG}(1, \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) \\
&k_{ij} = Z_{ij} - 1/2
\end{aligned}
$$

$$L_{ij} = k_{ij}/w_{ij}, \quad \mathbf{L} = \{L_{ij}\}$$

$$\boldsymbol{\Omega} = \text{diag}(\omega_{ij})$$

Therefore, reorganizing and collecting the terms with δ_1, we have the posterior sample of δ_1 from

$$\delta_1 \mid . \sim N\left(\frac{\boldsymbol{t}^T\boldsymbol{\Omega}(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1)}{\boldsymbol{t}^T\boldsymbol{\Omega}\boldsymbol{t} + 1/\sigma^2_{\delta_1}}, \frac{1}{\boldsymbol{t}^T\boldsymbol{\Omega}\boldsymbol{t} + 1/\sigma^2_{\delta_1}}\right)$$

The other data augmentation technique is implemented on probit models as described in Albert and Chib [2]. The main idea is to rewrite the probit likelihood as a Gaussian density based on the latent variable. The latent variables are generated for the desired parameters and then updated with posterior inference based on Gibbs sampling via explicit posterior densities. These strategies circumvent the heavy computation by using Metropolis-Hasting in a more efficient way. We provide the detailed algorithms for posterior computation for the remaining parameters associated with the proposed generalized skewed logit (and probit) link function in Appendix.

3.5 Simulations

3.5.1 Simulation Settings

To evaluate the performance of the proposed method, we perform a simulation study involving four factors that could impact propensity score analysis: 1) true model of the treatment assignment process (three levels), 2) number of clusters (m) (two levels), 3) cluster size (n_i) (two levels) and 4) percentage of treated observations (three levels); which results in the 36 scenarios displayed in Table 3.1. For each simulation scenario, 100 replications are generated. we include two pre-treatment covariates $\boldsymbol{X}_{ij} = (X_{1,ij}, X_{2,ij})$ that represent the two types of covariates which often occur in practice. One continuous covariate is generated from a normal distribution, the other categorical covariate is created from a multinomial distribution with three equally likely possible values. The specific generation distributions are summarized in Table 3.2 for the different models of generating the treatment assignments and the percentage of treated observations. Let the treatment assignment Z_{ij} follow a Bernoulli distribution with probability of success depending on the three link functions. More specifically, for Cloglog link $1 - \exp(-\exp(.))$, logit link $1/(1 + \exp(-.))$ and probit link $\Phi(.)$, with the linear components of these models as

Table 3.1: Simulation factorial experiment design based on four factors: 1) true model of treatment assignment process, 2) number of clusters (m), 3) cluster size (n_i) and 4) percentage of treated observations.

True model	Number of clusters (m)	Each cluster size (n_i)	Percentage of treated observations
Cloglog / logit / probit	50	Small	50%
	50	Small	75%
	50	Small	90%
	50	Large	50%
	50	Large	75%
	50	Large	90%
	100	Small	50%
	100	Small	75%
	100	Small	90%
	100	Large	50%
	100	Large	75%
	100	Large	90%

Table 3.2: Simulation setting: generation distribution for covariates $\boldsymbol{X}_{ij} = (X_{1,ij}, X_{2,ij})$ under different scenarios of true model of generating treatment assignment and percentage of treated observations.

true model	Percentage of treated observations	$X_{1,ij}$	$X_{2,ij}$
Cloglog	50%	$N(-1.65, 1)$	uniform$(-0.67, 0, 0.67)$
	75%	$N(0.55, \sqrt{3})$	uniform$(-1.7, 0, 0.5)$
	90%	$N(2, \sqrt{3})$	uniform$(-1.5, 0, 1)$
logit	50%	$N(-1.2, 1)$	uniform$(-0.5, 0, 1)$
	75%	$N(0.5, 1)$	uniform$(-0.5, 0, 1)$
	90%	$N(3, 1.5)$	uniform$(-2, 0, 1)$
probit	50%	$N(-1.1, 1)$	uniform$(-1, 0, 1)$
	75%	$N(0.4, 1)$	uniform$(-1, 0, 1)$
	90%	$N(2.4, \sqrt{3})$	uniform$(-1.5, 0, 1)$

$\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + h(U_i) = X_{1,ij}\beta_1 + X_{2,ij}\beta_2 + \beta^U U_i$, where assume $h(U_i) = \beta^U U_i$. The cluster level variable U_i follows a normal distribution where the first 60% of U_i from $N(0, 0.1^2)$, and the remaining from $N(3, 0.1^2)$. Accordingly, the two potential outcome values for each individual are generated from $Y_{ij}(z) = \boldsymbol{X}_{ij}^T\boldsymbol{\omega} + \gamma z + q(U_i) + N(0, \sigma_y^2) = X_{1,ij}\omega_1 + X_{2,ij}\omega_2 + \omega^U U_i + \gamma z + N(0, \sigma_y^2)$, for $z = 0, 1$, where we assume $q(U_i) = \omega^U U_i$. The values of β_1, β_2, β^U, ω_1, ω_2, ω^U and γ used in the simulation are provided in Table 3.3. The $Y_{ij}(0)$ and $Y_{ij}(1)$ are used to obtain the approximated ATE under each scenario by $\widetilde{\text{ATE}}_{1,0} = \sum_i \sum_j (Y_{ij}(1) - Y_{ij}(0))/N$.

Table 3.3: True parameters

parameter	value
β_1	1
β_2	1
β^U	1
ω_1	1
ω_2	1
ω^U	1
γ	2

3.5.2 Simulation Results

We run a single MCMC chain of 11,000 iterations with the first 1,000 as burn-ins, and set the hyper-parameters discussed in the previous section as $\sigma_{\delta_1}^2 = 100$, $a_v = a_{v^y} = 5$, $b_v = b_{v^y} = 50$, and $a_1 = b_1 = a_2 = b_2 = a_\beta = b_\beta = a_\alpha = b_\alpha = a_\gamma = b_\gamma = a_y = b_y = 0.001$ for each simulated data set. Based on 100 replications, the estimated ATE's ($\widehat{\text{ATE}}_{1,0}$ in 3.12) are reported in the following tables and are summarized in terms of the mean squared error (MSE) under different scenarios, defined as $\text{MSE} = \sum_{r=1}^{100} (\widehat{\text{ATE}}_{1,0,r} - \widetilde{\text{ATE}}_{1,0,r})^2/100$, where $\widehat{\text{ATE}}_{1,0,r}$ and $\widetilde{\text{ATE}}_{1,0,r}$ are the estimated and approximated ATE of the r^{th} replicate, respectively. Note that the $\widetilde{\text{ATE}}_{1,0,r}$ is considered as the target value to compare in each replicate, since the true ATE is not tractable based on sample data. For reporting results of the different analysis approaches, we denote "re" as using random effects with a normal prior and "nmig" as using the spike and slab prior. The proposed generalized skewed link with the nmig method is labeled as "gen skew nmig" with three specific models considered in the simulations: (1) the generalized skewed link with the random effect model (labeled as "gen skew re") that uses the normal priors for the cluster effects; (2) the standard symmetric link (logit or probit) with nmig model (labeled as "logit (or probit) nmig") that uses the nmig priors for the cluster effects; and (3) the standard symmetric link (logit or probit) with random effect model (labeled as

"logit (or probit) re") that uses the normal priors for the cluster effects.

Table 3.4: Simulation results under true Cloglog model of different scenarios: (a) $m = 50$, n_i uniformly ranges from 6 to 15; (b) $m = 50$, n_i uniformly ranges from 12 to 30; (c) $m = 100$, n_i uniformly ranges from 6 to 15; (d) $m = 100$, n_i uniformly ranges from 12 to 30 and three levels of percentage of treated observations in terms of MSE, including mean and standard deviation of $\widetilde{\text{ATE}}_{1,0}$. The proposed generalized skewed link with nmig method is labeled as "gen skew nmig" with three comparing models: (1) the generalized skewed link with random effect model (labeled as "gen skew re") that uses the normal priors for the cluster effects; (2) the standard logit link with nmig model (labeled as "logit nmig") that uses the nmig priors for the cluster effects; and (3) the standard logit link with random effect model (labeled as "logit re") that uses the normal priors for the cluster effects.

(a) $m = 50$, n_i uniformly ranges from 6 to 15									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.06)	-	-	2.00 (0.06)	-	-	1.99(0.06)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	2.01 (0.16)	0.022	0.148	1.96 (0.20)	0.036	0.189	1.84 (0.30)	0.108	0.329
gen skew re	1.92 (0.16)	0.026	0.162	1.73 (0.19)	0.105	0.324	1.98 (0.29)	0.083	0.288
logit nmig mcmc2	2.10 (0.17)	0.039	0.196	1.93 (0.19)	0.037	0.192	1.80 (0.29)	0.115	0.339
logit re	1.98 (0.16)	0.023	0.150	1.74 (0.20)	0.106	0.326	1.50 (0.28)	0.321	0.567
(b) $m = 50$, n_i uniformly ranges from 12 to 30									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	1.99 (0.05)	-	-	2.00 (0.05)	-	-	2.00 (0.05)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	1.98 (0.10)	0.008	0.089	2.00 (0.13)	0.014	0.118	2.09 (0.21)	0.052	0.229
gen skew re	1.94 (0.10)	0.011	0.105	1.81 (0.13)	0.050	0.224	1.83 (0.23)	0.079	0.281
logit nmig mcmc2	2.00 (0.10)	0.008	0.090	1.95 (0.14)	0.018	0.132	1.89 (0.21)	0.054	0.231
logit re	1.96 (0.10)	0.009	0.096	1.87 (0.13)	0.032	0.179	1.82 (0.21)	0.076	0.276
(c) $m = 100$, n_i uniformly ranges from 6 to 15									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.04)	-	-	2.00 (0.05)	-	-	1.99 (0.04)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	1.99 (0.10)	0.007	0.082	2.03 (0.16)	0.023	0.153	1.91 (0.23)	0.059	0.243
gen skew re	1.92 (0.10)	0.013	0.113	1.72 (0.14)	0.097	0.311	1.95 (0.24)	0.056	0.237
logit nmig mcmc2	2.12 (0.14)	0.031	0.175	1.96 (0.15)	0.022	0.148	1.87 (0.25)	0.075	0.274
logit re	1.98 (0.10)	0.007	0.086	1.72 (0.15)	0.098	0.313	1.45 (0.22)	0.338	0.581
(d) $m = 100$, n_i uniformly ranges from 12 to 30									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.03)	-	-	2.00 (0.03)	-	-	2.00 (0.03)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	2.00 (0.07)	0.004	0.067	2.03 (0.08)	0.007	0.088	2.13 (0.11)	0.027	0.165
gen skew re	1.96 (0.07)	0.006	0.079	1.82 (0.09)	0.040	0.199	1.74 (0.13)	0.082	0.287
logit nmig mcmc2	2.03 (0.07)	0.005	0.072	1.98 (0.08)	0.007	0.085	1.91 (0.12)	0.021	0.142
logit re	1.99 (0.07)	0.005	0.068	1.89 (0.09)	0.020	0.140	1.79 (0.12)	0.057	0.240

In Table 3.4, 3.5 and 3.6 summarize the distribution of $\widehat{\text{ATE}}_{1,0}$ with associated MSE under three true models of generating the treatment assignment (that is, Cloglog, logit and probit). Within each true model setting, there are four sub-scenarios (a) $m = 50$, n_i uniformly ranges from 6 to 15;

Table 3.5: Simulation results under true logit model of different scenarios: (a) $m = 50$, n_i uniformly ranges from 6 to 15; (b) $m = 50$, n_i uniformly ranges from 12 to 30; (c) $m = 100$, n_i uniformly ranges from 6 to 15; (d) $m = 100$, n_i uniformly ranges from 12 to 30 and three levels of percentage of treated units in terms of MSE, including mean and standard deviation of $\widetilde{\text{ATE}}_{1,0}$. The proposed generalized skewed link with nmig method is labeled as "gen skew nmig" with three comparing models: (1) the generalized skewed link with random effect model (labeled as "gen skew re") that uses the normal priors for the cluster effects; (2) the standard logit link with nmig model (labeled as "logit nmig") that uses the nmig priors for the cluster effects; and (3) the standard logit link with random effect model (labeled as "logit re") that uses the normal priors for the cluster effects.

(a) $m = 50$, n_i uniformly ranges from 6 to 15									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.01 (0.06)	-	-	2.01 (0.06)	-	-	2.01 (0.06)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	2.00 (0.12)	0.012	0.108	2.07 (0.13)	0.017	0.129	2.10 (0.23)	0.059	0.244
gen skew re	1.95 (0.12)	0.016	0.125	1.99 (0.13)	0.013	0.113	2.01 (0.29)	0.078	0.279
logit nmig mcmc2	2.01 (0.12)	0.011	0.107	2.05 (0.13)	0.015	0.122	2.01 (0.25)	0.059	0.243
logit re	1.98 (0.12)	0.012	0.112	1.91 (0.12)	0.020	0.140	1.77 (0.23)	0.104	0.322
(b) $m = 50$, n_i uniformly ranges from 12 to 30									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.05)	-	-	2.00 (0.05)	-	-	2.00 (0.05)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	2.00 (0.08)	0.006	0.074	2.00 (0.09)	0.006	0.080	2.04 (0.17)	0.027	0.165
gen skew re	1.97 (0.08)	0.006	0.078	1.95 (0.09)	0.009	0.094	1.81 (0.17)	0.064	0.253
logit nmig mcmc2	2.00 (0.08)	0.005	0.074	2.00 (0.09)	0.006	0.079	2.02 (0.17)	0.028	0.166
logit re	1.99 (0.08)	0.006	0.075	1.94 (0.09)	0.010	0.101	1.79 (0.17)	0.069	0.263
(c) $m = 100$, n_i uniformly ranges from 6 to 15									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.05)	-	-	2.00 (0.05)	-	-	2.00 (0.05)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skewed nmig mcmc2	2.00 (0.09)	0.006	0.077	2.03 (0.09)	0.009	0.096	2.03 (0.17)	0.027	0.163
gen skew re	1.95 (0.09)	0.009	0.093	1.93 (0.10)	0.013	0.116	1.93 (0.18)	0.037	0.192
logit nmig mcmc2	2.01 (0.08)	0.006	0.075	2.00 (0.10)	0.009	0.094	2.00 (0.16)	0.025	0.157
logit re	1.98 (0.09)	0.006	0.079	1.87 (0.10)	0.026	0.161	1.73 (0.17)	0.098	0.314
(d) $m = 100$, n_i uniformly ranges from 12 to 30									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.03)	-	-	2.00 (0.03)	-	-	2.00 (0.03)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	1.99 (0.06)	0.003	0.057	2.00 (0.07)	0.003	0.058	2.04 (0.11)	0.013	0.113
gen skew re	1.97 (0.06)	0.004	0.065	1.95 (0.07)	0.006	0.076	1.79 (0.12)	0.053	0.231
logit nmig mcmc2	1.99 (0.06)	0.003	0.057	2.00 (0.07)	0.003	0.058	2.03 (0.12)	0.013	0.114
logit re	1.98 (0.06)	0.004	0.060	1.94 (0.07)	0.007	0.083	1.78 (0.12)	0.057	0.238

Table 3.6: Simulation results under true probit model of different scenarios: (a) $m = 50$, n_i uniformly ranges from 6 to 15; (b) $m = 50$, n_i uniformly ranges from 12 to 30; (c) $m = 100$, n_i uniformly ranges from 6 to 15; (d) $m = 100$, n_i uniformly ranges from 12 to 30 and three levels of percentage of treated units in terms of MSE, including mean and standard deviation of $\widetilde{\text{ATE}}_{1,0}$. The proposed generalized skewed link with nmig method is labeled as "gen skew nmig" with three comparing models: (1) the generalized skewed link with random effect model (labeled as "gen skew re") that uses the normal priors for the cluster effects; (2) the standard probit link with nmig model (labeled as "probit nmig") that uses the nmig priors for the cluster effects; and (3) the standard probit link with random effect model (labeled as "probit re") that uses the normal priors for the cluster effects.

