

©Copyright 2025

Daniel Suen

Topics in Estimation and Inference with Multivariate Missing Data

Daniel Suen

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Yen-Chi Chen, Chair

Thomas Richardson

Kwun Chuen Gary Chan

Program Authorized to Offer Degree:

Statistics

University of Washington

Abstract

Topics in Estimation and Inference with Multivariate Missing Data

Daniel Suen

Chair of the Supervisory Committee:

Yen-Chi Chen

Department of Statistics

This dissertation discusses statistical methodologies for complex missing data problems with a focus on multivariate and partially observed structures. Chapter 2 introduces a unified framework for handling multiple missing covariates and partially observed responses using inverse probability weighting, regression adjustment, and a multiply-robust procedure. Applications include the Cox model for survival analysis, missing responses, and binary treatment in causal inference, along with supporting identification and asymptotic theory. Chapter 3 focuses on modeling multivariate bounded discrete outcomes such as those from neuropsychological tests in dementia studies. We propose a flexible modeling strategy based on mixtures of experts and latent class models, extended to handle missing at random outcomes via a nested EM algorithm. The joint model also allows for imputation and clustering. Chapter 4 addresses nonmonotone missing data under missing not at random (MNAR) mechanisms, extending the work in Chapter 3. A tree graph is a directed acyclic graph on the missing patterns, and each one represents a MNAR mechanism. Combining this with the idea of a conjugate odds property, we are able to preserve distributional structure across missing patterns and construct relatively straightforward models for the full data distribution. Throughout the dissertation, we also highlight practical relevance using an Alzheimer's disease data set.

TABLE OF CONTENTS

	Page
List of Figures	iii
List of Tables	v
Chapter 1: Introduction	1
Chapter 2: Multistage Estimators for Missing Covariates and Incomplete Outcomes	4
2.1 Introduction	4
2.2 Background and Framework	6
2.3 A Multistage Estimator	11
2.4 Simulation	28
2.5 Data Analysis	32
2.6 Discussion	38
Appendix A: Chapter 2 Appendices	40
A.1 A Multistage Estimator for the Missing Response Problem	40
A.2 Asymptotic Theory	44
A.3 Beyond CCMV	51
A.4 Sensitivity Analysis	54
A.5 Details on Survival Analysis Simulation and additional simulation studies . .	58
A.6 Additional Analysis of the NACC Data	71
A.7 Proofs	75
Chapter 3: Modeling Missing at Random Neuropsychological Test Scores Using a Mixture of Binomial Product Experts	108
1 Introduction	108
2 Mixture of Binomial Product Experts	114
3 Missingness in the Outcome Variables	123

4	Inference	131
5	Cluster Analysis	133
6	Simulations	135
7	Application to the NACC data	139
8	Discussion	146
	Appendix B: Chapter 3 Appendices	148
B.1	Detailed Derivation of the EM Algorithm Update Equations	148
B.2	Identification Theory	152
B.3	Comments on Code Implementation	153
B.4	Remarks on Alternative EM Algorithms	154
B.5	Simulations	156
B.6	Comments on Model Selection	160
B.7	Additional Comments on Clustering	162
B.8	Proofs	163
B.9	Further Comments on the Real Data Analysis	170
	Chapter 4: Modeling Multivariate Missingness with Tree Graphs and Conjugate Odds	172
1	Introduction	172
2	Tree Graphs and Identification Theory	178
3	Conjugate Odds Families and Domain Adaptation	185
4	Tree Graphs and Conjugate Odds	196
5	Strategies for Tree Graph Selection	202
6	Inference	210
7	Sensitivity Analysis	215
8	Simulations	218
9	Real Data Analysis	221
10	Discussion	233
	Appendix C: Chapter 3 Appendices	235
C.1	Proofs	235
C.2	Further Conjugate Odds Examples	249
C.3	Additional Comments on Tree Graphs	252

LIST OF FIGURES

Figure Number	Page
2.1 Estimation of β in a simulated Cox model for varying sample sizes.	30
2.2 Estimation of the ATE in a simulated causal inference problem for varying sample sizes.	32
2.3 Point estimates and confidence intervals for each β parameter	34
2.4 Point estimates and confidence intervals for the IPW estimator	36
2.5 Point estimates and confidence intervals for the RA estimator	36
2.6 The tradeoff between model complexity, computation, and efficiency in the CCMV setting.	38
2.7 Sensitivity analysis for the Cox model simulation with both the IPW and RA estimators.	61
2.8 Sensitivity analysis for the ATE simulation with both the IPW and RA estimators.	71
2.9 Sensitivity analysis results for the IPW estimator for β_4	72
2.10 Sensitivity analysis results for the RA estimator for β_4	73
2.11 Sensitivity analysis results for the IPW estimator for β_5	73
2.12 Sensitivity analysis results for the RA estimator for β_5	74
3.1 Project 2 inference flowchart	135
3.2 CDR score distributions based on the missing patterns	141
3.3 Complete data clustering	145
3.4 Missing at random clustering	145
3.5 AIC and BIC curves for a simulated data set	162
3.6 Histogram of final log-likelihood values for the real data	171
4.1 CCMV and NCMV tree graph examples	183
4.2 Example of exponential tilting	187
4.3 LNCMV and RNCMV examples	206
4.4 Example of asymptotic covariance matrix block structure for the MLE	212

4.5	Tree graphs identified in simulation from data-driven methods	221
4.6	Missing pattern distribution for the UDSBENTD variable over 5 years. . . .	223
4.7	Tree graphs for UDSBENTD data analysis	224
4.8	Fitted model versus the complete case distribution	226
4.9	Fitted distributions against the observed data distributions	228
4.10	Fitted trajectories for patterns 10000, 11000, 10100, and 10010 with parent- based and mice methods	229
4.11	Fitted KDEs on wine data set	232
4.12	Tree graph examples	253
4.13	Representor graph examples	253
4.14	Merge operation example	255

LIST OF TABLES

Table Number	Page
3.1 Simulation results after $M = 20$ imputations	136
3.2 Simulation results after $M = 20$ imputations with <i>mice</i>	136
3.3 Computation time in minutes for different methods	137
3.4 Estimated test score means for the real data analysis	143
3.5 Estimated regression coefficients for the real data analysis	143
3.6 Simulation results after $M = 10$ imputations	158
3.7 Adjusted Rand index with $M = 20$ imputations	159
3.8 Simulation results after $M = 50$ imputations	159
3.9 Complete case analysis simulations	160
3.10 AIC and BIC simulation results	161
4.1 Simulation results	220
4.2 Simulation results for the data-driven methods	220

ACKNOWLEDGMENTS

A PhD is a long and winding journey, and I feel incredibly fortunate and deeply grateful to have met so many people who supported and guided me along the way.

First and foremost, I am grateful to my advisor Prof. Yen-Chi Chen for his guidance, advice, and mentorship over the years. I am indebted to him in many ways. In my first year, I took two courses from him, and his Statistical Machine Learning course, in particular, offered a wonderful guided tour through the landscape of modern statistics for a young student like myself. Our conversations are always thoughtful and erudite, and I have long admired his wisdom, honesty, openness, and optimism. He was also the one who introduced me to the Alzheimer’s disease dataset that became central to my research. Another set of thank yous goes out to the rest of my committee. Early during my graduate studies, Prof. Thomas Richardson served on my prelim committee and also helped connect me with a summer internship, where I had the opportunity to explore causal inference in an applied setting. Furthermore, Prof. Thomas Richardson and Prof. Gary Chan gave generous feedback and suggestions before and during my general exam for which I am extremely thankful. Prof. Jing Tao and Prof. Chun Wang have also been very kind to serve as GSRs on my committee.

Beyond my committee, several other faculty and staff at UW made a lasting impact. I look back on those days when I was still taking classes with much fondness. An acknowledgment goes to Prof. Marco Carone, whose Advanced Statistical Inference

III class was extremely illuminating, and I really enjoyed the introduction to empirical process and semiparametric efficiency theories. Thank you to Prof. Ema Perkovic, who not only taught an insightful special topics course on causal inference but has also been a generous academic advisor. I am also thankful to Prof. Daniela Witten and Prof. Amy Willis, who helped encourage me to apply for the NSF Graduate Research Fellowship and provided support along the way. I was so lucky to be awarded the fellowship, and I do not take it for granted. Thank you to Ellen Reynolds, Tracy Pham, Kristine Chan, and Veronica Bae for all the behind-the-scenes work that takes place in the department.