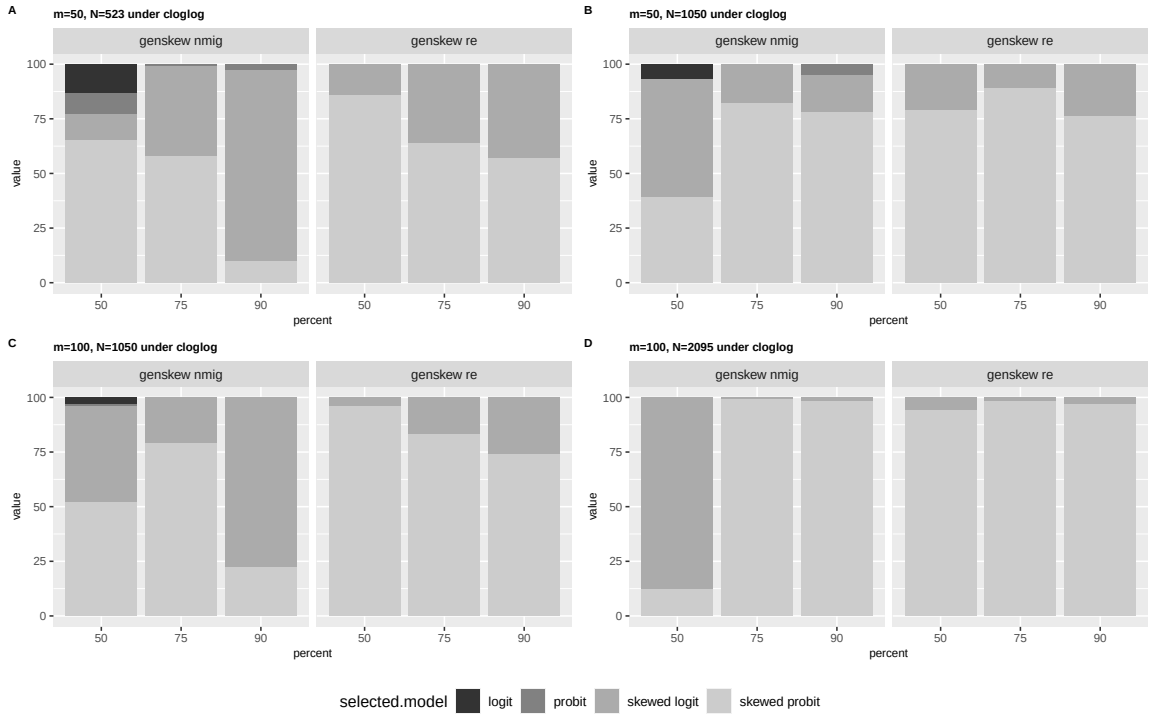
(a) $m = 50$, n_i uniformly ranges from 6 to 15									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.01 (0.06)	-	-	2.01 (0.06)	-	-	2.01 (0.06)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	2.02 (0.15)	0.019	0.138	2.05 (0.17)	0.027	0.163	2.01 (0.35)	0.117	0.341
gen skew re	1.91 (0.16)	0.030	0.174	1.99 (0.16)	0.022	0.150	2.03 (0.31)	0.095	0.308
probit nmig mcmc2	2.02 (0.15)	0.019	0.137	2.05 (0.17)	0.027	0.165	2.11 (0.22)	0.121	0.348
probit re	1.91 (0.15)	0.029	0.171	1.89 (0.16)	0.036	0.189	1.38 (0.29)	0.464	0.681
(b) $m = 50$, n_i uniformly ranges from 12 to 30									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.05)	-	-	2.00 (0.05)	-	-	1.99 (0.05)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	1.99 (0.10)	0.010	0.100	2.01 (0.11)	0.011	0.107	2.01 (0.18)	0.034	0.185
gen skew re	1.95 (0.11)	0.013	0.112	1.96 (0.11)	0.013	0.116	1.87 (0.20)	0.056	0.236
probit nmig mcmc2	1.99 (0.10)	0.010	0.100	2.01 (0.11)	0.011	0.107	2.00 (0.18)	0.035	0.187
probit re	1.95 (0.10)	0.012	0.111	1.94 (0.11)	0.015	0.123	1.65 (0.19)	0.158	0.398
(c) $m = 100$, n_i uniformly ranges from 6 to 15									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.05)	-	-	2.00 (0.05)	-	-	2.00 (0.04)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skewed nmig mcmc2	2.02 (0.11)	0.011	0.107	2.02 (0.11)	0.012	0.109	1.95 (0.22)	0.051	0.226
gen skew re	1.90 (0.10)	0.020	0.142	1.93 (0.12)	0.017	0.130	1.94 (0.18)	0.038	0.194
probit nmig mcmc2	2.03 (0.10)	0.011	0.107	2.02 (0.11)	0.012	0.109	2.04 (0.17)	0.032	0.180
probit re	1.90 (0.10)	0.019	0.138	1.86 (0.12)	0.032	0.180	1.35 (0.20)	0.456	0.675
(d) $m = 100$, n_i uniformly ranges from 12 to 30									
	50%			75%			90%		
mean (SD) of $\widetilde{\text{ATE}}_{1,0}$	2.00 (0.03)	-	-	2.00 (0.03)	-	-	2.00 (0.03)	-	-
	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB	mean (sd)	MSE	rMSB
gen skew nmig mcmc2	1.98 (0.07)	0.005	0.073	2.01 (0.08)	0.006	0.075	2.00 (0.14)	0.016	0.128
gen skew re	1.94 (0.07)	0.008	0.091	1.94 (0.08)	0.009	0.095	1.77 (0.14)	0.069	0.263
probit nmig mcmc2	1.98 (0.07)	0.005	0.073	2.01 (0.08)	0.006	0.075	2.00 (0.13)	0.016	0.125
probit re	1.94 (0.07)	0.008	0.091	1.93 (0.08)	0.009	0.097	1.62 (0.13)	0.163	0.404

(b) $m = 50$, n_i uniformly ranges from 12 to 30; (c) $m = 100$, n_i uniformly ranges from 6 to 15; (d) $m = 100$, n_i uniformly ranges from 12 to 30. It is worth noting some general findings. First, as the percentage of treated observations increases, the MSE of the proposed method and other methods increase monotonously. However, our proposed method still outperforms the others in most of the simulation scenarios, especially when the true model is not symmetric, that is, the true model of generating treatment assignment is the Cloglog link function. Second, as the sample size or the number of clusters increases and the percentage of treated observation remains the same, the proposed generalized skewed link with nmig performs better in terms of the smallest MSE in most of scenarios. This advantage becomes more pronounced when the number of clusters is large or sample size is small (sub scenarios (a) and (c)) with asymmetric generating true model for treatment assignment. On the other hand, MSE increases when the imbalance of treatment groups gets more extreme within the same m and n_i. Nevertheless, the proposed method still performs well compared to other methods in most of scenarios. Lastly, under the standard symmetric link function of generating the treatment assignment, the proposed method also performs well compared to the other methods. This is because the standard logit or probit link functions are special cases of the proposed generalized skewed link function.

We also investigate the proposed method with regard to the choice of symmetric and skewed link functions within the MCMC iterations. The correct choice for the different scenarios could be expected. The vertical stacked bar charts present the choices across 100 replications under different true generating models in Figures 3.1, 3.2 and 3.3. These graphs show the proposed generalized skewed link function with the nmig prior setting can determine the necessary of skewness, when the true model is asymmetric. Note that the generalized skewed link function with random effects is not as effective at making this determination. As discussed in [45], the fixed effect model could lead to unstable propensity score estimates with large number of clusters and small cluster size. Furthermore, it might produce unstable treatment effect estimation when using these unstable propensity score estimates. Also, the choice of the skewness parameter δ_1 is controlled by the Bernoulli binary variable η. The associated probability of the Bernoulli distribution for η is more likely to be inflated using random effect prior rather than nmig. More specifically, the binary variable η is more likely to generated as a 1 under random effect prior, than under the nmig prior setting; or, in order words, the proposed generalized skewed link function with random effects forces a skewed link function even it is not necessary. A detailed explanation with empirical results based on a

simulated data example is included in Appendix.

Figure 3.1: Simulation results: Number of each selected model out of 100 replications under true model is Cloglog, where the four selected models are: (1)the generalized skewed logit link model(labeled as "skewed logit"); (2) the generalized skewed probit link model(labeled as "skewed probit"); (3) the standard logit link model (labeled as "logit"); and (4) the standard probit link model (labeled as "probit"). For each plot, the x-axis denotes the different percentage level of treated observations, the y-axis is the numbers of replications out of 100 in each model.



3.6 Application of the Proposed Method

Numerous researchers have mentioned sleep insufficiency as a possible global health issue. Sleeping less than the recommended hours from National Sleep Foundation Sleep in America Poll may increase the risk for cardiometabolic diseases and obesity. There are some studies to investigate the relationship of sleep duration and HDL cholesterol, triglyceride and/or LDL cholesterol. Aho et.al [1] used an experimental and two epidemiological cohorts to specifically address how sleep insufficiency contributed to these factors. They found that prolonged sleep deprivation modifies inflammation, cholesterol pathways and serum lipoproteins, inducing potentially higher risk for cardiometabolic diseases. Broussard and Brady [11] pointed out that sleep has not only been considered necessary

Figure 3.2: Simulation results: Number of each selected model out of 100 replications under true model is logit, where the four selected models are: (1)the generalized skewed logit link model(labeled as "skewed logit"); (2) the generalized skewed probit link model(labeled as "skewed probit"); (3) the standard logit link model (labeled as "logit"); and (4) the standard probit link model (labeled as "probit"). For each plot, the x-axis denotes the different percentage level of treated observations, the y-axis is the numbers of replications out of 100 in each model.

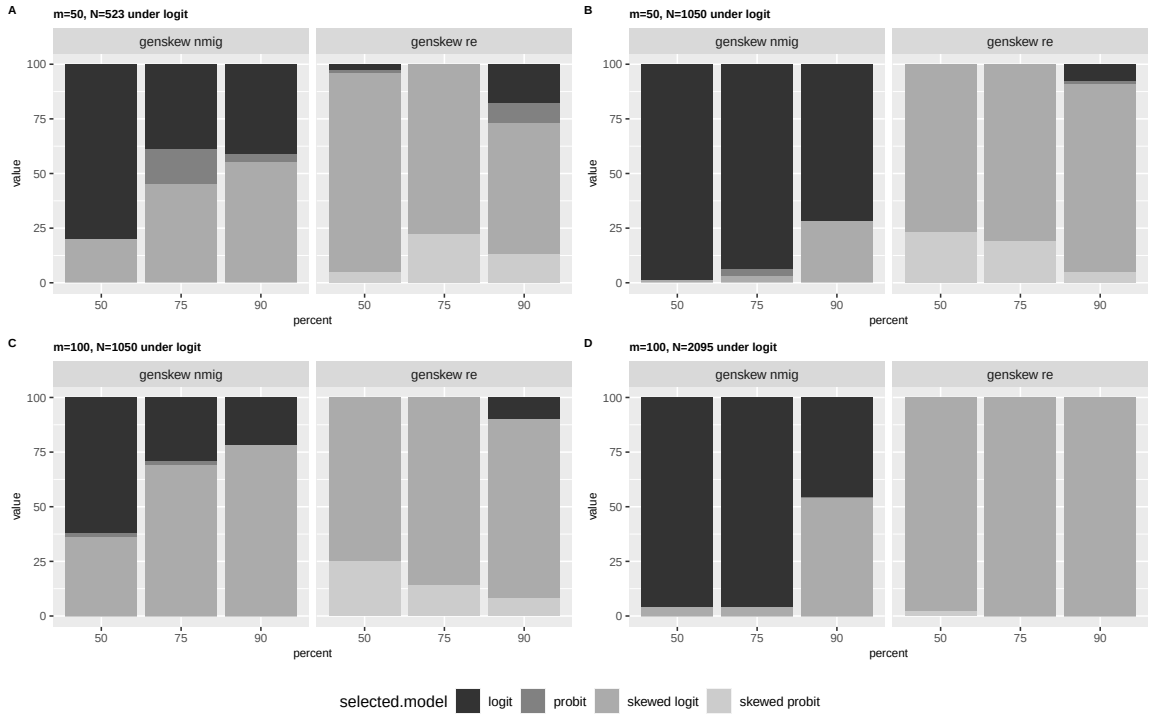
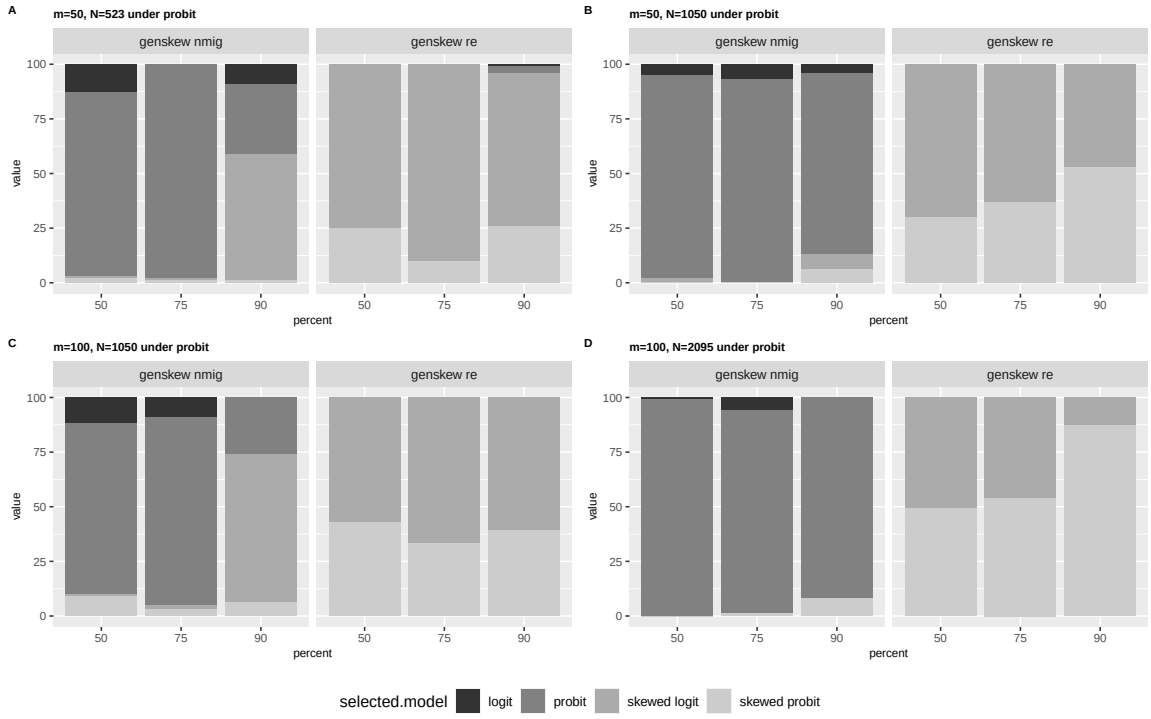


Figure 3.3: Simulation results: Number of each selected model out of 100 replications under true model is probit, where the four selected models are: (1)the generalized skewed logit link model(labeled as "skewed logit"); (2) the generalized skewed probit link model(labeled as "skewed probit"); (3) the standard logit link model (labeled as "logit"); and (4) the standard probit link model (labeled as "probit"). For each plot, the x-axis denotes the different percentage level of treated observations, the y-axis is the numbers of replications out of 100 in each model.



for the brain, but also plays a critical role in modulating energy metabolism in peripheral tissues. More importantly, they suggested these links may also have important clinical contributions to the global obesity epidemic. Kaneita et.al [37] examined the association between sleep duration and a high serum triglyceride, low HDL cholesterol, or high LDL cholesterol level at the individual level. They concluded there is a close association between sleep duration and serum lipid and lipoprotein levels. However, there are some researchers that suggest the association between sleep and different lipids groups is inconsistent. O'Keeffe et. al [60] claimed their results do not support that short-term, reduced sleep duration affects the lipid profile negatively for young and normal-weight individuals.

Another issue worth noting is the impact of age of these results. There are several studies in the literature focused on the relationships between sleep and lipid profile for specific age range. Kong et. al [40] studied the associations of sleep duration with obesity and serum lipid profiles in children and adolescents aged 6 to 20 years. Their results indicated decreased sleep duration was associated with obesity and atherogenic dyslipidemia in young school children. However, Lin et. al[50] performed a study restricted to middle-aged and elderly individuals. They concluded that there is a U-shaped relationship between sleep duration and HDL-C levels, that is, ¿ 7 hours or ¡ 6 hours sleep duration increased the risk of low serum HDL-C levels. In order to control for the unclear relationship between age and lipid profile outcomes, we consider the age ranges as clusters instead of including age as a covariate in the estimation models for this application.

The application used to illustrate the proposed method is a data set on lipid profile (HDL cholesterol/triglyceride/LDL cholesterol) and sleep duration with treatments defined as $Z_{ij} = 1$ for ≥ 6 hours sleep and $Z_{ij} = 0$ for < 6 hours sleep. The data are from the National Health and Nutrition Examination Survey (NHANES) for 2013-2014. After removing abnormal and missing values, we include 1204 participants. These participants have complete information including marital status, sitting time and race to be used as background covariates. Of the 1204 participants, 1065 (88%) have at least 6 hours of sleep ($Z_{ij} = 1$). Following the same hyper-parameter settings discussed in Section 3.5.2, and based on an MCMC sample of 11,000 iterations with a 1,000 burn-ins, the results of the estimated treatment effect, that is, $\hat{\alpha}$ in equation 3.12, are reported in Table 3.7. The shadowed 95% highest posterior density (HPD) interval of ATE are those that do not include zero and are considered significant. From Table 3.7, the proposed method results in all insignificant findings with the smaller standard deviations (SD) in (i) and (iii) cases, while the logit link approach with "nmig" or "re" yields a positive effect of the treatment for all three lipid profile outcomes. Additionally, we

report the selected link function for propensity score estimation based on 11,000 iterations MCMC sample with a 1,000 burn-ins for the proposed generalized skewed link function in Table 3.8. Since the true generating mechanism for treatment assignment in the data set is unknown, the simulation results are used as reference. From Table 3.8 we note that when the skewed link functions are selected, and majority of iterations also select the skewed logit link for propensity score estimation, no matter the "nmig" or "re" prior is used for cluster effects. As discussed in Section 3.5.2, the proposed model appears to correctly select the underlying true model generating the treatment assignment. This suggests that for this data set, an asymmetric link function is the true underlying model for treatment assignment.

From this data application, our proposed method serves as a suitable tool to evaluate the causal effects of the treatment with the existence of clusters and an unbalanced percentage of treated observations. Furthermore, we gain some insights into the actual treatment assignment mechanism which is impossible to observe in real-world observational studies, but more research is certainly needed.

Table 3.7: Summary Results of the estimated ATE for lipid profile data set with four different approaches under three outcome values separately: (i) HDL cholesterol is the outcome of interest, (ii) triglyceride cholesterol is the outcome of interest and (iii) LDL cholesterol is the outcome of interest to adjusting for covariates and clusters. Estimate (mean), standard deviation (SD) and 95% posterior credible intervals are based on 11,000 Monte Carlo samples after 1,000 burn-in periods.

(i) HDL cholesterol				
Approach	Mean	SD	**95% HPD Int**	
			Lower	Upper
gen skewed nmig MCMC2	0.551	1.464	-2.217	3.478
gen skewed re	0.490	1.437	-2.405	3.341
logit nmig MCMC2	4.299	1.465	1.452	7.100
logit re	4.220	1.450	1.364	7.032
probit nmig MCMC2	5.033	1.443	2.261	7.900
probit re	4.459	1.437	1.621	7.253
(ii) triglyceride cholesterol				
	Mean	SD	**95% HPD Int**	
			Lower	Upper
gen skewed nmig MCMC2	2.081	6.279	-10.459	14.051
gen skewed re	1.589	6.276	-10.229	14.367
logit nmig MCMC2	16.730	6.260	5.115	29.145
logit re	17.239	6.168	5.622	29.790
probit nmig MCMC2	20.382	6.186	8.260	32.478
probit re	18.788	6.099	6.448	30.424
(iii) LDL cholesterol				
	Mean	SD	**95% HPD Int**	
			Lower	Upper
gen skewed nmig MCMC2	-3.186	3.221	-9.353	3.213
gen skewed re	-3.345	3.151	-9.535	2.913
logit nmig MCMC2	8.283	3.345	2.078	14.955
logit re	7.711	3.254	1.318	13.981
probit nmig MCMC2	10.725	3.303	4.234	17.158
probit re	7.830	3.770	-0.432	14.160

Table 3.8: Summary results of number of iterations selected for each link function based on combination of skewness δ_1 and symmetric link functions δ_2 (logit or probit link) for propensity score estimation based on 11,000 iterations MCMC samples with a 1,000 burn-ins for the proposed generalized skewed link function.

	logit	probit	skewl	skewp
gen skewed nmig	0	0	9917	83
gen skewed re	0	0	8777	1223

Chapter 4

Summary and Conclusions

Two problems are addressed in this study. For the first problem, we address the issue of performing causal analysis when both post-treatment variable and clustering exist in a data set. Specifically, for this problem, following the idea of principal stratification, we define the clustering structure as random effects with a spike and slab prior in a Bayesian hierarchical model. With the spike and slab settings, the selection procedure for the cluster effects can be done by the posterior inclusion probabilities from a single MCMC sample. As a result, a parsimonious model which only contains clusters with large effects on the outcome can be obtained without much computational cost. We also develop an efficient MCMC sampling algorithm by using data augmentation. From the results of six simulation scenarios and two data applications, we confirm that the clustering structure must be considered to yield reasonable treatment effect estimates in the chosen post-treatment variable strata, and that the proposed method can handle more extreme cases (a large number of clusters or small sample sizes) than other competing methods.

For the second problem, we address the issue of performing causal inference when both clustering exists, and the number of observations in the treatment groups is very unbalanced in a data set. We develop a two-step Bayesian framework using a proposed generalized skewed link function, where the skewness follows a Dirac-spike prior. We define the clustering structure as random effects with a spike and slab prior in a Bayesian hierarchical model which allows selection of significant cluster effects from a single MCMC sample. We also develop an efficient MCMC sampling algorithm by using Pólya-Gamma latent variables for logit regression.

The proposed approaches can be extended to address even more complicated data struc-

tures for both methods. In modelling the latent strata, we follow the probit settings by [70] for the first problem mainly because of its tractable full conditionals after a series of data augmentation. Alternatively, in future research, we could consider the multinomial logit model because the augmented structure of Pólya–Gamma distributions [68] still leads to an easy MCMC sampling scheme. Another possible future research direction is to extend the method to the cases with multiple treatment groups or multilevel post-treatment variables. Due to an inflated number of principal strata, some model assumptions or regularizations on the number of latent groups may be needed to reduce the computational burden. For the second problem, an obvious future extension would be to allow correlation between covariates and unmeasured clustered-level confounders. As discussed in [28], He proposed a sufficient statistic method for handling such relationships in propensity score estimation. Multilevel post-treatment variables could also be considered and conceptually straightforward. But the computation burden could be very large in this setting and cause identification problems due to the exponentially increasing number of underlying principal strata. It would be useful to consider the Dirichlet process mixture (DPM) model discussed in [82] for such post-treatment variables. Another direction would also explore categorical potential outcomes or multivariate outcomes [57] under post-treatment variable and clustered-confounder structure.

Appendices

Appendix A Bayesian Inference under the Spike and Slab Priors

Let $\mathbb{1}\{A\}$ denote the usual indicator function with value 1 when A is true and 0 otherwise, and let $\pi_{slab}()$ and $\pi_{spike}()$ separately denote the density functions of slab and spike components. We update $G^*_{ij,l}$, $v_{i,l}$, and $w_{i,g}$ via

$$G^*_{ij,l} \mid . \sim \begin{cases} N(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_l + \zeta_{i,l}, 1)\mathbb{1}\{G^*_{ij,l} < 0\} & \text{if } G_{ij} = l, \\ N(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_l + \zeta_{i,l}, 1)\mathbb{1}\{G^*_{ij,l} > 0\} & \text{if } G_{ij} \neq l, \end{cases}$$

$$v_{i,l} \mid . \sim \text{Ber}\left(p^*_{v_{i,l}} = \frac{\pi_{slab}(\zeta_{i,l})p_{v_{i,l}}}{\pi_{slab}(\zeta_{i,l})p_{v_{i,l}} + \pi_{spike}(\zeta_{i,l})(1-p_{v_{i,l}})}\right),$$

and

$$w_{i,g} \mid . \sim \text{Ber}\left(p^*_{w_{i,g}} = \frac{\pi_{slab}(\xi_{i,g})p_{w_{i,g}}}{\pi_{slab}(\xi_{i,g})p_{w_{i,g}} + \pi_{spike}(\xi_{i,g})(1-p_{w_{i,g}})}\right),$$

respectively. Also, let $C = (C_{11}, C_{12}, \ldots, C_{mn_m})^T$ be a $N \times m$ indicator matrix with $C_{ij} = (\mathbb{1}\{i = 1\}, \mathbb{1}\{i = 2\}, \ldots, \mathbb{1}\{i = m\})^T$. The full conditional distributions of remaining parameters are then given by

$$\boldsymbol{\beta}_l \mid . \sim N\left(\left(\boldsymbol{X}_{\setminus l}^T\boldsymbol{X}_{\setminus l} + \sigma_{\beta_l}^{-2}I_p\right)^{-1}\boldsymbol{X}_{\setminus l}^T(G^*_l - C\boldsymbol{\zeta}_l), \left(\boldsymbol{X}_{\setminus l}^T\boldsymbol{X}_{\setminus l} + \sigma_{\beta_l}^{-2}I_p\right)^{-1}\right),$$

$$\boldsymbol{\zeta}_l \mid . \sim N\left(\left(C_{\setminus l}^TC_{\setminus l} + \boldsymbol{\Phi}_{v_l}^{-1}\right)^{-1}C_{\setminus l}^T\left(\mathbf{G}^*_l - \boldsymbol{X}\boldsymbol{\beta}_l\right), \left(C_{\setminus l}^TC_{\setminus l} + \boldsymbol{\Phi}_{v_l}^{-1}\right)^{-1}\right),$$

$$\boldsymbol{\alpha}_g \mid . \sim N\left(\left(\boldsymbol{X}_g^T\boldsymbol{X}_g + \sigma_{\alpha_g}^{-2}\sigma_g^2I_p\right)^{-1}\boldsymbol{X}_g^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{Z}\gamma_g - C\boldsymbol{\xi}_g\right), \left(\sigma_g^{-2}\boldsymbol{X}_g^T\boldsymbol{X}_g + \sigma_{\alpha_g}^{-2}I_p\right)^{-1}\right),$$

$$\gamma_g \mid . \sim N\left(\left(\boldsymbol{Z}_g^T\boldsymbol{Z}_g + \sigma_{\gamma_g}^{-2}\sigma_g^2\right)^{-1}\boldsymbol{Z}_g^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{X}\boldsymbol{\alpha}_g - C\boldsymbol{\xi}_g\right), \left(\sigma_g^{-2}\boldsymbol{Z}_g^T\boldsymbol{Z}_g + \sigma_{\gamma_g}^{-2}\right)^{-1}\right),$$

$$\boldsymbol{\xi}_g \mid . \sim N\left(\left(C_g^TC_g + \sigma_g^2\boldsymbol{\Phi}_{\boldsymbol{w}_g}^{-1}\right)^{-1}C_g^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{X}\boldsymbol{\alpha}_g - \boldsymbol{Z}\gamma_g\right), \left(\sigma_g^{-2}C_g^TC_g + \boldsymbol{\Phi}_{\boldsymbol{w}_g}^{-1}\right)^{-1}\right),$$

$$\sigma^2_{\beta_l} \mid . \sim \text{InvGamma}\left(a_\beta + \frac{p}{2}, b_\beta + \frac{\boldsymbol{\beta}_l^T\boldsymbol{\beta}_l}{2}\right),$$

$$\sigma^2_{\alpha_g} \mid . \sim \text{InvGamma}\left(a_\alpha + \frac{p}{2}, b_\alpha + \frac{\boldsymbol{\alpha}_g^T\boldsymbol{\alpha}_g}{2}\right),$$

$$\sigma^2_{\gamma_g} \mid . \sim \text{InvGamma}\left(a_\gamma + \frac{1}{2}, b_\gamma + \frac{{\gamma_g}^2}{2}\right),$$

$$\sigma_g^2 \mid . \sim \text{InvGamma}\left(a_0 + \frac{N_g}{2}, b_0 + \frac{(\boldsymbol{Y}_g^{obs} - \boldsymbol{X}_g\boldsymbol{\alpha}_g - \boldsymbol{Z}_g\gamma_g - C_g\boldsymbol{\xi}_g)^T(\boldsymbol{Y}_g^{obs} - \boldsymbol{X}_g\boldsymbol{\alpha}_g - \boldsymbol{Z}_g\gamma_g - C_g\boldsymbol{\xi}_g)}{2}\right),$$

$$p_{v_{i,l}} \mid . \sim \text{Beta}\left(a_v + v_{i,l}, b_v + (1 - v_{i,l})\right),$$

$$p_{w_{i,g}} \mid . \sim \text{Beta}\left(a_w + w_{i,g}, b_w + (1 - w_{i,g})\right),$$

$$\sigma_\zeta^2 \mid . \sim \text{InvGamma}\left(a_\zeta + \frac{3m}{2}, b_\zeta + \frac{1}{2}\sum_{\forall i}\sum_{\forall l}\frac{\zeta_{i,l}^2}{\delta(v_{i,l})}\right),$$

$$\sigma_\xi^2 \mid . \sim \text{InvGamma}\left(a_\xi + \frac{4m}{2}, b_\xi + \frac{1}{2}\sum_{\forall i}\sum_{\forall g}\frac{\xi_{i,g}^2}{\delta(w_{i,g})}\right),$$

where $\mathbf{G}_l^* = (G_{11,l}^*, G_{12,l}^*, \ldots, \ldots, G_{mn_m,l}^*)^T$, $\boldsymbol{\zeta}_l = (\zeta_{1l}, \ldots, \zeta_{ml})^T$, $\boldsymbol{\xi}_g = (\xi_{1g}, \ldots, \xi_{mg})^T$, $N_g = \sum_{\forall i}\sum_{\forall j}\mathbb{1}\{G_{ij} = g\}$, $\boldsymbol{Y}_g^{obs} = (Y_{11}^{obs} \times \mathbb{1}\{G_{11} = g\}, Y_{12}^{obs} \times \mathbb{1}\{G_{12} = g\}, \ldots, Y_{mn_m}^{obs} \times \mathbb{1}\{G_{mn_m} = g\})^T$, $\boldsymbol{Z}_g = (Z_{11}^{obs} \times \mathbb{1}\{G_{11} = g\}, Z_{12}^{obs} \times \mathbb{1}\{G_{12} = g\}, \ldots, Z_{mn_m}^{obs} \times \mathbb{1}\{G_{mn_m} = g\})^T$, $\boldsymbol{X}_g = (\boldsymbol{X}_{11} \times \mathbb{1}\{G_{11} = g\}, \boldsymbol{X}_{12} \times \mathbb{1}\{G_{12} = g\}, \ldots, \boldsymbol{X}_{mn_m} \times \mathbb{1}\{G_{mn_m} = g\})^T$, $C_g = (C_{11} \times \mathbb{1}\{G_{11} = g\}, C_{12} \times \mathbb{1}\{G_{12} = g\}, \ldots, C_{mn_m} \times \mathbb{1}\{G_{mn_m} = g\})^T$, $\boldsymbol{X}_{\setminus l} = \boldsymbol{X}_l + \boldsymbol{X}_{l+1} + \cdots + \boldsymbol{X}_4$, $C_{\setminus l} = C_l + C_{l+1} + \cdots + C_4$, $\boldsymbol{\Phi}_{v_l} = \text{diag}(\delta(v_{1,l})\sigma_\zeta^2, \ldots, \delta(v_{m,l})\sigma_\zeta^2)$, and $\boldsymbol{\Phi}_{\boldsymbol{w}_g} = \text{diag}(\delta(w_{1,g})\sigma_\xi^2, \ldots, \delta(w_{m,g})\sigma_\xi^2)$ with $\delta(0) = \delta$ and $\delta(1) = 1$.