The road to graduate school started in undergrad. I am grateful to Prof. Sean Meyn, Prof. Tan Wong, and Prof. Alina Zare for their classes, mentorship, and exposing me to undergraduate research. I am also grateful to Dr. Michael Hurley and Dr. Edward Kao from MIT Lincoln Laboratory, who introduced me to missing data and encouraged me to pursue a PhD in statistics. I treasure those formative years very much.

I would be remiss not to acknowledge my wonderful friends and peers. Over the past six years, they have stood by me through the highs and lows with unwavering support, generosity, and care. I began graduate school in Autumn 2019, amid a tumultuous time in my personal life. Not long after, the world was upended by the Covid-19 pandemic, adding new and unprecedented challenges to an already demanding journey. And even as the world slowly emerged from the pandemic, life continued to throw its share of unexpected and difficult challenges. Through all of it, their kindness helped me endure, and their presence brought light in the moments I needed it most.

First, thank you to Yikun, Haoyang, Yanjiao, Steve, Jerry, and Zhen, who were great companions and have been wonderful to bounce ideas off of and get help with my research. Thank you in particular to Yikun for his always helpful discussions and for pointing me to many invaluable resources. To Alex (Ziyu), Antonio, Nina, David, Jess, and the many wonderful members of our statistics and biostatistics cohorts, your camaraderie has meant the world to me. The joys of graduate school and the beauty of learning could not have been felt without such dear friends. From exploring Seattle to attending conferences together, these shared experiences brought both laughter and inspiration. I will always cherish the memories we made together in Seattle, both inside and outside the classroom.

And my friends outside of the department, you have meant a lot to me too. Michael, Tianbo, Alvin, Susanne, and Abby, thank you for keeping me grounded and reminding me of the world beyond academia. Whether it was me visiting some of you in New York or us spending time together in Seattle, those experiences were always a welcome reminder of the friendships that have stayed steady through the years. I often reminisce on many of these times together. Midway through my PhD, I also discovered my love for ceramics, and I am grateful to my friends Ayò, Lara, Rosa, Alyssa, Zach, and Naomi that I was able to meet through the hobby. They helped me discover and nurture a new artistic side of myself outside of my STEM life. Their encouragement and companionship made that endeavor all the more meaningful.

I am eternally thankful to my Mom and Dad for their love, their sacrifices, and the values they instilled in me. I owe them so, so much. Their presence lives on in all that I strive to become. Thank you to my brother Johann for his support and unwavering belief in me. I hope to make my family proud, not just through what I

accomplish, but through the kind of person I continue to grow into.

The past six years have been marked by growth, challenges, and change. As I sat down to reflect, I often found myself searching for the right words and falling short. I only wish I could fully express the depth of gratitude I carry in my heart. I truly could not have done it with you all, and for that, I am forever touched.

DEDICATION

to the memory of my parents

Chapter 1

INTRODUCTION

Missing data is a fundamental issue in statistics and data science, arising in nearly every field of empirical research. Whether due to nonresponse in surveys, dropout in longitudinal studies, or incomplete electronic health records, missingness introduces uncertainty that can bias estimates, reduce efficiency, and complicate interpretation. As datasets become increasingly large with complex structures, the challenge of appropriately handling missing data has only grown more urgent. Addressing missing data is not merely a technical necessity; it is critical for drawing valid conclusions from incomplete information. Naively ignoring missingness or relying on ad hoc imputation methods can lead to misleading inferences, particularly when data are missing in a systematic or informative way. To this end, the development of principled, flexible, and computationally tractable methodologies remains a central goal in statistical research.

This dissertation investigates several interrelated problems in modern missing data methodology with a focus on multivariate settings that arise frequently in biomedical and social science applications. We develop and analyze new approaches tailored to the complexities of real-world data such as multiple missing covariates, nonmonotone patterns of missingness, bounded discrete outcomes, and missing at random and missing not at random mechanisms. Many common methods that practitioners use to handle missing data can be unsatisfactory for a number of reasons. For example, sometimes an imputation method is a naive deterministic single imputation or can even suffer from model uncongeniality problems (Meng, 1994). Rubin (1976) outlined three different kinds of missingness mechanisms: missing completely

at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This dissertation primarily focuses on the latter two.

In Chapter 2, we begin by proposing a general framework for estimation when covariates are partially missing and the outcome may itself be unobserved or incomplete. Drawing on techniques such as inverse probability weighting, regression adjustment, and multiply-robust estimation, we present a suite of multistage estimators. This framework is applied to three classical problems: Cox proportional hazards models in survival analysis, missing response, and binary treatment effects in causal inference. Each of these problems presents unique challenges when covariates are not fully observed. We derive identification theory as well as associated estimation procedures and asymptotic properties. We demonstrate performance through simulations and a real-world Alzheimer’s disease data set. This work builds on the complete-case missing value assumption that has been used in the literature (Little, 1993; Tchetgen Tchetgen et al., 2018), but the methodology can be generalized to other assumptions as well.

In Chapter 3, we consider modeling multivariate bounded discrete outcomes that can be subject to missingness. This problem is very general, but we are primarily motivated by neuropsychological test scores collected in dementia research. These outcomes are often collected alongside covariates and may be partially missing. We propose a model that flexibly captures the joint distribution conditional on covariates, motivated by the latent class model and the mixture of experts (Goodman, 1974; Jacobs et al., 1991). We extend this model to accommodate data missing at random using a nested EM algorithm, enabling both imputation and clustering.

Lastly, in Chapter 4, we address the challenging setting of nonmonotone and nonignorable (MNAR) missingness. Building on work from Chen (2022a), we introduce a class of missing not at random assumptions called tree graphs. Tree graphs through a directed acyclic graph

on missing patterns encode a nonparametric identifying restriction for the full data distribution. Within this framework, we consider a conjugate odds family of selection models that preserve the data distributional form across patterns. This property allows us to model the full data distribution elegantly and perform imputation. This work also builds off the work of Chapter 3 and [Suen and Chen \(2023\)](#) by demonstrating how models in those works can now be fit under a MNAR assumption.

Through theory, simulation studies, and real-data applications, we show that the proposed methods offer improvements over existing techniques and provide new methods for analyzing incomplete data. Collectively, the work in this dissertation advances missing data methodology, offering practical solutions for real data settings.

Chapter 2

MULTISTAGE ESTIMATORS FOR MISSING COVARIATES AND INCOMPLETE OUTCOMES

2.1 *Introduction*

Missing data occurs across many different fields, and as the size of datasets becomes ever larger, it becomes increasingly unrealistic to expect fully observed datasets. In many statistical problems, we have outcome variables of interest and a set of associated covariates. However, these outcome variables may not be completely observed. For example, in a survival problem, either the censoring time or the time-to-event are observed. In a missing response problem, a response may be subject to missingness and only be observed for a subset of the sample. In the binary treatment problem, only one of the two potential outcomes will be observed. In a traditional setting, all the covariates are assumed to be completely measured and there has been a vast amount of literature on how to estimate a parameter of interest. However, little is known when the covariates could be missing.

Some early attempts to handle missing covariates rely on ignorability assumptions, utilize estimation procedures in which the assumptions are not well-specified, or assume that there is only one missing pattern (one covariate subject to missing). For example, many standard assumptions to identify an effect of interest take the form of a conditional independence statement, where the conditioning is done on the covariates X . In the case of missing covariates, strong ignorability assumptions have been proposed by conditioning on the observed covariates X_R and the missing covariate pattern R and using generalized propensity scores with propensity score matching ([Rosenbaum and Rubin, 1984, 1985](#)). Other approaches in-

clude augmenting the support of the covariates with a new category for missing data $\{\text{NA}\}$ (Mayer et al., 2020, 2021). One downside to both of these methods is that in some situations, the interpretation of these missing data assumptions may not be so natural as there are effectively different sets of covariates being conditioned on for different individuals. Various imputation methods have also been proposed in the Cox model such as those in Paik (1997); White and Royston (2009); Beesley et al. (2016), but those proposed methods rely on specific assumptions that may not extend to more general settings. The multiple imputation approach discussed in Qu and Lipkovich (2009); Crowe et al. (2010) rely on missing data assumptions that are not clearly stated, which can lead to difficulty in interpretation. Berg and Yu (2021) also recently examined estimation under missing-not-at-random (MNAR) but they only consider a single covariate that are subject to missingness.