Appendix B True Parameters

Table 1: parameters and true value

(a) latent structure in principal strata

Parameter	Value
$\beta_{1,1}$	1.5
$\beta_{1,2}$	1.5
$\beta_{1,u}$	1
$\beta_{2,1}$	1
$\beta_{2,2}$	1
$\beta_{2,u}$	1.5
$\beta_{3,1}$	0.5
$\beta_{3,2}$	0.5
$\beta_{3,u}$	2

(b) outcome models

Parameter	Value
$\alpha_{1,1}$	2.5
$\alpha_{1,2}$	2.5
$\alpha_{1,z}$	2.5
$\alpha_{1,u}$	2.5
$\alpha_{2,1}$	1.5
$\alpha_{2,2}$	1.5
$\alpha_{2,z}$	1.5
$\alpha_{2,u}$	1.5
$\alpha_{3,1}$	2
$\alpha_{3,2}$	2
$\alpha_{3,z}$	2
$\alpha_{3,u}$	2
$\alpha_{4,1}$	1
$\alpha_{4,2}$	1
$\alpha_{4,z}$	1
$\alpha_{4,u}$	1
σ_1^2	1
σ_2^2	1.25
σ_3^2	1.5

Appendix C Bayesian Inference under Lasso Priors

With the same prior settings as Section 2.3 except

$$\zeta_{i,l} \mid \sigma_\zeta^2, \lambda_1 \sim \text{Laplace}\left(0, \frac{\sqrt{\sigma_\zeta^2}}{\lambda_1}\right),$$

$$\xi_{i,g} \mid \sigma_\xi^2, \lambda_2 \sim \text{Laplace}\left(0, \frac{\sqrt{\sigma_\xi^2}}{\lambda_2}\right),$$

$$\pi(\sigma_\xi^2) = \frac{1}{\sigma_\xi^2},$$

$$\pi(\sigma_\zeta^2) = \frac{1}{\sigma_\zeta^2},$$

$$\lambda_1^2 \sim \text{Gamma}(a_{\lambda_1}, b_{\lambda_1}),$$

$$\lambda_2^2 \sim \text{Gamma}(a_{\lambda_1}, b_{\lambda_1}),$$

and data augmentation in [64], the full conditionals are as follows:

$$G_{ij,l}^* \mid . \sim \begin{cases} N(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_l + \zeta_{i,l}, 1)\mathbb{1}\{G_{ij,l}^* < 0\} & \text{if } G_{ij} = l, \\ N(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_l + \zeta_{i,l}, 1)\mathbb{1}\{G_{ij,l}^* > 0\} & \text{if } G_{ij} \neq l, \end{cases}$$

$$\boldsymbol{\beta}_l \mid . \sim N\left(\left(\boldsymbol{X}_{\setminus l}^T\boldsymbol{X}_{\setminus l} + \sigma_{\beta_l}^{-2}I_p\right)^{-1}\boldsymbol{X}_{\setminus l}^T(G_l^* - C\boldsymbol{\zeta}_l), \left(\boldsymbol{X}_{\setminus l}^T\boldsymbol{X}_{\setminus l} + \sigma_{\beta_l}^{-2}I_p\right)^{-1}\right),$$

$$\boldsymbol{\zeta}_l \mid . \sim N\left(\left(C_{\setminus l}^TC_{\setminus l} + \sigma_\zeta^{-2}\boldsymbol{D}_{\tau_l}^{-1}\right)^{-1}C_{\setminus l}^T\left(\mathbf{G}_l^* - \boldsymbol{X}\boldsymbol{\beta}_l\right), \left(C_{\setminus l}^TC_{\setminus l} + \sigma_\zeta^{-2}\boldsymbol{D}_{\tau_l}^{-1}\right)^{-1}\right),$$

$$\boldsymbol{\alpha}_g \mid . \sim N\left(\left(\boldsymbol{X}_g^T\boldsymbol{X}_g + \sigma_{\alpha_g}^{-2}\sigma_g^2I_p\right)^{-1}\boldsymbol{X}_g^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{Z}\gamma_g - C\boldsymbol{\xi}_g\right), \left(\sigma_g^{-2}\boldsymbol{X}_g^T\boldsymbol{X}_g + \sigma_{\alpha_g}^{-2}I_p\right)^{-1}\right),$$

$$\gamma_g \mid . \sim N\left(\left(\boldsymbol{Z}_g^T\boldsymbol{Z}_g + \sigma_{\gamma_g}^{-2}\sigma_g^2\right)^{-1}\boldsymbol{Z}_g^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{X}\boldsymbol{\alpha}_g - C\boldsymbol{\xi}_g\right), \left(\sigma_g^{-2}\boldsymbol{Z}_g^T\boldsymbol{Z}_g + \sigma_{\gamma_g}^{-2}\right)^{-1}\right),$$

$$\boldsymbol{\xi}_g \mid . \sim N\left(\left(C_g^TC_g + \sigma_\xi^{-2}\sigma_g^2\boldsymbol{D}_{\varphi_g}^{-1}\right)^{-1}C_g^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{X}\boldsymbol{\alpha}_g - \boldsymbol{Z}\gamma_g\right), \left(\sigma_g^{-2}C_g^TC_g + \sigma_\xi^{-2}\boldsymbol{D}_{\varphi_g}^{-1}\right)^{-1}\right),$$

$$\sigma_{\beta_l}^2 \mid . \sim \text{InvGamma}\left(a_\beta + \frac{p}{2}, b_\beta + \frac{\boldsymbol{\beta}_l^T\boldsymbol{\beta}_l}{2}\right),$$

$$\sigma_{\alpha_g}^2 \mid . \sim \text{InvGamma}\left(a_\alpha + \frac{p}{2}, b_\alpha + \frac{\boldsymbol{\alpha}_g^T\boldsymbol{\alpha}_g}{2}\right),$$

$$\sigma_{\gamma_g}^2 \mid . \sim \text{InvGamma}\left(a_\gamma + \frac{1}{2}, b_\gamma + \frac{{\gamma_g}^2}{2}\right),$$

$$\sigma_g^2 \mid . \sim \text{InvGamma}\left(a_0 + \frac{N_g}{2}, b_0 + \frac{(\boldsymbol{Y}_g^{obs} - \boldsymbol{X}_g\boldsymbol{\alpha}_g - \boldsymbol{Z}_g\gamma_g - C_g\boldsymbol{\xi}_g)^T(\boldsymbol{Y}_g^{obs} - \boldsymbol{X}_g\boldsymbol{\alpha}_g - \boldsymbol{Z}_g\gamma_g - C_g\boldsymbol{\xi}_g)}{2}\right),$$

$$\sigma_{\zeta}^{2} \mid . \sim \text{InvGamma}\left(\frac{3m}{2}, \frac{1}{2}\sum_{\forall i}\sum_{\forall l}\frac{\zeta_{i,l}^{2}}{\tau_{i,l}}\right),$$

$$\sigma_{\xi}^{2} \mid . \sim \text{InvGamma}\left(\frac{4m}{2}, \frac{1}{2}\sum_{\forall i}\sum_{\forall g}\frac{\xi_{i,g}^{2}}{\varphi_{i,g}}\right),$$

$$\lambda_{1}^{2} \mid . \sim \text{Gamma}\left(a_{\lambda_1} + 3m, b_{\lambda_1} + \frac{1}{2}\sum_{\forall i}\sum_{\forall l}\tau_{i,l}\right),$$

$$\lambda_{2}^{2} \mid . \sim \text{Gamma}\left(a_{\lambda_2} + 4m, b_{\lambda_2} + \frac{1}{2}\sum_{\forall i}\sum_{\forall g}\varphi_{i,g}\right),$$

$$\tau_{i,l}^{-1} \mid . \sim \text{InvGaussian}\left(\sqrt{\frac{\lambda_{1}^{2}\sigma_{\zeta}^{2}}{\zeta_{i,l}^{2}}}, \lambda_{1}^{2}\right),$$

$$\varphi_{i,g}^{-1} \mid . \sim \text{InvGaussian}\left(\sqrt{\frac{\lambda_{2}^{2}\sigma_{\xi}^{2}}{\xi_{i,g}^{2}}}, \lambda_{2}^{2}\right),$$

where $\tau_{i,l}$ and $\varphi_{i,g}$ are non-negative latent variables, $\boldsymbol{D}_{\tau_l} = \text{diag}(\tau_{1,l}, \tau_{2,l}, \ldots, \tau_{m,l})$, and $\boldsymbol{D}_{\varphi_g} = \text{diag}(\varphi_{1,g}, \varphi_{2,g}, \ldots, \varphi_{m,g})$.

Appendix D The Pólya-Gamma technique in Bayesian inference for logistic regression

The Pólya-Gamma variables for Bayesian inference of logistic regression is proposed by Polson [68].

$$X \sim PG(b,c) \longrightarrow X \overset{D}{=} \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{g_k}{(k-1/2)^2+c^2/(4\pi^2)}$$

This technique avoid the traditional heavy computation of Metropolis-Hasting for logistic regression in Bayesian inference. The main result is representing binomial likelihood of log-odds as mixtures of Gaussians with respect to a Pólya-Gamma distribution as following

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b}e^{k\psi}\int_0^{\infty} e^{-w\psi^2/2}p(w)dw,$$

where $k = a - b/2$ and $w \sim \mathrm{PG}(b,0)$, which denoted w is generated from the Pólya-Gamma distribution. When ψ is a linear function of factors, saying $\psi = \boldsymbol{X}\boldsymbol{\beta}$, then the integrand part is the Gaussian kernel in $\boldsymbol{\beta}$. In [68] they also show that conditional distribution of $w \mid \psi$ is also followed the Pólya-Gamma distribution, which gives a strategy based on Gibbs sampling for posterior inference of $\boldsymbol{\beta}$. In the second problem, specifically for illustration, applying for logit link with naive method, notationally as in the following equation

$$PS_{ij}^{\text{naive}} = \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}\right)}$$

The posterior distribution of $\boldsymbol{\beta}$ conditional on data is

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \cdot) &\propto \pi(\boldsymbol{\beta}) \prod_{i,j} \left(\frac{\exp\left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}\right)} \right)^{z_{ij}} \left(\frac{1}{1 + \exp\left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}\right)} \right)^{1 - z_{ij}} \\
&= \pi(\boldsymbol{\beta}) \prod_{i,j} \frac{\left(\exp\left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}\right)\right)^{z_{ij}}}{\left(1 + \exp\left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}\right)\right)^{1}} \propto \pi(\boldsymbol{\beta}) \prod_{i,j} \exp\left(k_{ij} \boldsymbol{X}_{ij}^T \boldsymbol{\beta} - w_{ij} (\boldsymbol{X}_{ij}^T \boldsymbol{\beta})^2 / 2\right) \\
&\propto \pi(\boldsymbol{\beta}) \prod_{i,j} \exp\left(\frac{w_{ij}}{2} \left(\boldsymbol{X}_{ij}^T \boldsymbol{\beta} - k_{ij} / w_{ij} \right)^2 \right)
\end{aligned}
$$

given $\pi(\boldsymbol{\beta}) \sim \text{MVN}(\mathbf{0}, \sigma_\beta^{-2} I_p)$

$$
\propto \pi(\boldsymbol{\beta}) \exp\left(-\frac{1}{2} \left(L - \boldsymbol{X}\boldsymbol{\beta} \right)^T \boldsymbol{\Omega} \left(L - \boldsymbol{X}\boldsymbol{\beta} \right) \right)
$$

Reorganized the terms with $\boldsymbol{\beta}$, we will have,

$$
\begin{aligned}
w_{ij} \mid . &\sim \text{PG}(1, \boldsymbol{X}_{ij}^T \boldsymbol{\beta}) \\
\boldsymbol{\beta} \mid . &\sim \text{N}(\mu^*, \Sigma^*) \\
\Sigma^* &= \left(\boldsymbol{X}^T \boldsymbol{\Omega} \boldsymbol{X} + 1/\sigma_\beta^2 I_p \right)^{-1} \\
\mu^* &= \Sigma^* \left(\boldsymbol{X}^T k \right)
\end{aligned}
$$

where $w = \{w_{ij}\}_{i,j}$, $\boldsymbol{\Omega}$ is diagonal matrix with elements w_{ij}, and $k = \{k_{ij} \mid k_{ij} = z_{ij} - 1/2\}$

Appendix E Posterior sampling distributions

We proposed two step Bayesian inference for propensity score estimation and outcome model respectively.

E.1 Propensity score modeling

There will be four possible cases we consider based on the combination values of δ_1 and δ_2. For simplicity, we use F_1 and F_2 as logit link and probit link. If δ_1 is 0, then it reduces to corresponding traditional symmetric logit/probit link. If δ_1 is nonzero, we consider the skewed logit/probit link similarly as discussed in [13]. The detailed processes are described in following sections. To update δ_2, we draw from a Bernoulli distribution where the probability is calculated as, if $\delta_1 = 0$,

$$p(\delta_2^{(k)} = 1) = \frac{A}{A+B} = \frac{1}{1+\exp(\log(B) - \log(A))}$$

where

$$A = \prod_{i,j}[F_{logit}(\boldsymbol{\beta}^{(k)}\boldsymbol{X}_{ij} + \zeta_i^{(k)})]^{z_{ij}}[1 - F_{logit}(\boldsymbol{\beta}^{(k)}\boldsymbol{X}_{ij} + \zeta_i^{(k)})]^{1-z_{ij}}$$

$$B = \prod_{i,j}[F_{probit}(\boldsymbol{\beta}^{(k)}\boldsymbol{X}_{ij} + \zeta_i^{(k)})]^{z_{ij}}[1 - F_{probit}(\boldsymbol{\beta}^{(k)}\boldsymbol{X}_{ij} + \zeta_i^{(k)})]^{1-z_{ij}}$$

Otherwise, if δ_1 is nonzero, then, we can sample from Bernoulli with probability

$$p(\delta_2^{(k)} = 1) = \frac{A}{A+B} = \frac{1}{1+\exp(\log(B) - \log(A))}$$

where

$$A = \prod_{i,j}(ps_{ij}^{skewl})^{z_{ij}}(1 - (ps_{ij}^{skewl}))^{1-z_{ij}}$$

$$B = \prod_{i,j}(ps_{ij}^{skewp})^{z_{ij}}(1 - (ps_{ij}^{skewp}))^{1-z_{ij}}$$

We will utilize Bayesian technique with conjugate prior distributions to estimate the causal effects. Markov chain Monte Carlo (MCMC) methods are used to implemented posterior samples based on a class of chains for the Bayesian system. Conditioning on the value of δ_1 and δ_2, after taking series of data augmentation techniques as discussed in previous section, we could updated the remaining parameters from their posterior distribution at each iteration of MCMC. Within 10000

iterations after 1000 burn-in period, the corresponding posterior samples of most often selected link model were used to make inference. For instance, 60% of iterations are based on skewed probit link model, the remaining 40% iterations were split by skewed logit link, logit link and probit link. We only focused on the iterations based on skewed probit link model for inference.

E.1.1 Posterior samples for skewed logit link

When $F = \exp(.)/(1+\exp(.))$, equation 3.6 gives skewed logit model, which is another special case from a class of scale mixtures of normal distributions [12]. Under proposed Bayesian inference framework, Pólya-Gamma strategy introduced by Polson [68] is used. It is very beneficial for releasing computational burden for binomial likelihoods. It is essential for many complicated cases and diseases studies in clinical experiments.

The joint posterior distribution 3.10 becomes

$$
\begin{aligned}
&p(\boldsymbol{t}, \boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1 \mid \boldsymbol{Z}, \boldsymbol{X}) \\
&\propto \left(\prod_{i,j} \left[\frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1+\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)} \right]^{z_{ij}} \left[1 - \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1+\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)} \right]^{1-z_{ij}} g(t_{ij}) \right) p(\boldsymbol{\zeta}, \boldsymbol{\beta}, \delta)
\end{aligned}
$$

For the full conditional posterior distribution of $t_{ij} \mid .$ is given by

$$
\begin{aligned}
p(t_{ij} \mid .) &\propto \left[\frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1+\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)} \right]^{z_{ij}} \left[1 - \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1+\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)} \right]^{1-z_{ij}} g(t_{ij}) \\
&\propto \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)^{z_{ij}}}{\left(1+\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)\right)^{1}} e^{-(t_{ij}^2/2)}
\end{aligned}
$$

using Polygamma rewritten $\propto \exp\left(-\frac{1}{2}\left(L_{ij} - \boldsymbol{X}_{ij}^T\boldsymbol{\beta} - \zeta_{i,1} - \delta_1 t_{ij}\right)^2 \omega_{ij}\right) e^{-(t_{ij}^2/2)}$

where

$$
\begin{aligned}
&\omega_{ij} \mid . \sim \mathrm{PG}(1, \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) \\
&k_{ij} = z_{ij} - 1/2 \\
&L_{ij} = k_{ij}/w_{ij}
\end{aligned}
$$

therefore the posterior sample of t_{ij} from truncated normal as follows

$$t_{ij} \mid . \sim N(\frac{(L_{ij} - \boldsymbol{X}_{ij}^T\boldsymbol{\beta} - \zeta_{i,1})\delta_1\omega_{ij}}{\delta_1^2\omega_{ij} + 1}, \frac{1}{\delta_1^2\omega_{ij} + 1})\mathbb{1}_{(0,\infty)}(t_{ij})$$

Correspondingly, for the posterior distribution of $\delta_1 \mid .$, if $\eta = 0$, then posterior sample $\delta_1 \mid . = 0$. If $\eta = 1$, then

$$\begin{aligned} p(\delta_1 \mid .) &\propto \prod_{i,j} \left[\frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}\right]^{z_{ij}} \left[1 - \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}\right]^{1-z_{ij}} p(\delta_1) \\ &\propto \prod_{i,j} \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)^{z_{ij}}}{\left(1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)\right)^1} p(\delta_1) \end{aligned}$$

using Polygamma rewritten $\propto p(\delta_1)\exp\left(-1/2(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})^T\boldsymbol{\Omega}(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})\right)$

where

$$\begin{aligned} &\omega_{ij} \mid . \sim \mathrm{PG}(1, \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) \\ &k_{ij} = z_{ij} - 1/2 \\ &L_{ij} = k_{ij}/w_{ij}, \quad \mathbf{L} = \{L_{ij}\} \\ &\boldsymbol{\Omega} = \mathrm{diag}(\omega_{ij}) \end{aligned}$$

therefore the posterior sample of δ_1 from

$$\delta_1 \mid . \sim N\left(\frac{\boldsymbol{t}^T\boldsymbol{\Omega}(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1)}{\boldsymbol{t}^T\boldsymbol{\Omega}\boldsymbol{t} + 1/\sigma_{\delta_1}^2}, \frac{1}{\boldsymbol{t}^T\boldsymbol{\Omega}\boldsymbol{t} + 1/\sigma_{\delta_1}^2}\right)$$

The posterior sample for η is needed to integrate out δ_1, based on the marginal distribution

$$p(\eta \mid .) \propto \pi(\eta)\int p(\mathbf{L}, \boldsymbol{Z} \mid \boldsymbol{\zeta}_1, \delta_1, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{X})p(\delta_1 \mid \eta)d\delta_1$$

Focus on the integral part,

$$\int p(\mathbf{L}, \boldsymbol{Z} \mid \boldsymbol{\zeta}_1, \delta_1, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{X})p(\delta_1 \mid \eta)d\delta_1$$

$$= \int \exp\left(-1/2\left(\mathbf{L} - \boldsymbol{X\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right)^T \boldsymbol{\Omega}(\mathbf{L} - \boldsymbol{X\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})\right) \frac{1}{\sqrt{2\pi}\sigma_{\delta_1}} \exp(-\frac{1}{2\sigma_{\delta_1}^2}\delta_1^2)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_{\delta_1}} \int \exp\left(-1/2\left(\mathbf{L}^* - \delta_1\boldsymbol{t}\right)^T \boldsymbol{\Omega}\left(\mathbf{L}^* - \delta_1\boldsymbol{t}\right) + \frac{1}{2\sigma_{\delta_1}^2}\delta_1^2\right) d\delta_1, \quad \mathbf{L}^* = \mathbf{L} - \boldsymbol{X\beta} - C\boldsymbol{\zeta}_1$$

$$\propto \frac{1}{\sigma_{\delta_1}} \exp(\frac{1}{2}\frac{((\mathbf{L}^*)^T\boldsymbol{\Omega t})^2}{\boldsymbol{t}^T\boldsymbol{\Omega t} + 1/\sigma_{\delta_1}^2}) \exp(-\frac{1}{2}(\mathbf{L}^*)^T\boldsymbol{\Omega}\mathbf{L}^*)(\frac{1}{\boldsymbol{t}^T\boldsymbol{\Omega t} + 1/\sigma_{\delta_1}^2})^{1/2}$$

Therefore the posterior sample of η is given by, $\eta \sim \text{Ber}(Prob)$ where

$$prob = \frac{1}{1 + \frac{1-p^\eta}{p^\eta}\frac{1}{R}}, \quad R = \frac{1}{\sigma_{\delta_1}} \exp(\frac{1}{2}\frac{((\mathbf{L}^*)^T\boldsymbol{\Omega t})^2}{\boldsymbol{t}^T\boldsymbol{\Omega t} + 1/\sigma_{\delta_1}^2})(\frac{1}{\boldsymbol{t}^T\boldsymbol{\Omega t} + 1/\sigma_{\delta_1}^2})^{1/2}$$

Finally for the last piece of Dirac spike, the posterior sample of $p^\eta \mid \eta$ is given by

$$\begin{aligned} p(p^\eta \mid \eta) &\propto \pi(p^\eta)p(\eta \mid p^\eta) \\ &\propto (p^\eta)^{a-1}(1-p^\eta)^{b-1}(p^\eta)^\eta(1-p^\eta)^{1-\eta} \end{aligned}$$

Therefore, the posterior sample $p^\eta \mid \eta \sim \text{Beta}(a + \eta, b + 1 - \eta)$.

Secondly, the full conditional posterior distribution of $\boldsymbol{\beta} \mid .$ is given by

$$\begin{aligned} p(\boldsymbol{\beta} \mid .) &\propto \prod_{i,j} \left[\frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}\right]^{z_{ij}} \left[1 - \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}\right]^{1-z_{ij}} p(\boldsymbol{\beta}) \\ &\propto \prod_{i,j} \frac{\exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij})^{z_{ij}}}{(1 + \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}))^1} p(\boldsymbol{\beta}) \end{aligned}$$

using Polygamma rewritten $\propto \exp(-\frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}^T\boldsymbol{\beta}) \exp\left(-1/2(\mathbf{L} - \boldsymbol{X\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})^T\boldsymbol{\Omega}(\mathbf{L} - \boldsymbol{X\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})\right)$

where

$$\begin{aligned} &\omega_{ij} \mid . \sim \text{PG}(1, \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) \\ &k_{ij} = z_{ij} - 1/2 \\ &L_{ij} = k_{ij}/w_{ij}, \quad \mathbf{L} = \{L_{ij}\} \\ &\boldsymbol{\Omega} = \text{diag}(\omega_{ij}) \end{aligned}$$

therefore the posterior sample of $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta} \mid . \sim MVN\left(\left(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X} + \frac{1}{\sigma_\beta^2 I_p}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}\left(\mathbf{L} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right), \left(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X} + \frac{1}{\sigma_\beta^2 I_p}\right)^{-1}\right)$$

Next, the full conditional posterior distribution of $(\boldsymbol{\zeta}_1 \mid .)$ is given by

$$\begin{aligned} p(\boldsymbol{\zeta}_1 \mid .) &\propto \prod_{i,j}\left[\frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}\right]^{z_{ij}}\left[1 - \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}{1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)}\right]^{1-z_{ij}} p(\boldsymbol{\zeta}_1) \\ &\propto \prod_{i,j}\frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)^{z_{ij}}}{\left(1 + \exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}\right)\right)^1} p(\boldsymbol{\zeta}_1) \end{aligned}$$

using Polygamma rewritten $\propto \exp\left(-1/2\boldsymbol{\zeta}_1^T\boldsymbol{\Phi}_v^{-1}\boldsymbol{\zeta}_1\right)\exp\left(-1/2\left(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right)^T\boldsymbol{\Omega}\left(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right)\right)$

where $\boldsymbol{\Phi}_v = \text{diag}(r(v_1)\sigma_1^2, \ldots, r(v_m)\sigma_1^2)$ be the variance-covariance matrix of prior distribution, where $r(0) = r$ and $r(1) = 1$. And

$$\begin{aligned} &\omega_{ij} \mid . \sim \text{PG}(1, \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}) \\ &k_{ij} = z_{ij} - 1/2 \\ &L_{ij} = k_{ij}/w_{ij}, \quad \mathbf{L} = \{L_{ij}\} \\ &\boldsymbol{\Omega} = \text{diag}(\omega_{ij}) \end{aligned}$$

Therefore, posterior sample $\boldsymbol{\zeta}_1$ from

$$\boldsymbol{\zeta}_1 \mid . \sim MVN\left((C^T\boldsymbol{\Omega}C + \boldsymbol{\Phi}_v^{-1})^{-1}C^T\boldsymbol{\Omega}(\mathbf{L} - \boldsymbol{X}\boldsymbol{\beta} - \delta_1\boldsymbol{t}), (C^T\boldsymbol{\Omega}C + \boldsymbol{\Phi}_v^{-1})^{-1}\right)$$

The remaining full conditional of parameters are as follows

$$\begin{aligned} &v_i \mid . \sim \text{Ber}\left(p_{v_i}^* = \frac{\pi_{slab}(\zeta_{i,1})p_{v_i}}{\pi_{slab}(\zeta_{i,1})p_{v_i} + \pi_{spike}(\zeta_{i,1})(1 - p_{v_i})}\right), \\ &\sigma_\beta^2 \mid . \sim \text{InvGamma}\left(a_\beta + \frac{p}{2}, b_\beta + \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2}\right), \\ &p_{v_i} \mid . \sim \text{Beta}\left(a_v + v_i, b_v + (1 - v_i)\right), \end{aligned}$$

$$\sigma_1^2 \mid . \sim \text{InvGamma}\left(a_1 + \frac{m}{2}, b_1 + \frac{1}{2}\sum_{\forall i}\frac{\zeta_{i,1}^2}{r(v_i)}\right),$$

E.1.2 Posterior samples for skewed probit link

When $F = \Phi(.)$, equation 3.6 gives skewed probit, which is one special cases from a class of scale mixtures of normal distributions [12]. As introduced previously for probit likelihood we introduce latent variable $\boldsymbol{w}$, then the joint posterior distribution of $(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1)$ is given by

$$\begin{aligned}
p(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1 \mid \boldsymbol{X}, \boldsymbol{Z}) &\propto p(\boldsymbol{t}, \boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1) p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1, \boldsymbol{X}) p(\boldsymbol{Z} \mid \boldsymbol{w}) \\
&\propto p(\boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1) g(\boldsymbol{t}) p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{\zeta}_1, \boldsymbol{\beta}, \delta_1, \boldsymbol{X}) p(\boldsymbol{Z} \mid \boldsymbol{w}) \\
&= p(\boldsymbol{\zeta}_1) p(\boldsymbol{\beta}) p(\delta_1) \prod_{i,j} g(t_{ij}) \prod_{i,j} N(w_{ij} \mid \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) p(z_{ij} \mid w_{ij})
\end{aligned}$$

The latent variable w_{ij} is sampled from

$$w_{ij} \mid . \sim \begin{cases} N(\boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) \mathbb{1}\{w_{ij} > 0\} & \text{if} \quad z_{ij} = 1 \\ N(\boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) \mathbb{1}\{w_{ij} < 0\} & \text{if} \quad z_{ij} = 0 \end{cases}$$

For the full conditional posterior distribution of $t_{ij} \mid .$ is given by

$$\begin{aligned}
p(t_{ij} \mid .) = p(t_{ij} \mid w_{ij}, \zeta_{i,1}, \boldsymbol{\beta}, \delta_1, \boldsymbol{X}_{ij}) &\propto g(t_{ij}) N(w_{ij} \mid \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) \\
&= \frac{2}{\sqrt{2\pi}} e^{-(t_{ij}^2/2)} N(w_{ij} \mid \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) \\
&\propto \exp(-t_{ij}^2/2) \exp\{-1/2(w_{ij} - \boldsymbol{X}_{ij}^T \boldsymbol{\beta} - \zeta_{i,1} - \delta_1 t_{ij})^2\}
\end{aligned}$$

therefore, posterior sample from truncated normal distribution as $t_{ij} \mid . \sim N\left(\frac{(w_{ij} - \boldsymbol{X}_{ij}^T \boldsymbol{\beta} - \zeta_{i,1})\delta_1}{1+\delta_1^2}, \frac{1}{1+\delta_1^2}\right) \mathbb{1}_{(0,\infty)}(t_{ij})$,

Correspondingly, for the posterior distribution of $\delta_1 \mid .$, if $\eta = 0$, then posterior sample $\delta_1 \mid . = 0$. If $\eta = 1$, then

$$\begin{aligned}
p(\delta_1 \mid .) &\propto p(\delta_1) \prod_{i,j} N(w_{ij} \mid \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) \\
&\propto \exp\left(-\frac{1}{2\sigma_{\delta_1}^2}\delta_1^2\right) \exp\left(-1/2(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1 \boldsymbol{t})^T (\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1 \boldsymbol{t})\right)
\end{aligned}$$

where $C = (C_{11}, C_{12}, \ldots, C_{mn_m})^T$ be a $N \times m$ indicator matrix with $C_{ij} = (\mathbb{1}\{i = 1\}, \mathbb{1}\{i = 2\}, \ldots, \mathbb{1}\{i = m\})^T$. Therefore, posterior sample from $\delta_1 \mid . \sim N(\frac{(\boldsymbol{w}-\boldsymbol{X}\boldsymbol{\beta}-C\boldsymbol{\zeta}_1)^T\boldsymbol{t}}{1/\sigma^2_{\delta_1}+\boldsymbol{t}^T\boldsymbol{t}}, \frac{1}{1/\sigma^2_{\delta_1}+\boldsymbol{t}^T\boldsymbol{t}})$

The posterior sample for η is needed to integrate out δ_1, based on the marginal distribution

$$p(\eta \mid .) \propto \pi(\eta) \int p(\boldsymbol{w}, \boldsymbol{Z} \mid \boldsymbol{\zeta}_1, \delta_1, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{X}) p(\delta_1 \mid \eta) d\delta_1$$

Focus on the integral part,

$$\begin{aligned}
&\int p(\boldsymbol{w}, \boldsymbol{Z} \mid \boldsymbol{\zeta}_1, \delta_1, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{X}) p(\delta_1 \mid \eta) d\delta_1 \\
&= \int \exp\left(-1/2\left(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right)^T\left(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right)\right) \frac{1}{\sqrt{2\pi}\sigma_{\delta_1}} \exp(-\frac{1}{2\sigma^2_{\delta_1}}\delta_1^2) \\
&= \frac{1}{\sqrt{2\pi}\sigma_{\delta_1}} \int \exp\left(-1/2(\boldsymbol{w}^* - \delta_1\boldsymbol{t})^T(\boldsymbol{w}^* - \delta_1\boldsymbol{t}) + \frac{1}{2\sigma^2_{\delta_1}}\delta_1^2\right) d\delta_1, \quad \boldsymbol{w}^* = \boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 \\
&= \frac{1}{\sigma_{\delta_1}} \exp(\frac{1}{2}\frac{((\boldsymbol{w}^*)^T\boldsymbol{t})^2}{\boldsymbol{t}^T\boldsymbol{t} + 1/\sigma^2_{\delta_1}}) \exp(-\frac{1}{2}(\boldsymbol{w}^*)^T\boldsymbol{w}^*)(\frac{1}{\boldsymbol{t}^T\boldsymbol{t} + 1/\sigma^2_{\delta_1}})^{1/2}
\end{aligned}$$