In the case of multiple missing variables, the missing-at-random (MAR) assumption becomes hard to interpret (Robins and Gill, 1997), so a nonignorable/MNAR missing data assumption may be preferred. There have been previous developments using nonignorable missing data assumptions in the literature. For example, Robins and Gill (1997) proposed the permutation missingness process as a generalization of MAR processes. Also, Tchetgen Tchetgen et al. (2018) used discrete-choice models as a way to generate a class of MNAR assumptions. Shpitser (2016); Sadinle and Reiter (2017); Malinsky et al. (2020) considered a “no self-censoring” or “itemwise conditionally independent nonresponse” assumption. ? introduced the idea of a pattern graph to generate further MNAR assumptions. Other approaches include various graphical ones: Mohan et al. (2013); Mohan and Pearl (2014); Nabi et al. (2020); Mohan and Pearl (2021).

In this chapter, we build on this previous work by introducing a general framework for constructing a set of compatible assumptions on problems with missing covariates and incomplete outcomes. This framework has several advantages. First, one can construct iden-

tifying restrictions on the missing covariates that are compatible with common identifying restrictions for the outcome variables. This ensures that the final outcome of interest is nonparametrically identifiable (Robins et al., 2000). Also, because we start with a clear set of assumptions and then develop an estimation method based on those assumptions, this framework provides a principled approach for estimating the final outcome of interest. Previous work on MNAR mechanisms can be utilized in our estimation procedure. Lastly, our framework can be applied to various settings including but not limited to: survival analysis with censoring, binary treatment from causal inference, and missing response.

Outline. In Section 2.2, we introduce our framework and several problems of interest. In Section 2.3, we outline how to construct a multistage estimator with the Cox model as a motivating example. In Sections 6 and 2.5, we explore simulation results and include a specific data analysis. As a matter of theoretical interest, we include details on a multistage estimator for the missing response problem in Appendix A.1. We discuss the asymptotic theory of our estimators and demonstrate the $\sqrt{n}$ -consistency of our estimators in Appendix A.2. Further simulations, data analysis results, and theoretical details such as sensitivity analyses are provided in Appendices A.3, A.4, A.5, A.6, and B.8.

2.2 Background and Framework

In this section, we discuss the problem setup and introduce notation that will be helpful in describing our framework. We also discuss the setting of three classical problems from the statistics literature: the Cox model from survival analysis, missing response, and binary treatment from causal inference.

2.2.1 Problem Setup

In our problems, there are both covariate and outcome-related variables. Let $Z = (X, R)$ represent the covariate-related variables, where $X \in \mathbb{R}^d$ is a set of covariate variables. Since X is multivariate, it can have an arbitrary missing pattern. We use the binary vector $R \in \{0, 1\}^d$ to represent the missing response for X . If $R_j = 1$, then the variable X_j is observed; otherwise, it is missing. Let $1_d = (1, 1, \dots, 1)^\top$ denote the case where X_j is observed for all j . We denote $\bar{r} = 1_d - r$ as the flipped version of pattern r . Thus, the notation $X_r = (X_j : r_j = 1)$ and $X_{\bar{r}} = (X_j : r_j = 0)$ refer to observed and missing covariate variables, respectively.

Next, we define $S = (W, A)$ as the outcome-related variables. Here $W \in \mathbb{R}^k$ is a set of outcome variables, and $A \in \mathcal{A}$ is a random variable over a discrete support $\mathcal{A}$ that represents the observed pattern of the outcome variables. Note that A has a more general interpretation than the missing pattern R for X because A does not necessarily correspond to a missing pattern for W in the traditional sense. This idea will be made more precise in the following subsections. Similar to the previous case with the covariates, we will use the notation W_A to denote the set of outcome variables that are observed under the observed pattern A .

Therefore, with a slight abuse of notation, the observed dataset consists of n IID observations of the form $\{(X_{i,R_i}, R_i, W_{i,A_i}, A_i)\}_{i=1}^n$. Here R_i is the covariate missingness pattern for the i th individual. As X_i consists of the covariates for the i th individual, some of its entries may be subject to missingness; thus, we use X_{i,R_i} to denote the observed covariates for the i th individual. Similarly, the observed outcome variables may differ between individuals, so we use W_{i,A_i} to denote the observed outcome variables for the i th individual, where A_i is the observed pattern of the outcome variables for the i th individual.

To further describe our problem formulation, we provide a brief description of each of the three problems here (Cox model, missing response, and binary treatment).

2.2.2 Cox model

In survival analysis, we typically wish to determine the effect of covariates on a time-to-event variable. However, the event time is often not known for some individuals due to censoring. For example, people participating in a medical trial may dropout prematurely, which may mean that the event of interest is not observed for those individuals. Thus, the observed data is of the form

$$(X_{1,R_1}, R_1, Y_1, \Delta_1), \dots, (X_{n,R_n}, R_n, Y_n, \Delta_n),$$

where $Y_i = \min\{T_i, C_i\} \in \mathbb{R}$ is the observed time, and $\Delta_i = 1(Y_i = T_i) = 1(C_i \geq T_i)$ is the status indicator for a time-to-event T_i and a censoring time C_i . Thus, the outcome variables are $W = (C, T)$ and the observed outcome pattern corresponds to $A = \Delta$. We see that $W_0 = C$ and $W_1 = T$, as the observed outcome pattern denotes whether or not the time-to-event variable was censored. One can see that the observed outcome pattern A is not associated with a missingness pattern. Importantly, the outcome variables $W = (C, T)$ are never completely observed for any individuals.

The Cox proportional hazards model assumes that the effect of the covariates is multiplicative on $\lambda(t|x)$ through a baseline hazard function $\lambda(t|x) = \lambda_0(t) \exp(\beta^\top x)$. Under the Cox model and an independent censoring assumption, the parameter of interest β^* satisfies the population estimating equation:

$$\begin{aligned} U(\beta) &= \mathbb{E} \left[\Delta \left(X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \right) \right] = 0, \\ s^{(1)}(t; \beta) &= \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X], \\ s^{(0)}(t; \beta) &= \mathbb{E}[1(Y \geq t) \exp(\beta^\top X)]. \end{aligned} \tag{2.1}$$

Our goal is to estimate β^* in this semiparametric setting. When the covariates X are fully observed, under suitable regularity conditions, it is known that finding a near-solution $\hat{\beta}$ to

the estimating equation

$$\begin{aligned}
U_n(\beta) &= \frac{1}{n} \sum_{i=1}^n \Delta_i \left(X_i - \frac{\hat{s}^{(1)}(Y_i; \beta)}{\hat{s}^{(0)}(Y_i; \beta)} \right) = o_p(1/\sqrt{n}), \\
\hat{s}^{(1)}(t; \beta) &= \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) \exp(\beta^\top X_i) X_i, \\
\hat{s}^{(0)}(t; \beta) &= \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) \exp(\beta^\top X_i)
\end{aligned}$$

provides a consistent and asymptotically normal estimator of β^* . Since X can now be subject to missingness, we need a way to estimate these nuisance functions.

2.2.3 Missing response

In the missing response problem, we assume that we have a single outcome variable of interest $Y \in \mathbb{R}$ that is possibly subject to missingness. We use the binary random variable $A \in \{0, 1\}$, which is 1 if and only if Y is observed. The observed data is either in the form of $(X_{i,R_i}, R_i, Y_i, A_i = 1)$ or $(X_{i,R_i}, R_i, A_i = 0)$. In the second case (Y_i is missing), we can think of the data as $(X_{R_i}, R_i, \mathbf{NA}, A_i = 0)$. Using our aforementioned notations, $W_0 = \mathbf{NA}$ and $W_1 = Y$. Since Y can be subject to missingness, A takes a traditional interpretation as a missing response pattern. Then, Y_{A_i} is the observed outcome variable for the i th individual. In a typical case, the parameter of interest is the marginal mean $\mathbb{E}[Y]$ and we often assume $Y \perp A|X$ to identify the marginal mean. Under this assumption, we have

$$\mathbb{E}[Y] = \mathbb{E} \left[\frac{YA}{\pi_A(X)} \right] \quad (2.2)$$

$$= \mathbb{E}[m_1(X)], \quad (2.3)$$

where $\pi_A(X) = P(A = 1|X)$ is the propensity score and $m_1(X) = \mathbb{E}[Y|A = 1, X]$ is the outcome regression. The expressions (2.2) and (2.3) are associated with the *inverse probability weighting* and *regression adjustment* estimators, respectively. When the covariates X are always observed, these quantities can be easily estimated from the observed data, but the situation becomes more complicated when the covariates are multivariate and subject to nonmonotone missingness.