Therefore the posterior sample of η is given by, $\eta \sim \text{Ber}(Prob)$ where

$$prob = \frac{1}{1 + \frac{1-p^\eta}{p^\eta}\frac{1}{R}}, \quad R = \frac{1}{\sigma_{\delta_1}} \exp(\frac{1}{2}\frac{((\boldsymbol{w}^*)^T\boldsymbol{t})^2}{\boldsymbol{t}^T\boldsymbol{t} + 1/\sigma^2_{\delta_1}})(\frac{1}{\boldsymbol{t}^T\boldsymbol{t} + 1/\sigma^2_{\delta_1}})^{1/2}$$

Finally for the last piece of Dirac spike, the posterior sample of $p^\eta \mid \eta$ is given by

$$\begin{aligned}
p(p^\eta \mid \eta) &\propto \pi(p^\eta) p(\eta \mid p^\eta) \\
&\propto (p^\eta)^{a-1}(1 - p^\eta)^{b-1}(p^\eta)^\eta(1 - p^\eta)^{1-\eta}
\end{aligned}$$

Therefore, the posterior sample $p^\eta \mid \eta \sim \text{Beta}(a + \eta, b + 1 - \eta)$

Secondly, the full conditional posterior distribution of $\boldsymbol{\beta} \mid .$ is given by

$$\begin{aligned}
p(\boldsymbol{\beta} \mid .) &\propto p(\boldsymbol{\beta}) \prod_{i,j} N(w_{ij} \mid \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) \\
&\propto \exp(-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^T\boldsymbol{\beta}) \exp\left(-1/2(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})^T(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t})\right)
\end{aligned}$$

Therefore, posterior sample from

$$\boldsymbol{\beta} \mid . \sim MVN\left(\left(\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma^2_\beta I_p}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{w} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right), \left(\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma^2_\beta I_p}\right)^{-1}\right)$$

Next, the full conditional posterior distribution of $(\boldsymbol{\zeta}_1 \mid .)$ is given by

$$\begin{aligned} p(\boldsymbol{\zeta}_1 \mid .) &\propto p(\boldsymbol{\zeta}_1)\prod_{i,j} N(w_{ij} \mid \boldsymbol{X}^T_{ij}\boldsymbol{\beta} + \zeta_{i,1} + \delta_1 t_{ij}, 1) \\ &\propto \exp\left(-1/2\boldsymbol{\zeta}_1^T\boldsymbol{\Phi}_v^{-1}\boldsymbol{\zeta}_1\right)\exp\left(-1/2\left(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right)^T\left(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - C\boldsymbol{\zeta}_1 - \delta_1\boldsymbol{t}\right)\right) \end{aligned}$$

where $\boldsymbol{\Phi}_v = \text{diag}(r(v_1)\sigma^2_{\zeta_1}, \ldots, r(v_m)\sigma^2_{\zeta_1})$ be the variance-covariance matrix of prior distribution, where $r(0) = r$ and $r(1) = 1$. Therefore, posterior sample from

$$\boldsymbol{\zeta}_1 \mid . \sim MVN\left((C^TC + \boldsymbol{\Phi}_v^{-1})^{-1}C^T(\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta} - \delta_1\boldsymbol{t}), (C^TC + \boldsymbol{\Phi}_v^{-1})^{-1}\right)$$

The remain full conditional of parameter are as follows

$$\begin{aligned} v_i \mid . &\sim \text{Ber}\left(p^*_{v_i} = \frac{\pi_{slab}(\zeta_{i,1})p_{v_i}}{\pi_{slab}(\zeta_{i,1})p_{v_i} + \pi_{spike}(\zeta_{i,1})(1 - p_{v_i})}\right), \\ \sigma^2_\beta \mid . &\sim \text{InvGamma}\left(a_\beta + \frac{p}{2}, b_\beta + \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2}\right), \\ p_{v_i} \mid . &\sim \text{Beta}\left(a_v + v_i, b_v + (1 - v_i)\right), \\ \sigma^2_1 \mid . &\sim \text{InvGamma}\left(a_1 + \frac{m}{2}, b_1 + \frac{1}{2}\sum_{\forall i}\frac{\zeta^2_{i,1}}{r(v_i)}\right), \end{aligned}$$

E.2 Outcome modeling

After estimated propensity score, we consider it as latent variable in the outcome modeling using covariate adjustment regression strategy. The potential outcome model in terms of treatment

assignment and estimated propensity scores written as

$$
\begin{aligned}
&Y_{ij}(z) \mid ps_{ij}, U_i \sim N(\alpha z + \gamma ps_{ij} + \zeta_{i,2}, \sigma_y^2),\\
&v_{i,y} \mid p_{v_{i,y}} \sim \text{Ber}(p_{v_{i,y}}),\\
&p_{v_{i,y}} \sim \text{Beta}(a_{v_y}, b_{v_y}),\\
&\sigma_2^2 \sim \text{invgamma}(a_2, b_2),\\
&\sigma_y^2 \sim \text{invgamma}(a_y, b_y)
\end{aligned}
$$

Through MCMC Gibbs' sampling, the posterior samples of outcome model related parameters were generated from following conditional distributions, where $\mid .$ denote the conditional on all the other parameters.

$$
\begin{aligned}
&\alpha \mid . \sim N\left(\left(\boldsymbol{Z}^T\boldsymbol{Z} + \sigma_\alpha^{-2}\sigma_y^2\right)^{-1}\boldsymbol{Z}^T\left(\boldsymbol{Y}^{obs} - C\boldsymbol{\zeta}_2 - \boldsymbol{ps}\gamma\right), \left(\sigma_y^{-2}\boldsymbol{Z}^T\boldsymbol{Z} + \sigma_\alpha^{-2}\right)^{-1}\right)\\
&\gamma \mid . \sim N\left(\left(\boldsymbol{ps}^T\boldsymbol{ps} + \sigma_\gamma^{-2}\sigma_y^2\right)^{-1}\boldsymbol{ps}^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{Z}\alpha - C\boldsymbol{\zeta}_2\right), \left(\sigma_y^{-2}\boldsymbol{ps}^T\boldsymbol{ps} + \sigma_\gamma^{-2}\right)^{-1}\right),\\
&\boldsymbol{\zeta}_2 \mid . \sim N\left(\left(C^TC + \sigma_y^2\Phi_{v_y}^{-1}\right)^{-1}C^T\left(\boldsymbol{Y}^{obs} - \boldsymbol{Z}\alpha - \boldsymbol{ps}\gamma\right), \left(\sigma_y^{-2}C^TC + \boldsymbol{\Phi}_{v_y}^{-1}\right)^{-1}\right)\\
&v_{i,y} \mid . \sim \text{Ber}\left(p^*_{v_{i,y}} = \frac{\pi_{slab}(\zeta_{i,2})p_{v_{i,y}}}{\pi_{slab}(\zeta_{i,2})p_{v_{i,y}} + \pi_{spike}(\zeta_{i,2})(1-p_{v_{i,y}})}\right),\\
&\sigma_\alpha^2 \mid . \sim \text{invGamma}\left(a_\alpha + \frac{1}{2}, b_\alpha + \frac{\alpha^2}{2}\right),\\
&\sigma_\gamma^2 \mid . \sim \text{InvGamma}\left(a_\gamma + \frac{1}{2}, b_\gamma + \frac{\gamma^2}{2}\right),\\
&\sigma_y^2 \mid . \sim \text{invGamma}\left(a_y + \frac{N}{2}, b_y + \frac{(Y^{obs} - \boldsymbol{Z}\alpha - \boldsymbol{ps}\gamma - C\boldsymbol{\zeta}_2)^T(Y^{obs} - \boldsymbol{Z}\alpha - \boldsymbol{ps}\gamma - C\boldsymbol{\zeta}_2)}{2}\right),\\
&p_{v_{i,y}} \mid . \sim \text{Beta}\left(a_{v_y} + v_{i,y}, b_{v_y} + (1 - v_{i,y})\right),\\
&\sigma_2^2 \mid . \sim \text{invgamma}\left(a_2 + \frac{m}{2}, b_2 + \frac{1}{2}\sum_{\forall i}\frac{\zeta_{i,2}^2}{r(v_{i,y})}\right)
\end{aligned}
$$

where $\boldsymbol{\Phi}_{v_y} = \text{diag}(r(v_{1,y})\sigma_2^2, \ldots, r(v_{m,y})\sigma_2^2)$ be the variance-covariance matrix of prior distribution, where $r(0) = r$ and $r(1) = 1$.

Appendix F Simulation results

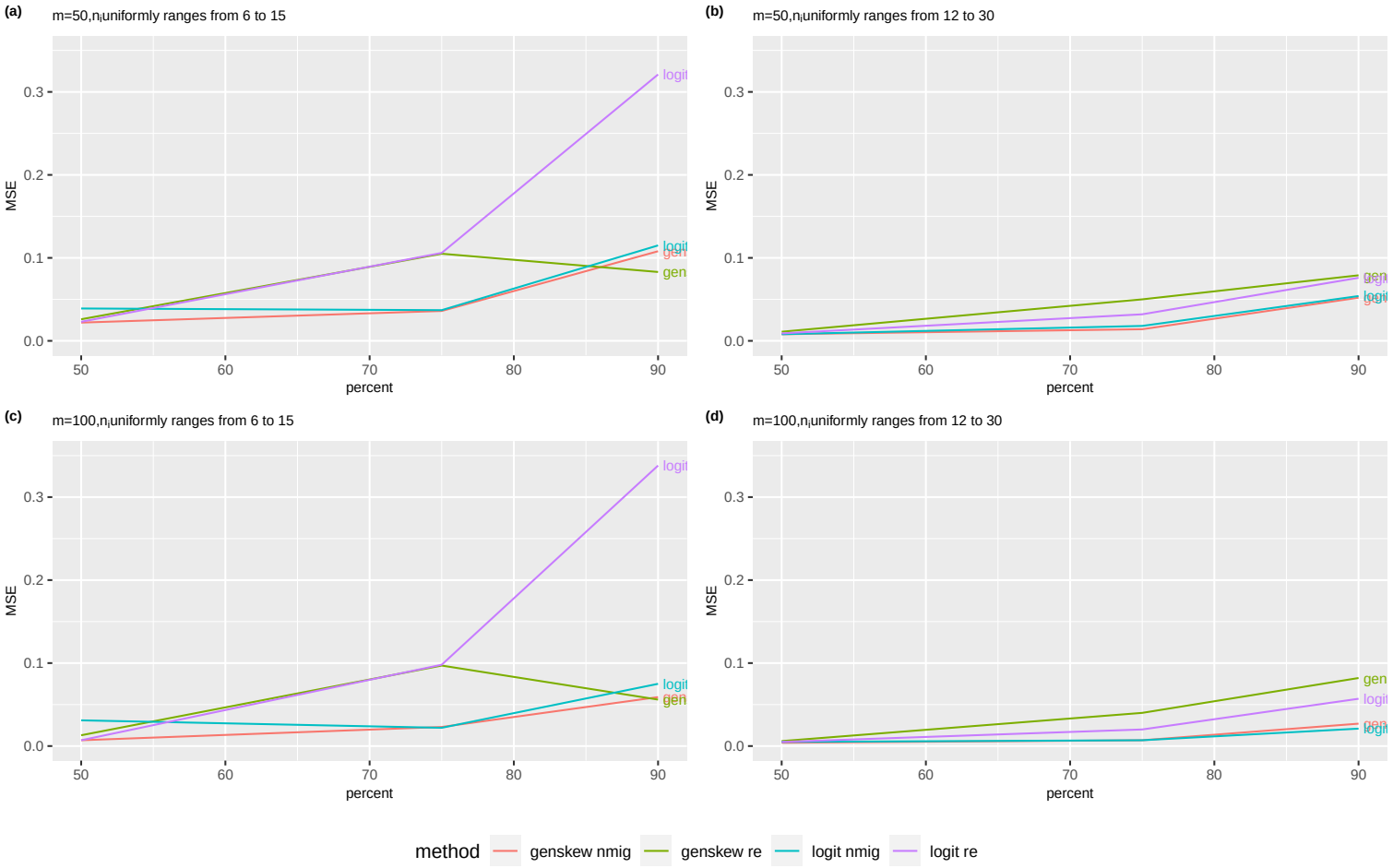


Figure 1: Plots summary under Cloglog

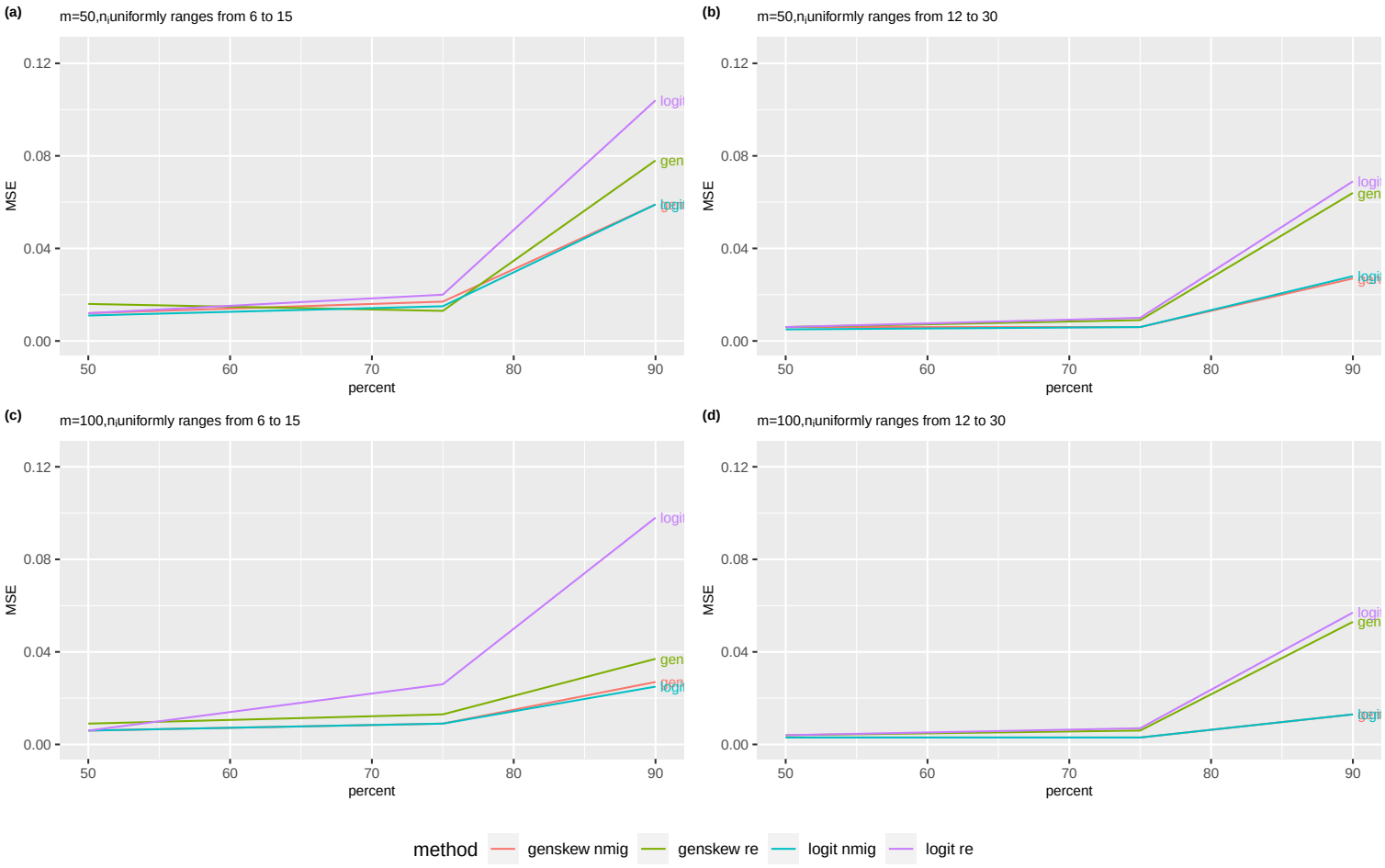


Figure 2: Plots summary under logit

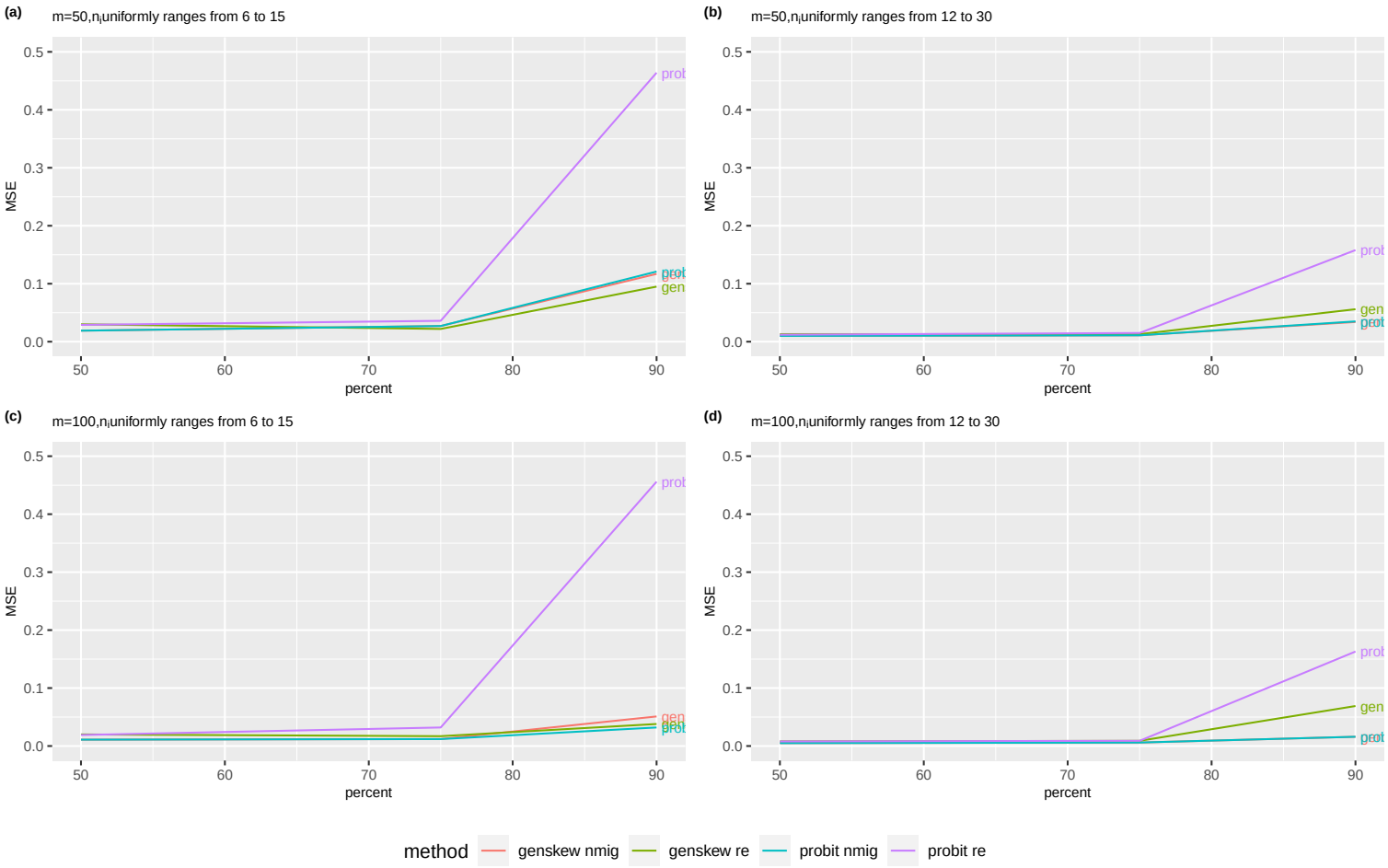


Figure 3: Plots summary under probit

Appendix G Skewed link model selection

$$\delta_1 \sim (1-\eta)\Delta_0(\delta_1) + \eta N(0, 100),$$

$$\eta \mid p^\eta \sim \text{Ber}(p^\eta)$$

$$p^\eta \sim \text{Beta}(a, b)$$

The posterior sample for η is needed to integrate out δ_1, based on the marginal distribution, the posterior sample of η is given by, $\eta \sim \text{Ber}(Prob)$ where

$$prob = \frac{1}{1 + \frac{1-p^\eta}{p^\eta}\frac{1}{R}}, \quad R = \frac{1}{10}\exp(\frac{1}{2}\frac{((\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{\beta} - C\boldsymbol{\zeta})^T\boldsymbol{t})^2}{\boldsymbol{t}^T\boldsymbol{t} + 1/100})(\frac{1}{\boldsymbol{t}^T\boldsymbol{t} + 1/100})^{1/2}$$

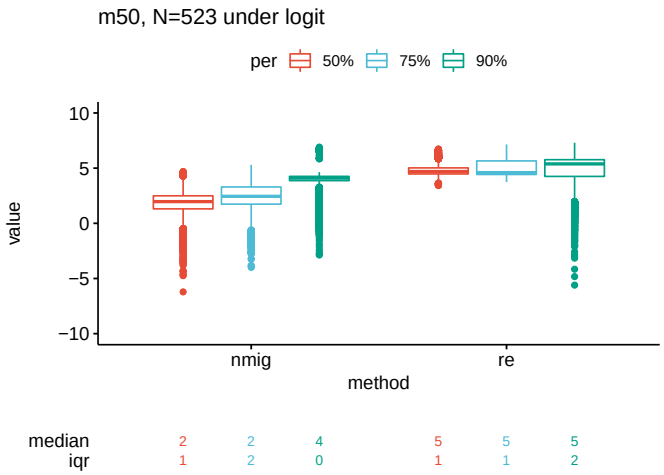
when using random effect for $\boldsymbol{\zeta}$, closed-to-zero clusters has negative values, while closed-to-zero clusters $\boldsymbol{\zeta}$ has really small values (nearly 0) using nmig, therefore

$$((\boldsymbol{w}_{\text{re}} - \boldsymbol{X}^T\boldsymbol{\beta}_{\text{re}} - \boldsymbol{C}\boldsymbol{\zeta}_{\text{re}})^T\boldsymbol{t}_{\text{re}})^2 > ((\boldsymbol{w}_{\text{nmig}} - \boldsymbol{X}^T\boldsymbol{\beta}_{\text{nmig}} - \boldsymbol{C}\boldsymbol{\zeta}_{\text{nmig}})^T\boldsymbol{t}_{\text{nmig}})^2 \xrightarrow{t>0,} \boldsymbol{R}_{\text{re}} > \boldsymbol{R}_{\text{nmig}}$$

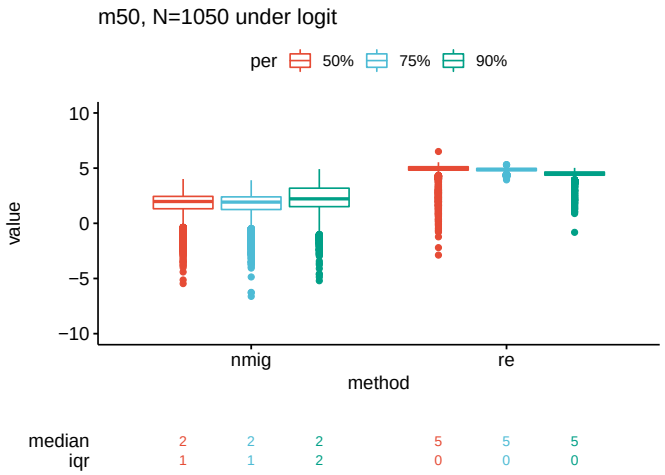
So *prob* is closer to 1, then η is more likely to be 1, then δ_1 is more likely to be nonzero.

Figure 4: Different scenarios under logit true model
value = $\mathbf{log_{10}((L - X^T\beta - C\zeta)^T\Omega t)^2}$

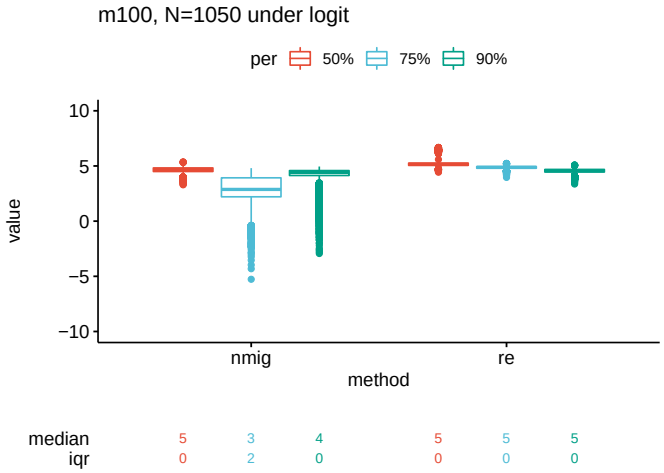
(a)



(b)



(c)



(d)

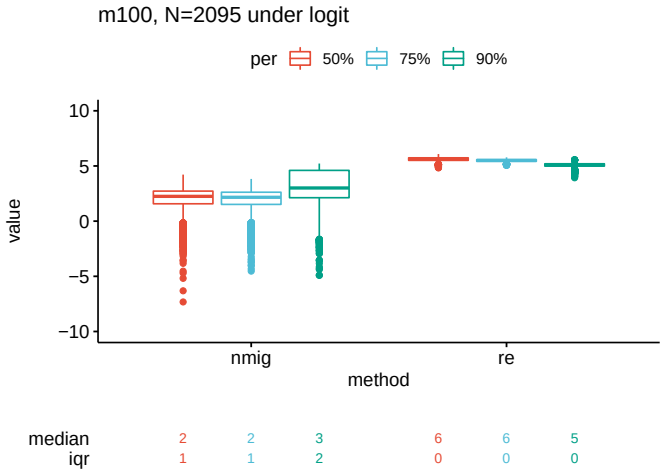
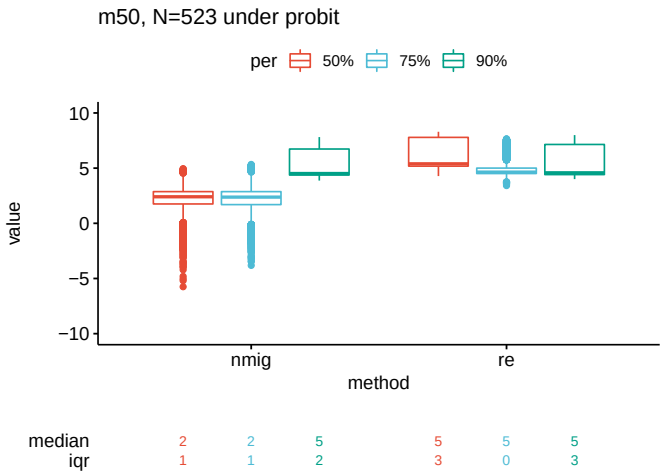
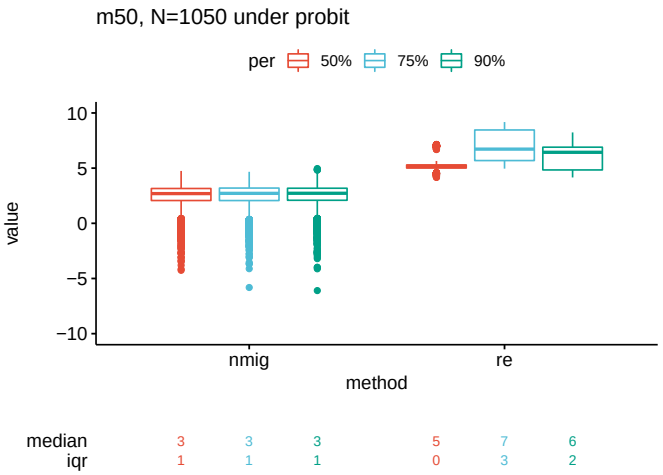


Figure 5: Different scenarios under probit true model
value = $\mathbf{log_{10}}((\boldsymbol{w} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} - \boldsymbol{C}\boldsymbol{\zeta})^{\boldsymbol{T}}\boldsymbol{t})^{\mathbf{2}}$

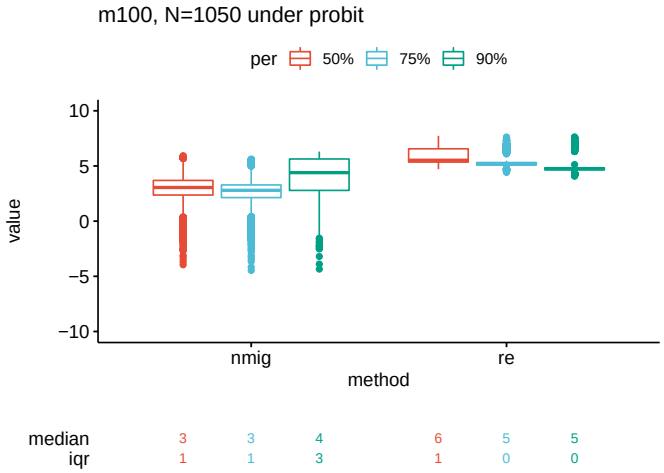
(a)



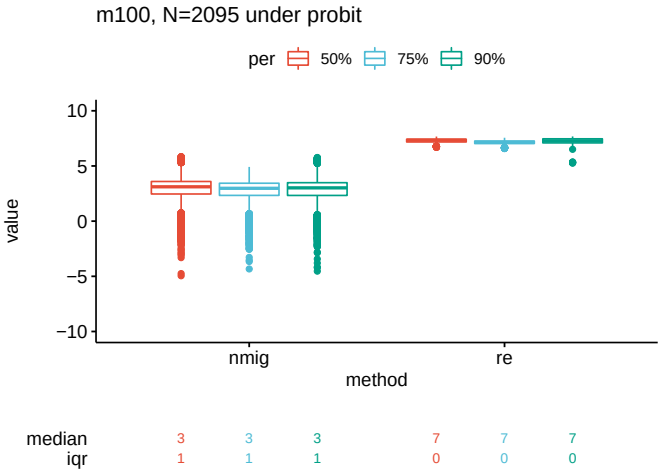
(b)



(c)



(d)



Bibliography

[1] Vilma Aho, Hanna M Ollila, Erkki Kronholm, Isabel Bondia-Pons, Pasi Soininen, Antti J Kangas, Mika Hilvo, Ilkka Seppälä, Johannes Kettunen, Mervi Oikonen, et al. Prolonged sleep restriction induces changes in pathways involved in cholesterol metabolism and inflammatory responses. Scientific reports, 6(1):1–14, 2016.

[2] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88(422):669–679, 1993.

[3] Weihua An. 4. bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. Sociological Methodology, 40(1):151–189, 2010.

[4] Cande V Ananth, Abraham Peedicayil, and David A Savitz. Effect of hypertensive diseases in pregnancy on birthweight, gestational duration, and small-for-gestational-age births. Epidemiology, pages 391–395, 1995.

[5] Bruno Arpino and Massimo Cannas. Propensity score matching with clustered data. an application to the estimation of the impact of caesarean section on the apgar score. Statistics in medicine, 35(12):2074–2091, 2016.

[6] Bruno Arpino and Fabrizia Mealli. The specification of the propensity score in multilevel observational studies. Computational Statistics & Data Analysis, 55(4):1770–1780, 2011.

[7] Peter C Austin. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and monte carlo simulations. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 51(1):171–184, 2009.

[8] Peter C Austin. The performance of different propensity score methods for estimating marginal hazard ratios. Statistics in medicine, 32(16):2837–2849, 2013.

[9] Peter C Austin, Nathaniel Jembere, and Maria Chiu. Propensity score matching and complex surveys. Statistical methods in medical research, 27(4):1240–1257, 2018.

[10] Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine, 34(28):3661–3679, 2015.

[11] Josiane Broussard and Matthew J Brady. The impact of sleep disturbances on adipocyte function and lipid metabolism. Best Practice & Research Clinical Endocrinology & Metabolism, 24(5):763–773, 2010.

[12] Ming-Hui Chen and Dipak K Dey. Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. Sankhyā: The Indian Journal of Statistics, Series A, pages 322–343, 1998.

[13] Ming-Hui Chen, Dipak K Dey, and Qi-Man Shao. A new skewed link model for dichotomous quantal response data. Journal of the American Statistical Association, 94(448):1172–1186, 1999.