2.2.4 Binary treatment

Under the Neyman-Rubin causal model (Rubin, 1978; Splawa-Neyman et al., 1990), we have a binary treatment $A \in \{0, 1\}$ that corresponds to potential outcomes $Y(0)$ and $Y(1)$, respectively. Here the outcome variables are $W = (Y(0), Y(1))$. Similar to the survival analysis problem with censoring, only one of the outcome variables is observed for a given individual. The observed outcome variables are $W_0 = Y(0)$ and $W_1 = Y(1)$.

A common parameter of interest is the average treatment effect (ATE) $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ and under the classical assumptions (consistency, SUTVA, weak ignorability; (Rubin, 1978)) the ATE admits the following expression

$$\begin{aligned} \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] &= \mathbb{E}\left[\frac{YA}{\pi_A(X)}\right] - \mathbb{E}\left[\frac{Y(1-A)}{1 - \pi_A(X)}\right] \\ &= \mathbb{E}[m_1(X)] - \mathbb{E}[m_0(X)], \end{aligned}$$

where $\pi_A(X) = P(A = 1|X)$ is the propensity score and $m_a(X) = \mathbb{E}[Y|A = a, X]$ is the outcome regression. Thus, this leads to the inverse probability weighting or regression adjustment estimators for inferring the ATE.

2.3 A Multistage Estimator

The intermediate quantities that we described in the previous section such as $s^{(1)}(t; \beta)$ and $\pi_A(X)$ are calculated using the covariates X and the outcome-related variables (W, A) . Thus, due to this shared property, before we construct our multistage estimator, we first consider a general real-valued function h of the covariate-related variables X and the outcome-related variables (W, A) , as defined below.

Definition 2.1 (Observed-decomposable functions). Let $h : \mathbb{R}^d \times \mathbb{R}^j \times \mathcal{A} \mapsto \mathbb{R}$ be a function of interest. Call h an observed-decomposable function (ODF) if

$$h(X, W, A) = \sum_{a \in \mathcal{A}} f_a(X, W_a) 1(A = a),$$

where $f_a : \mathbb{R}^d \times \mathbb{R}^{\dim(W_a)} \mapsto \mathbb{R}$ is a fixed (non-random) real-valued function for all $a \in \mathcal{A}$.

The interpretation of an ODF is that for each observed outcome pattern a , we have a separate real-valued function f_a that is dependent on all of the covariates X and the observed outcome variables W_a under pattern a . Since h has this decomposition, it is computable in the traditional setting with incomplete outcomes and completely observed covariates. As we will see in the construction of a multistage estimator, many quantities of interest can be written in this way, so this is a very natural decomposition. We will either typically be interested in estimating $\theta = \mathbb{E}[h(X, W, A)]$ or $\mathbb{E}[h(X, W, A)] = 0$ will be an estimating equation. We provide an illustrative example of h using the missing response problem.

Example 1 (Missing response). *In the traditional missing response problem without missing covariates where we commonly wish to estimate $\mathbb{E}[Y]$, there are 3 traditional estimators: regression adjustment on Y , inverse probability weighting on Y , and doubly robust on Y estimators. Because our goal is to estimate $\mathbb{E}[Y]$, the corresponding h , f_0 , and f_1 functions*

take the following forms in these three situations. Note that in the case $A = 0$ (Y is missing), we simplify notation such that $f_0(X, W_0) = f_0(X, \mathbf{NA}) = f_0(X)$ since W_0 can only take one possible outcome $\mathbf{NA}$.

1. *Regression adjustment on Y .* We have $h(X, W, A) = \mathbb{E}[Y|A = 1, X]$, which implies that $f_0(X) = f_1(X, Y) = \mathbb{E}[Y|A = 1, X]$.
2. *Inverse probability weighting on Y .* We have $h(X, W, A) = \frac{1(A=1)Y}{P(A=1|X)}$, which implies that $f_0(X) = 0$ and $f_1(X, Y) = \frac{Y}{P(A = 1|X)}$.
3. *Doubly-robust on Y .* We have $h(X, W, A) = \frac{1(A=1)Y}{P(A=1|X)} + \mathbb{E}[Y|A = 1, X] \left(1 - \frac{1(A=1)}{P(A=1|X)}\right)$, which implies that $f_0(X) = \mathbb{E}[Y|A = 1, X]$ and $f_1(X, Y) = \frac{Y - \mathbb{E}[Y|A = 1, X]}{P(A = 1|X)} + \mathbb{E}[Y|A = 1, X]$.

Example 2 (Cox model). In the Cox model, $\Delta = A$ and $W_0 = W_1 = Y$ and there are three key equations as described in equation (2.1): $U(\beta)$, $s^{(1)}(t; \beta)$, $s^{(0)}(t; \beta)$.

1. $U(\beta)$. For this quantity, $U(\beta) = \mathbb{E}[h(X, W, A)]$, so $f_0(X, W_0) = 0$ and $f_1(X, W_1) = \left(X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)}\right)$.
2. $s^{(1)}(t; \beta)$. In this case, by setting $s^{(1)}(t; \beta) = \mathbb{E}[h(X, W, A)]$, we obtain $f_0(X, W_0) = 1(Y \geq t) \exp(\beta^T X) X$ and $f_1(X, W_1) = 1(Y \geq t) \exp(\beta^T X) X$.
3. $s^{(0)}(t; \beta)$. This case is similar to $s^{(1)}(t; \beta)$, so that $f_0(X, W_0) = 1(Y \geq t) \exp(\beta^T X)$ and $f_1(X, W_1) = 1(Y \geq t) \exp(\beta^T X)$.

2.3.1 The IPW approach for missing covariates

To handle missing covariates, we start with an IPW procedure based on using the *complete odds* as defined below.

Definition 2.2 (Complete odds). For all missingness patterns r and observed outcome pat-

terns a , the complete odds is

$$O_{r,a}(X, W_a) = \frac{P(R = r|X, W_a, A = a)}{P(R = 1_d|X, W_a, A = a)}.$$

Thus, these complete odds $O_{r,a}(X, W_a)$ are a function of all of the outcome variables completely observed under outcome pattern a and the covariates. With the complete odds, $\mathbb{E}[f_a(X, W_a)1(A = a)]$ can be decomposed via the following lemma.

Lemma 2.1 (Complete odds decomposition). *The complete odds can be used to decompose the following two expectations as follows*

$$\mathbb{E}[f_a(X, W_a)1(A = a)1(R = r)] = \mathbb{E}[f_a(X, W_a)1(A = a)O_{r,a}(X, W_a)1(R = 1_d)],$$

$$\mathbb{E}[f_a(X, W_a)1(A = a)] = \sum_r \mathbb{E}[f_a(X, W_a)1(A = a)O_{r,a}(X, W_a)1(R = 1_d)]. \quad (2.4)$$

This lemma shows that the expectation can be estimated using the complete cases since we have the indicator $1(R = 1_d)$ in the expression on the RHS. Thus, we see that $\mathbb{E}[f_a(X, W_a)1(A = a)]$ is identifiable, provided that the complete odds $O_{r,a}(X, W_a)$ are identifiable for all r . So, it is clear what we need to show before estimation. Additionally, note that

$$\sum_r O_{r,a}(X, W_a) = \sum_r \frac{P(R = r|X, W_a, A = a)}{P(R = 1_d|X, W_a, A = a)} = \frac{1}{P(R = 1_d|X, W_a, A = a)} \quad (2.5)$$

is the inverse propensity score, so we see that (2.4) recovers the form of the IPW estimator. In light of this, one can view the complete odds as an alternate version of a propensity score. In this situation, it is not required to use the complete odds approach since one may use the

propensity score directly. However, modeling via complete odds has several nice properties; we discuss this in Remark 1.