[14] Francesco Cottone, Amelie Anota, Franck Bonnetain, Gary S Collins, and Fabio Efficace. Propensity score methods and regression adjustment for analysis of nonrandomized studies with health-related quality of life outcomes. Pharmacoepidemiology and drug safety, 28(5):690–699, 2019.

[15] John L Czajka, Sharon M Hirabayashi, Roderick JA Little, and Donald B Rubin. Projecting from advance data using propensity modeling: An application to income and tax statistics. Journal of Business & Economic Statistics, 10(2):117–131, 1992.

[16] Saddam Adams Damisa, Musa Tasi'u, Salamatu Yusuf Bello, Farouq Ndamadu Musa, Nurudeen Ayobami Ajadi, and Samson Agboola. On the comparison of some link functions of binary response analysis under symmetric and asymmetric assumptions. Biomedical Statistics and Informatics, 2(4):145, 2017.

[17] Peng Ding and Jiannan Lu. Principal stratification analysis using principal scores. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):757–777, 2017.

[18] Jeanette H Elam. An examination of single-gender and coeducational classes: Their impact on the academic achievement of middle school students enrolled in mathematics and science at selected schools in georgia. ProQuest LLC, 2009.

[19] Margaret M Ferrara. The student and the teacher—making a match in a single-gender classroom. Advances in Gender and Education,(1), 1421, 2009.

[20] National Center for Health Statistics. Nchs' vital statistics natality birth data. `http://data.nber.org/data/vital-statistics-natality-data.html`.

[21] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. Biometrics, 58(1):21–29, 2002.

[22] Constantine E Frangakis, Donald B Rubin, and Xiao-Hua Zhou. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. Biostatistics, 3(2):147–164, 2002.

[23] Jessica M Franklin, Wesley Eddings, Peter C Austin, Elizabeth A Stuart, and Sebastian Schneeweiss. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. Statistics in medicine, 36(12):1946–1963, 2017.

[24] Robert Gallop, Dylan S Small, Julia Y Lin, Michael R Elliott, Marshall Joffe, and Thomas R Ten Have. Mediation analysis with principal stratification. Statistics in medicine, 28(7):1108–1130, 2009.

[25] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. Journal of the American Statistical Association, 88(423):881–889, 1993.

[26] Matthew W Gillman, Sheryl Rifas-Shiman, Catherine S Berkey, Alison E Field, and Graham A Colditz. Maternal gestational diabetes, birth weight, and adolescent obesity. Pediatrics, 111(3):e221–e226, 2003.

[27] Anthony F Grasha. Teaching with style: A practical guide to enhancing learning by understanding teaching and learning styles. Alliance publishers, 1996.

[28] Zhulin He. Inverse conditional probability weighting with clustered data in causal inference. arXiv preprint arXiv:1808.01647, 2018.

[29] James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The review of economic studies, 64(4):605–654, 1997.

[30] J Hintze. Data matching–optimal and greedy. NCSS User's Guide, 2007.

[31] Paul W Holland. Statistics and causal inference. Journal of the American statistical Association, 81(396):945–960, 1986.

[32] David E Hunt. Matching models in education: The coordination of teaching methods with student characteristics. Ontario Institute for Studies in Education, Monograph, 1971.

[33] Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

[34] Hemant Ishwaran, J Sunil Rao, et al. Spike and slab variable selection: frequentist and bayesian strategies. Annals of Statistics, 33(2):730–773, 2005.

[35] Xun Jiang, Dipak K Dey, Rachel Prunier, Adam M Wilson, and Kent E Holsinger. A new class of flexible link functions with application to species co-occurrence in cape floristic region. The Annals of Applied Statistics, pages 2180–2204, 2013.

[36] Hui Jin and Donald B Rubin. Principal stratification for causal inference with extended partial compliance. Journal of the American Statistical Association, 103(481):101–111, 2008.

[37] Yoshitaka Kaneita, Makoto Uchiyama, Nobuo Yoshiike, and Takashi Ohida. Associations of usual sleep duration with serum lipid and lipoprotein levels. Sleep, 31(5):645–652, 2008.

[38] Bryan Keller, Jee-Seon Kim, and Peter M Steiner. Neural networks for propensity score estimation: Simulation results and recommendations. In Quantitative psychology research, pages 279–291. Springer, 2015.

[39] Gary King and Langche Zeng. Logistic regression in rare events data. Political analysis, 9(2):137–163, 2001.

[40] Alice P Kong, Yun-Kwok Wing, Kai C Choi, Albert M Li, Gary TC Ko, Ronald C Ma, Peter C Tong, Chung-Shun Ho, Michael H Chan, Margaret H Ng, et al. Associations of sleep duration with obesity and serum lipid profile in children and adolescents. Sleep medicine, 12(7):659–665, 2011.

[41] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. Statistics in medicine, 29(3):337–346, 2010.

[42] Soohyung Lee, Lesley J Turner, Seokjin Woo, and Kyunghee Kim. All or nothing? the impact of school and classroom gender composition on effort and academic achievement. Technical report, National Bureau of Economic Research, 2014.

[43] Clémence Leyrat, Shaun R Seaman, Ian R White, Ian Douglas, Liam Smeeth, Joseph Kim, Matthieu Resche-Rigon, James R Carpenter, and Elizabeth J Williamson. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Statistical methods in medical research, 28(1):3–19, 2019.

[44] Fan Li, Alan M Zaslavsky, and Mary Beth Landrum. Propensity score weighting with multilevel data. Statistics in medicine, 32(19):3373–3387, 2013.

[45] Fan Li, Alan M Zaslavsky, and Mary Beth Landrum. Propensity score weighting with multilevel data. Statistics in medicine, 32(19):3373–3387, 2013.

[46] Liang Li and Tom Greene. A weighting analogue to pair matching in propensity score analysis. The international journal of biostatistics, 9(2):215–234, 2013.

[47] Yanping Li, Sylvia H Ley, Deirdre K Tobias, Stephanie E Chiuve, Tyler J VanderWeele, Janet W Rich-Edwards, Gary C Curhan, Walter C Willett, JoAnn E Manson, Frank B Hu, et al. Birth weight and later life adherence to unhealthy lifestyles in predicting type 2 diabetes: prospective cohort study. BMJ, 351, 2015.

[48] Yun Li, Jeremy MG Taylor, and Michael R Elliott. A bayesian approach to surrogacy assessment using principal stratification in clinical trials. Biometrics, 66(2):523–531, 2010.

[49] Minlei Liao, Yunfeng Li, Farid Kianifard, Engels Obi, and Stephen Arcona. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. BMC nephrology, 17(1):1–14, 2016.

[50] Pu Lin, Kai-Ting Chang, Yan-An Lin, I-Shiang Tzeng, Hai-Hua Chuang, and Jau-Yuan Chen. Association between self-reported sleep duration and serum lipid profile in a middle-aged and elderly population in taiwan: a community-based, cross-sectional study. BMJ open, 7(10):e015964, 2017.

[51] Ariel Linden and Paul R Yarnold. Using classification tree analysis to generate propensity score weights. Journal of evaluation in clinical practice, 23(4):703–712, 2017.

[52] Chuanhai Liu. Robit regression: a simple robust alternative to logistic and probit regression. Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives, pages 227–238, 2004.

[53] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine, 23(19):2937–2960, 2004.

[54] Rashidul Alam Mahumud, Marufa Sultana, and Abdur Razzaque Sarker. Distribution and determinants of low birth weight in developing countries. Journal of Preventive Medicine and Public Health, 50(1):18, 2017.

[55] M Makgoba, MD Savvidou, and PJ Steer. The effect of maternal characteristics and gestational diabetes on birthweight. BJOG: An International Journal of Obstetrics & Gynaecology, 119(9):1091–1097, 2012.

[56] Gertraud Malsiner-Walli and Helga Wagner. Comparing spike and slab priors for bayesian variable selection. arXiv preprint arXiv:1812.07259, 2018.

[57] Alessandra Mattei, Fan Li, Fabrizia Mealli, et al. Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program. The Annals of Applied Statistics, 7(4):2336–2360, 2013.

[58] Lawrence C McCandless, Paul Gustafson, and Peter C Austin. Bayesian propensity score analysis for observational data. Statistics in medicine, 28(1):94–112, 2009.

[59] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. Journal of the American Statistical Association, 83(404):1023–1032, 1988.

[60] Majella O'Keeffe, Amy L Roberts, Michael Kelleman, Arindam RoyChoudhury, and Marie-Pierre St-Onge. No effects of short-term sleep restriction, in a controlled feeding setting, on lipid profiles in normal-weight adults. Journal of sleep research, 22(6):717–720, 2013.

[61] World Health Organization et al. Physical status: The use of and interpretation of anthropometry, Report of a WHO Expert Committee. World Health Organization, 1995.

[62] Wei Pan and Haiyan Bai. Propensity score analysis: Concepts and issues. Propensity score analysis: Fundamentals and developments, pages 3–19, 2015.

[63] Georgia Papadogeorgou, Christine Choirat, and Corwin M Zigler. Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. Biostatistics, 20(2):256–272, 2018.

[64] Trevor Park and George Casella. The bayesian lasso. Journal of the American Statistical Association, 103(482):681–686, 2008.

[65] Lesley H Parker and Léonie J Rennie. Teachers' implementation of gender-inclusive instructional strategies in single-sex and mixed-sex science classrooms. International Journal of Science Education, 24(9):881–897, 2002.

[66] Sepideh Pashami, Anders Holst, Juhee Bae, and Sławomir Nowaczyk. Causal discovery using clusters from observational data. In FAIM'18 Workshop on CausalML, Stockholm, Sweden, July 15, 2018, 2018.

[67] Judea Pearl. An introduction to causal inference. The international journal of biostatistics, 6(2), 2010.

[68] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. Journal of the American Statistical Association, 108(504):1339–1349, 2013.

[69] Rindang Bangun Prasetyo, Heri Kuswanto, Nur Iriawan, and Brodjol Sutijo Suprih Ulama. Binomial regression models with a flexible generalized logit link function. Symmetry, 12(2):221, 2020.

[70] Federico Ricciardi, Alessandra Mattei, and Fabrizia Mealli. Bayesian inference for sequential treatments under latent sequential ignorability. Journal of the American Statistical Association, pages 1–20, 2019.

[71] Janet W Rich-Edwards, Graham A Colditz, Meir J Stampfer, Walter C Willett, Matthew W Gillman, Charles H Hennekens, Frank E Speizer, and JoAnn E Manson. Birthweight and the risk for type 2 diabetes mellitus in adult women. Annals of Internal Medicine, 130(4_Part_1):278–284, 1999.

[72] James M Robins, Steven D Mark, and Whitney K Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics, pages 479–495, 1992.

[73] Paul R Rosenbaum. Model-based direct adjustment. Journal of the American Statistical Association, 82(398):387–394, 1987.

[74] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

[75] Evan Rosenman, Art B Owen, Michael Baiocchi, and Hailey Banack. Propensity score methods for merging observational and experimental datasets. arXiv preprint arXiv:1804.07863, 2018.

[76] Jason Roy, Joseph W Hogan, and Bess H Marcus. Principal stratification with predictors of compliance for randomized trials with 2 active treatments. Biostatistics, 9(2):277–289, 2008.

[77] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American Statistical Association, 75(371):591–593, 1980.

[78] Donald B Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. Journal of the American Statistical Association, 81(396):961–962, 1986.

[79] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. Annals of internal medicine, 127(8_Part_2):757–763, 1997.

[80] Donald B Rubin et al. Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. Statistical Science, 21(3):299–309, 2006.

[81] Fabian Scheipl. spikeslabgam: Bayesian variable selection, model choice and regularization for generalized additive mixed models in r. arXiv preprint arXiv:1105.5253, 2011.

[82] Scott L Schwartz, Fan Li, and Fabrizia Mealli. A bayesian semiparametric approach to intermediate variables in causal inference. Journal of the American Statistical Association, 106(496):1331–1344, 2011.

[83] Soko Setoguchi, Sebastian Schneeweiss, M Alan Brookhart, Robert J Glynn, and E Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and drug safety, 17(6):546–555, 2008.

[84] Atul Singhal, Jonathan Wells, Tim J Cole, Mary Fewtrell, and Alan Lucas. Programming of lean body mass: a link between birth weight, obesity, and cardiovascular disease? The American Journal of Clinical Nutrition, 77(3):726–730, 2003.

[85] Chris J Skinner. Inverse probability weighting for clustered nonresponse. Biometrika, 98(4):953–966, 2011.

[86] Robert E Slavin. Class size and student achievement: Small effects of small classes. Educational Psychologist, 24(1):99–110, 1989.

[87] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82(398):528–540, 1987.

[88] Misato Terada, Yoshio Matsuda, Masaki Ogawa, Hideo Matsui, and Shoji Satoh. Effects of maternal factors on birth weight in japan. Journal of Pregnancy, 2013, 2013.

[89] Felix J Thoemmes and Eun Sook Kim. A systematic review of propensity score methods in the social sciences. Multivariate behavioral research, 46(1):90–118, 2011.

[90] Felix J Thoemmes and Stephen G West. The use of propensity scores for nonrandomized designs with clustered data. Multivariate Behavioral Research, 46(3):514–543, 2011.

[91] Helga Wagner and Christine Duller. Bayesian model selection for logistic regression models with random intercept. Computational Statistics & Data Analysis, 56(5):1256–1274, 2012.

[92] Xia Wang, Dipak K Dey, et al. Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. The Annals of Applied Statistics, 4(4):2000–2023, 2010.

[93] Yu-Bo Wang, Cuilin Zhang, and Zhen Chen. Intergenerational associations between maternal diet and childhood adiposity: A bayesian regularized mediation analysis. Statistics in Biosciences, pages 1–19, 2021.

[94] Mi-Ja Woo, Jerome P Reiter, and Alan F Karr. Estimation of propensity scores using generalized additive models. Statistics in medicine, 27(19):3805–3816, 2008.

[95] Mollie E Wood, Stavroula Chrysanthopoulou, Hedvig ME Nordeng, and Kate L Lapane. The impact of nondifferential exposure misclassification on the performance of propensity scores for continuous and binary outcomes. Medical care, 56(8):e46–e53, 2018.

[96] Shu Yang. Propensity score weighting for causal inference with clustered data. Journal of Causal Inference, 6(2), 2018.

[97] Junni L Zhang and Donald B Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". Journal of Educational and Behavioral Statistics, 28(4):353–368, 2003.

[98] Junni L Zhang, Donald B Rubin, and Fabrizia Mealli. Likelihood-based analysis of causal effects of job-training programs using principal stratification. Journal of the American Statistical Association, 104(485):166–176, 2009.

[99] Peng Zhao, Xiaogang Su, Tingting Ge, and Juanjuan Fan. Propensity score and proximity matching using random forest. Contemporary clinical trials, 47:85–92, 2016.

[100] Qi Zhou, Yoo-Mi Chin, James D Stamey, and Joon Jin Song. Bayesian misclassification and propensity score methods for clustered observational studies. Journal of Applied Statistics, 45(9):1547–1560, 2018.

[101] Yeyi Zhu, Sjurdur F Olsen, Pauline Mendola, Thorhallur I Halldorsson, Shristi Rawal, Stefanie N Hinkle, Edwina H Yeung, Jorge E Chavarro, Louise G Grunnet, Charlotta Granström, et al. Maternal consumption of artificially sweetened beverages during pregnancy, and offspring growth through 7 years of age: a prospective cohort study. International Journal of Epidemiology, 46(5):1499–1508, 2017.

[102] Yeyi Zhu, Sjurdur F Olsen, Pauline Mendola, Thorhallur I Halldorsson, Edwina H Yeung, Charlotta Granström, Anne A Bjerregaard, Jing Wu, Shristi Rawal, Jorge E Chavarro, et al. Maternal dietary intakes of refined grains during pregnancy and growth through the first 7 y of life among children born to women with gestational diabetes. The American Journal of Clinical Nutrition, 106(1):96–104, 2017.

[103] Corwin M Zigler, Krista Watts, Robert W Yeh, Yun Wang, Brent A Coull, and Francesca Dominici. Model feedback in bayesian propensity score estimation. Biometrics, 69(1):263–273, 2013.

[104] Corwin Matthew Zigler and Francesca Dominici. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. Journal of the American Statistical Association, 109(505):95–107, 2014.