In the following proposition, we show that a set of identifiable complete odds is equivalent to a set of identifiable selection probabilities. This connection offers one key insight: modeling the covariate missingness mechanism through the complete odds is equivalent to modeling the covariate missingness mechanism using the selection model (Little and Rubin, 2002). Specific missing data assumptions may make the former approach easier than the latter, so this can be a useful result.

Proposition 2.1 (Equivalence of complete odds and selection probabilities). *Fix an observed outcome pattern a . The complete odds $O_{r,a}(X, W_a)$ are identifiable for all r if and only if the selection probabilities $P(R = r|X, W_a, a)$ are identifiable for all r .*

Furthermore, another consequence of this proposition is that a set of identifiable complete odds implies that the distribution $p(x, r, w_a, a)$ is identifiable for all a . Marginalizing out r implies that the distribution $p(x, w_a, a)$ is nonparametrically identified (Robins et al., 2000), which is sufficient for showing the effect of interest is identifiable under the traditional assumptions that assume fully observed covariates. This implies that with a set of identifiable complete odds, we can use existing problem-specific assumptions to identify the parameter of interest.

Remark 1 (Advantages to modeling the complete odds). *A modeling approach using the complete odds can be preferable to one using the selection probabilities directly. Suppose r and a are fixed. Then, $O_{r,a}(X, W_a)$ is variation independent of any $O_{r',a}(X, W_a)$ for $r' \neq r$. Indeed, $O_{r,a}(X, W_a)$ can take any value in $[0, \infty)$ and is not constrained by any other odds function $O_{r',a}(X, W_a)$. On the other hand, the selection probabilities $P(R = r|x, w_a, a)$ have an explicit sum to 1 constraint, so the selection probabilities for all missing patterns r need to be modeled at the same time. In principle, one can use different methods to model each*

odds $O_{r,a}(X, W_a)$, provided it is identifiable under a missing data assumption for X . These allows for some flexibility because one can choose to have different binary classifiers for each pattern r . Due to the variation independence, they can all be compatible with each other.

Because of Proposition 2.1, we only need to identify the complete odds. Here we consider a (conditional) complete case missing value (CCMV) assumption (Little, 1993; Tchetgen Tchetgen et al., 2018).

Definition 2.3 (CCMV missing data assumption). Using a pattern mixture model factorization, the complete case missing value (CCMV) assumption equates the following two distributions

$$p(x_{\bar{r}}|x_r, r, w_a, a) \stackrel{\text{CCMV}}{=} p(x_{\bar{r}}|x_r, R = 1_d, w_a, a) \quad (2.6)$$

for all covariate missingness patterns r and observed outcome patterns a .

Note that the above condition is a conditional CCMV in the sense that we are using the CCMV for the missing X conditioned on the observed outcomes. We call it CCMV for abbreviation. Under Proposition 2.2, this particular missing data assumption implies a simple closed form expression $Q_{r,a}(X_r, W_a)$ for the odds.

Proposition 2.2 (Odds under CCMV). *Under the CCMV assumption for the covariates X , the odds $O_{r,a}(X, W_a)$ has the following simplified form in terms of observable quantities X_r and W_a , namely*

$$O_{r,a}(X, W_a) := \frac{P(R = r|X, W_a, a)}{P(R = 1_d|X, W_a, a)} \stackrel{\text{CCMV}}{=} \frac{P(R = r|X_r, W_a, a)}{P(R = 1_d|X_r, W_a, a)} =: Q_{r,a}(X_r, W_a). \quad (2.7)$$

Because each $Q_{r,a}(X_r, W_a)$ is identifiable from the data, Proposition 2.2 shows that the CCMV assumption allows us to estimate the parameter of interest under missing covariates. The modeling $Q_{r,a}(X_r, W_a)$ is easy because it is comparing the proportions between two missing patterns.

We provide some comments on generalizing the CCMV assumption to other assumptions in Appendix A.3. Other missing data assumptions can yield other identifiable and estimable expressions for the odds, but one would simply obtain a different equality than the one in (2.7). For example, ? exploits a pattern graph identifying restriction to present a recursive form for computing the odds.

Remark 2 (Relation to density ratio estimation). Wang and Kim (2021) have examined the problem of estimating an inverse propensity score using a density ratio representation under a missing at random assumption. Interestingly, because they assume the variables can be partitioned into one set of always missing variables and one set of always observed variables, the missing at random assumption is equivalent to CCMV. Additionally, our complete odds can be expressed as a density ratio as well, so these problems are related. We elect to model our complete odds directly rather than reexpress it as a ratio of densities, but note that one can borrow methods from the density ratio literature to also tackle this problem. See Sugiyama et al. (2012) for various approaches.

Theorem 2.1 (Nonparametric identification of CCMV). *The CCMV assumption in Definition 2.3 nonparametrically identifies the full-covariate distribution $p(x, r, w_a, a)$.*

With Theorem 2.1, we are able to identify the full-covariate distribution $p(x, r, w_a, a)$. This implies that the problem reduces to scenarios where the covariates are observed. Thus, we can estimate the parameter of interest as if all covariates are observed. In what follows, we give an example of Cox model, and in Appendix A.1, we provide examples of missing response problems.

Example: Cox model and survival problems

Recall that in the survival analysis problem, $\Delta = A$ and $W_0 = W_1 = Y$. In the case of $s^{(1)}(t; \beta)$, we have

$$f_0(X, W_0) = f_1(X, W_1) = 1(Y \geq t) \exp(\beta^\top X) X. \quad (2.8)$$

For any fixed time point t , a direct application of Lemma 2.1 yields

$$\begin{aligned} s^{(1)}(t; \beta) &= \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X] \\ &= \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X 1(\Delta = 1)] + \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X 1(\Delta = 0)] \\ &= \sum_r \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X 1(\Delta = 1) O_{r,1}(X, Y) 1(R = 1_d)] \\ &\quad + \sum_r \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X 1(\Delta = 0) O_{r,0}(X, Y) 1(R = 1_d)] \\ &= \sum_{r, \delta} \mathbb{E}[1(Y \geq t) X e^{\beta^\top X} O_{r,\delta}(X, Y) 1(\Delta = \delta) 1(R = 1_d)], \end{aligned}$$

where $O_{r,\delta}(X, Y) = P(R = r|X, Y, \delta)/P(R = 1_d|X, Y, \delta)$ are the complete odds. A similar expression can be obtained for $s^{(0)}(t; \beta)$. Next, we consider the estimating equation, which rewrites as

$$\begin{aligned} U(\beta) &= \mathbb{E} \left[\Delta \left(X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \right) \right] \\ &= \sum_r \mathbb{E} \left[\Delta \left(X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \right) O_{r,1}(X, Y) 1(R = 1_d) \right] \end{aligned} \quad (2.9)$$

using Lemma 2.1.

With Proposition 2.2, $O_{r,\delta}(X, Y) = Q_{r,\delta}(X_r, Y)$, which provides an elegant way of computing the odds because the quantity $Q_{r,\delta}(X_r, Y)$ is a ratio of two conditional probabilities

that no longer depend on all of the covariates but rather observed covariates X_r . We now discuss the construction of a multistage estimator under CCMV using the complete odds. Henceforth, we will use $Q_{r,\delta}(X_r, Y)$ rather than $O_{r,\delta}(X, Y)$ because they are equal by Proposition 2.2. This choice of an identifying restriction allows for us to compute the inverse propensity score in a fairly straightforward manner.

There are three steps to our multistage estimator, which we outline below.

1. **STAGE 1 (MODELING THE MISSINGNESS OF X USING THE COMPLETE ODDS).** We start by modeling $Q_{r,\delta}(X_r, Y)$ for every r and δ . For example, we can impose a parametric model $Q_{r,\delta}(X_r, Y; \alpha_{r,\delta})$ such as a logistic regression. Because this essentially becomes a binary classification problem (comparing the classes $R = r$ and $R = 1_d$ based on the observed data X_r and Y), one may also appeal to various machine learning methods such as decision trees and support vector machines. Other parametric binary classification tools are also available.
2. **STAGE 2 (ESTIMATING INTERMEDIATE QUANTITIES).** Next, we estimate the intermediate quantities of interest that are dependent on the potentially missing covariates X . Because of the model described in Stage 1, this estimation becomes possible. For example, assuming that the model for $Q_{r,\delta}(X_r, Y; \alpha_{r,\delta})$ is correctly specified and $\hat{\alpha}_{r,\delta}$ is a consistent estimator of $\alpha_{r,\delta}^*$, under the suitable regularity conditions, the following is a consistent estimator for $s^{(1)}(t; \beta)$

$$\begin{aligned}
\hat{s}^{(1)}(t; \beta) &= \sum_{r,\delta} \left(\frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) X_i e^{\beta^\top X_i} Q_{r,\delta}(X_{i,r}, Y_i; \hat{\alpha}_{r,\delta}) 1(\Delta_i = \delta) 1(R_i = 1_d) \right) \\
&= \frac{1}{n} \sum_{i=1}^n \left(1(Y_i \geq t) X_i e^{\beta^\top X_i} \cdot \underbrace{\left[\sum_{r,\delta} Q_{r,\delta}(X_{i,r}, Y_i; \hat{\alpha}_{r,\delta}) 1(\Delta_i = \delta) 1(R_i = 1_d) \right]}_{\lambda_i} \right)
\end{aligned} \tag{2.10}$$

$$= \frac{1}{n} \sum_{i=1}^n \lambda_i \cdot 1(Y_i \geq t) X_i e^{\beta^\top X_i},$$

where λ_i is a weight for the i th individual, as defined in (2.10). Similarly, we can mimic the same construction to estimate $s^{(0)}(t; \beta)$, yielding

$$\widehat{s}^{(0)}(t; \beta) = \frac{1}{n} \sum_{i=1}^n \lambda_i \cdot 1(Y_i \geq t) e^{\beta^\top X_i}.$$

3. **STAGE 3 (CONSTRUCTING THE FINAL ESTIMATOR).** Lastly, we assemble the intermediate quantities of interest together to create the final estimator. Following the decomposition in (2.9), we have

$$\begin{aligned} U_n(\beta) &= \sum_{r,\delta} \left(\frac{1}{n} \sum_{i=1}^n \Delta_i \left(X_i - \frac{\widehat{s}^{(1)}(Y_i; \beta)}{\widehat{s}^{(0)}(Y_i; \beta)} \right) Q_{r,\delta}(X_{i,r}, Y_i; \widehat{\alpha}_{r,1}) 1(R_i = 1_d) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\Delta_i \left(X_i - \frac{\widehat{s}^{(1)}(Y_i; \beta)}{\widehat{s}^{(0)}(Y_i; \beta)} \right) \cdot \underbrace{\sum_r Q_{r,1}(X_{i,r}, Y_i; \widehat{\alpha}_{r,1}) 1(\Delta_i = 1) 1(R_i = 1_d)}_{\lambda_{i,1}} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \lambda_i \cdot \Delta_i \left(X_i - \frac{\widehat{s}^{(1)}(Y_i; \beta)}{\widehat{s}^{(0)}(Y_i; \beta)} \right), \end{aligned}$$

where the second to last equality follows because $\Delta_i \in \{0, 1\}$. The last equality follows because $\lambda_i \Delta_i = \lambda_{i,1} \Delta_i$. Thus, $U_n(\beta) = 0$ is a reweighted estimating equation and can be solved to provide a consistent estimator $\widehat{\beta}$ for β^* .

A nice property is that the weights λ_i and $\lambda_{i,1}$ need only to be computed from the data once and can be reused thereafter. We summarize the implementation of this reweighted Cox model in Algorithm 1.

Algorithm 1 Fitting a weighted Cox model under CCMV for X

1. Input: $\{(X_{i,R_i}, R_i, Y_i, \Delta_i)\}_{i=1}^n$
2. For each pair (r, δ) , fit the parametric model $Q_{r,\delta}(X_r, Y; \alpha_{r,\delta})$.
3. For $i = 1, \dots, n$, do
 - 3-1. Form the weights

$$\lambda_i = \sum_{\delta} \sum_r Q_{r,\delta}(X_{i,r}, Y_i; \hat{\alpha}_{r,\delta}) 1(\Delta_i = \delta) 1(R_i = 1_d).$$

4. Estimate the intermediate quantities

$$\hat{s}^{(0)}(t; \beta) = \frac{1}{n} \sum_{i=1}^n \lambda_i \cdot 1(Y_i \geq t) e^{\beta^\top X_i},$$

$$\hat{s}^{(1)}(t; \beta) = \frac{1}{n} \sum_{i=1}^n \lambda_i \cdot 1(Y_i \geq t) X_i e^{\beta^\top X_i}.$$

5. Find $\hat{\beta}$ by finding a near-solution to the finite sample estimating equation

$$U_n(\beta) = \frac{1}{n} \sum_{i=1}^n \lambda_i \cdot \Delta_i \left(X_i - \frac{\hat{s}^{(1)}(Y_i; \beta)}{\hat{s}^{(0)}(Y_i; \beta)} \right) = 0.$$

6. Return: $\hat{\beta}$.
-

2.3.2 Regression adjustment and imputation for missing covariates

In addition to the IPW approach for handling missing covariates, we can use a regression adjustment approach as well. Moreover, this regression adjustment naturally leads to an imputation procedure. The following lemma is a key step to the regression adjustment (RA).

Lemma 2.2 (RA decomposition). *Regression adjustment can be used to decompose the following two expectations*

$$\mathbb{E}[f_a(X, W_a) 1(A = a) 1(R = r)] = \mathbb{E}[m_{r,a}(X_r, W_a) 1(R = r) 1(A = a)],$$

$$\mathbb{E}[f_a(X, W_a)1(A = a)] = \sum_r \mathbb{E}[m_{r,a}(X_r, W_a)1(R = r)1(A = a)],$$

where $m_{r,a}(X_r, W_a) = \mathbb{E}[f_a(X, W_a)|X_r, R = r, W_a, A = a]$ is a regression function.

Thus, in order to estimate the expectation, it suffices to model the regression function $m_{r,a}(X_r, W_a)$, which is a conditional expectation.

In practice, placing a parametric model for each regression function via $m_{r,a}(x_r, w_a; \beta_{r,a})$ may be problematic because naively fitting models can lead to incompatible regression functions and uncongenial sources of input (Meng, 1994). Instead, we recommend placing models on the extrapolation densities $p(x_{\bar{r}}|x_r, r, a, w_a)$ for all r . This ensures that the modeled regression functions will be compatible with each other.

Under the CCMV assumption (equation 2.6), we have

$$p(x_{\bar{r}}|x_r, r, w_a, a) = p(x_{\bar{r}}|x_r, R = 1_d, w_a, a),$$

which further implies

$$m_{r,a}(X_r, W_a) = \mathbb{E}[f_a(X, W_a)|X_r, R = r, W_a, A = a] \stackrel{\text{CCMV}}{=} \mathbb{E}[f_a(X, W_a)|X_r, R = 1_d, W_a, A = a].$$

The RHS is identifiable using observations with complete covariates. In this case, there are two ways of estimating the regression function:

1. **Method 1: Closed-form computation.** In some good scenarios, we may be able to derive the closed-form of the expectation $m_{r,a}(X_r, W_a) = \mathbb{E}[f_a(X, W_a)|X_r, R = r, W_a, A = a]$ from a parametric model of the extrapolation densities $p(x_{\bar{r}}|x_r, r, a, w_a)$. For instance, a multivariate Gaussian model implies that the conditional distributions will be conditional Gaussians.

2. Method 2: Imputation (Monte Carlo approximation). Alternatively, we can approximate the conditional expectation by a Monte Carlo method. This is equivalent to the frequentist multiple imputation. We generate from the extrapolation density $p(x_{\bar{r}}|x_r, r, a, w_a)$ and then evaluate the conditional expectation $\mathbb{E}[f_a(X, W_a)|X_r, R = r, W_a, A = a]$. Under CCMV (2.6), the extrapolation density $p(x_{\bar{r}}|x_r, r, y, \delta)$ is equated to the identifiable distribution $p(x_{\bar{r}}|x_r, R = 1_d, y, \delta)$. The latter distribution can be estimated using the complete cases.

Example: Cox model and survival problems

Using Lemma 2.2 with the functions defined in (2.8), we have

$$\begin{aligned} s^{(1)}(t; \beta) &= \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X] \\ &= \sum_{r, \delta} \mathbb{E}[m_{r, \delta}^{(1)}(X_r, Y; t, \beta) 1(R = r) 1(\Delta = \delta)]. \end{aligned}$$

Thus, to estimate $s^{(1)}(t; \beta)$, it suffices to model $m_{r, \delta}^{(1)}(X_r, Y; t, \beta) := \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) X | X_r, R = r, Y, \Delta = \delta]$ for all (r, δ) pairs. We obtain a similar expression for $s^{(0)}(t; \beta)$ via

$$s^{(0)}(t; \beta) = \sum_{r, \delta} \mathbb{E}[m_{r, \delta}^{(0)}(X_r, Y; t, \beta) 1(R = r) 1(\Delta = \delta)]$$

for $m_{r, \delta}^{(0)}(X_r, Y; t, \beta) := \mathbb{E}[1(Y \geq t) \exp(\beta^\top X) | X_r, R = r, Y, \Delta = \delta]$.

Applying Lemma 2.2 on the overall estimating equation, we obtain

$$U(\beta) = \sum_r \mathbb{E} \left[\mathbb{E} \left[X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \middle| X_r, R, Y, \Delta \right] \cdot 1(R = r) \cdot 1(\Delta = 1) \right].$$

Therefore, the regression adjustment method is equivalent to finding a near-solution to the

following finite sample estimating equation

$$\begin{aligned}
U_n(\beta) &= \sum_r \left[\frac{1}{n} \sum_{i=1}^n \left(\widehat{\mathbb{E}} \left[X - \frac{\widehat{s}^{(1)}(Y; \beta)}{\widehat{s}^{(0)}(Y; \beta)} \middle| X_{i,r}, R_i, Y_i, \Delta_i \right] \right) \cdot 1(R_i = r) \cdot 1(\Delta_i = 1) \right] = o_p(1/\sqrt{n}), \\
\widehat{s}^{(1)}(t; \beta) &= \sum_{r, \delta} \left[\frac{1}{n} \sum_{i=1}^n \widehat{m}_{r, \delta}^{(1)}(X_{i,r}, Y_i; t, \beta) 1(R_i = r) 1(\Delta_i = \delta) \right], \\
\widehat{s}^{(0)}(t; \beta) &= \sum_{r, \delta} \left[\frac{1}{n} \sum_{i=1}^n \widehat{m}_{r, \delta}^{(0)}(X_{i,r}, Y_i; t, \beta) 1(R_i = r) 1(\Delta_i = \delta) \right].
\end{aligned} \tag{2.11}$$

This can be achieved using a Monte Carlo procedure. We describe the multistage process below, and we summarize our sampling method in Algorithm 2. This algorithm is the same as performing a frequentist (type B) multiple imputation approach described in Wang and Robins (1998a) and Little and Rubin (2002).

1. STAGE 1 (MODELING THE MISSINGNESS OF X WITH EXTRAPOLATION DENSITIES).

As previously discussed, instead of modeling the regression functions $m_{r, \delta}^{(1)}(X_r, Y; t, \beta)$ and $m_{r, \delta}^{(0)}(X_r, Y; t, \beta)$ for all r and δ directly, we will instead focus on the extrapolation densities. Under CCMV, we simply need to model the joint distribution $p(x, R = 1_d, y, \delta)$. This will induce a model $p(x_{\bar{r}} | x_r, R = 1_d, y, \delta)$ for any covariate missing pattern r . By CCMV, that conditional distribution equates exactly to the extrapolation distribution for the missing data.

2. STAGE 2 (ESTIMATING INTERMEDIATE QUANTITIES). Next, we again outline how

to estimate the nuisance functions $s^{(1)}(t; \beta)$ and $s^{(0)}(t; \beta)$ that are dependent on the potentially missing covariates X . We estimate these using a Monte Carlo procedure through iterative imputation of the missing covariates.

By estimating the regression functions using M draws from the extrapolation distri-

bution, the following is a consistent estimator for $s^{(1)}(t; \beta)$

$$\begin{aligned}
\widehat{s}^{(1)}(t; \beta) &= \sum_{r, \delta} \left[\frac{1}{n} \sum_{i=1}^n \widehat{m}_{r, \delta}^{(1)}(X_{i, r}, Y_i; t, \beta) 1(R_i = r) 1(\Delta_i = \delta) \right] \\
&= \frac{1}{n} \sum_{i=1}^n \left(\sum_{r, \delta} \widehat{m}_{r, \delta}^{(1)}(X_{i, r}, Y_i; t, \beta) 1(R_i = r) 1(\Delta_i = \delta) \right) \\
&= \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{M} \sum_{j=1}^M 1(Y_i \geq t) \widetilde{X}_i^{(j)} e^{\beta^\top \widetilde{X}_i^{(j)}} \right), \\
&= \frac{1}{Mn} \sum_{j=1}^M \sum_{i=1}^n 1(Y_i \geq t) \widetilde{X}_i^{(j)} e^{\beta^\top \widetilde{X}_i^{(j)}},
\end{aligned}$$

where $\widetilde{X}_i^{(j)} = (X_{i, R_i}, \widetilde{X}_{i, \bar{R}_i}^{(j)})$. Here $\widetilde{X}_{i, \bar{R}_i}^{(j)}$ denotes the j th imputed tuple of missing covariates for the i th observation, and $\widetilde{X}_i^{(j)}$ is the result of attaching the imputed variables to the observed covariates X_{i, R_i} . A similar construction applied to $s^{(0)}(t; \beta)$ yields

$$\widehat{s}^{(0)}(t; \beta) = \frac{1}{Mn} \sum_{j=1}^M \sum_{i=1}^n 1(Y_i \geq t) e^{\beta^\top \widetilde{X}_i^{(j)}}.$$

3. **STAGE 3 (CONSTRUCTING THE FINAL ESTIMATOR).** Lastly, we again merge everything together in the final stage. Following the decomposition in (2.11), we have

$$U_n(\beta) = \frac{1}{Mn} \sum_{j=1}^M \sum_{i=1}^n \Delta_i \left(\widetilde{X}_i^{(j)} - \frac{\widehat{s}^{(1)}(Y_i; \beta)}{\widehat{s}^{(0)}(Y_i; \beta)} \right).$$

We solve the above estimating equation to obtain the final estimate $\widehat{\beta}$.

The terms we have derived suggest a simple procedure for the estimation. These estimators $\widehat{s}^{(1)}(t; \beta)$, $\widehat{s}^{(0)}(t; \beta)$, and $U_n(\beta)$ are all formed by merging the M imputed datasets together to form one large completed dataset with Mn entries. Thus, this regression adjustment method with a Monte Carlo approximation is equivalent to a stacked multiple

imputation approach utilized by [Robins and Wang \(2000\)](#); [Beesley and Taylor \(2020\)](#). Note that this approach differs from a traditional imputation approach where M imputed datasets are generated separately, analysis is performed on each one, and each analysis is pooled to form a final estimator. In contrast, in the stacked imputation approach, the pooling operation is done before the analysis such that M imputed datasets are pooled first and analysis is performed on all Mn observations to form a final point estimator.

An advantage of the traditional imputation approach is that the separate M analyses allows for parallelization. On the other hand, the stacked imputation does not exhibit this nice property, but a single analysis yields a more straightforward implementation, which can be helpful in situations where we perform the bootstrap.

Algorithm 2 Cox model using multiple imputation under CCMV for X

1. Input: $\{(X_{i,R_i}, R_i, Y_i, \Delta_i)\}_{i=1}^n$
 2. For $\delta = 0, 1$, fit the parametric model $p(x|R = 1_d, y, \delta)$.
 3. For $j = 1, \dots, M$, do
 - 3-1. For $i = 1, \dots, n$, do
 - i. Impute the missing covariates $\widetilde{X}_{i,R_i}^{(j)}$ for observation $(X_{i,R_i}, R_i, Y_i, \Delta_i)$ using the estimated extrapolation distribution $p(x_{\bar{r}}|x_r, R = 1_d, y, a)$.
 - ii. Set the j th imputation for the i th observation to be $(\widetilde{X}_i^{(j)}, R_i, Y_i, \Delta_i)$, where $\widetilde{X}_i^{(j)} = (X_{i,R_i}, \widetilde{X}_{i,R_i}^{(j)})$.
 - 3-2. Let $\widetilde{D}^{(j)} = \{(\widetilde{X}_i^{(j)}, R_i, Y_i, \Delta_i)\}_{i=1}^n$ denote the j th imputed data set.
 4. Pool all $\widetilde{D}^{(j)}, j = 1, \dots, M$ data sets together into a single large one $\check{D}$.
 5. Fit a Cox model using the completed data set $\check{D}$ to get an estimator $\widehat{\beta}$.
 6. Return: $\widehat{\beta}$.
-

In a frequentist multiple imputation approach, we first use the observed data to estimate the extrapolation densities and thereby, perform the imputation. Therefore, all further analysis is based on the estimated extrapolation densities, and it is important to incorporate this uncertainty in the final variance calculation. Rubin's estimator for the variance is commonly

used in practice, but it will underestimate the variance in this situation. Specifically, there will be an additional term in the asymptotic variance related to the estimator for the extrapolation distribution that will not be included. See Sections 14.4 and 14.5 of [Tsiatis \(2007\)](#) for more details.

To avoid this issue in practice, we recommend using the nonparametric bootstrap to estimate the variance. We bootstrap the data (including the missing values) many times and repeat the same analysis on each bootstrapped sample to obtain many point estimates. Each analysis on a bootstrapped sample will utilize a different estimated extrapolation distribution, so this uncertainty will be taken into account. [Chen and Sadinle \(2019\)](#) provide additional related details in their work.

2.3.3 A multiply-robust estimator

In the previous two sections, we have described two approaches based on an inverse probability weighting and a regression adjustment. In the seminal paper by [Robins et al. \(1994\)](#), the authors introduced a general framework of an augmented inverse probability weighted (AIPW) estimating function under a MAR assumption for the covariates, and thus, the selection probability did not depend on any missing covariates. [Wang and Chen \(2001\)](#) discussed further details on constructing an AIPW estimator with augmented functions for the Cox model under MAR. Additionally, in their paper, the covariates took the form of $(X_{\text{miss}}, X_{\text{obs}})$, where X_{miss} was a covariate vector possibly missing for some individuals and X_{obs} was a covariate vector that was always observed, thereby inducing only two possible missingness patterns for the covariates. These AIPW methods rely on two parametric models for the selection probabilities and the missing covariates. To relax this parametric assumption, [Qi et al. \(2005\)](#) outlined a fully augmented weighted estimator using a kernel regression approach for a one-dimensional covariate. However, it is unclear how to handle the case with

multiple covariates that may all be missing.

We extend the previous work by describing a multiply-robust estimator that allows missing not at random (MNAR) for a multidimensional covariate and arbitrary missing patterns. The following lemma shows that there exists an estimator that is robust to the misspecification of at most one of $Q_{r,a}(X_r, W_a)$ or $m_{r,a}(X_r, W_a)$ for each r .

Lemma 2.3 (Doubly robust linear form under CCMV). *Fix a missing pattern r and an observed outcome pattern a . Define*

$$g_r(X, W_a) = f_a(X, W_a)1(A = a)1(R = 1_d)Q_{r,a}(X_r, W_a) \quad (2.12)$$

$$+ m_{r,a}(X_r, W_a)1(R = r)1(A = a) \quad (2.13)$$

$$- m_{r,a}(X_r, W_a)1(A = a)1(R = 1_d)Q_{r,a}(X_r, W_a). \quad (2.14)$$

If $Q_{r,a}(X_r, W_a)$ or $m_{r,a}(X_r, W_a)$ is correctly specified, then under a CCMV assumption of equation (2.6), we have

$$\mathbb{E}[f_a(X, W_a)1(A = a)] = \sum_r \mathbb{E}[g_r(X, W_a)].$$

The expression of Lemma 2.3 shows an interesting connection to equation (10.95) of Tsiatis (2007). In the case where there is only two missing patterns of the covariates, i.e., the covariates can be expressed as $(X_{\text{miss}}, X_{\text{obs}})$, where X_{miss} are possibly missing covariates and X_{obs} are completely observed covariates, the MAR and CCMV assumptions on the covariates can be shown to be equivalent. Our expression and equation (10.95) in Tsiatis (2007) will be the same.

Theorem 2.2 (Multiply-robust Cox model). *Under CCMV, the estimator $\hat{\beta}_{mr}$ that solves the augmented estimating function is consistent provided either $Q_{r,\delta}(X_r, Y)$ or $m_{r,\delta}(X_r, Y)$*

is correctly specified for each r .

Thus, this implies that we can construct a multistage estimator based on Lemma 2.3 that is multiply-robust since we have doubly robustness for each covariate pattern $r \neq 1_d$. However, this comes at a cost. We see that the doubly robust expression above contains the terms from Lemmas 2.1 and 2.2 along with an additional term. Term (2.14) is of interest because it shows that in order to compute it, we will have to perform imputation even for cases with completely observed covariates. Therefore, we can see explicitly that this doubly robust estimator is even more computationally intensive than the regression adjustment estimator in Lemma 2.2.

2.4 Simulation

We examine two scenarios of the Cox model and the binary treatment problem to observe the consistency of our proposed estimators as well as behavior of the estimators under perturbations of the assumed identifying restriction.

2.4.1 Cox model simulation

We consider a Cox model with two binary covariates $X = (X_1, X_2) \in \{0, 1\}^2$ from the following generative model:

1. Sample $X = (X_1, X_2)$ such that $P(X_1 = 1) = 1/2$, $P(X_2 = 1) = 3/10$, and $\text{Corr}(X_1, X_2) = 0.3$.
2. For T and C (and Y and Δ), we do the following:
 - Generate $T|X$ from an Exponential distribution with rate $\exp(x^\top \beta)$ and $\beta = \begin{bmatrix} -0.5 \\ 2 \end{bmatrix}$. This is based on the baseline hazard $\lambda_0(t) = 1$.
 - Generate $C|X$ from an distribution with rate 2.

- Compute $Y = \min\{T, C\}$ and $\Delta = 1(T \leq C)$.

3. Generate the missing pattern of covariates based on CCMV:

$$\begin{aligned}\frac{P(R = 01|x, y, \delta)}{P(R = 11|x, y, \delta)} &= \frac{P(R = 01|x_2, y, \delta)}{P(R = 11|x_2, y, \delta)} = \exp(-0.5x_2 + 0.75y + 0.5\delta), \\ \frac{P(R = 10|x, y, \delta)}{P(R = 11|x, y, \delta)} &= \frac{P(R = 10|x_1, y, \delta)}{P(R = 11|x_1, y, \delta)} = \exp(-1 + 0.5x_1 + y + \delta),\end{aligned}$$

and remove the covariates accordingly.

Roughly, we expect approximately 40%, 30%, and 30% of the observations to fall in the classes $R = 01$, $R = 10$, and $R = 11$, respectively. This data generating mechanism completely specifies the joint model, and we consider three different estimators. The first estimator is the IPW estimator, which is based off the weighted Cox model from Section 2.3.1, and the complete odds are estimated using logistic regression under CCMV. The second estimator is multiple imputation (MI) / regression adjustment estimator as discussed in Section 2.3.2 with $M = 50$ imputations. Notably, the complete odds and the imputation distribution are not variation independent in this problem (and will generally not be in most problems of interest), so we have to fit the imputation distribution using knowledge of the complete odds. The third estimator is called the transformed MLE estimator. The specific parametric model assumed for $p_{Y,\Delta|X}(y, \delta|x)$ implies that we can recover the terms

$$\nu_1, \quad \nu_1 \exp(\beta_1), \quad \nu_1 \exp(\beta_2), \quad \text{and} \quad \nu_1 \exp(\beta_1 + \beta_2)$$

via maximum likelihood through the EM algorithm (details provided in Appendix A.5.1).

Let $\hat{\gamma}_i$ denote the estimator for the i th term above ($i = 1, 2, 3, 4$). Two naive estimators,

which we term the transformed MLEs, are as follows:

$$\begin{aligned}\hat{\beta}_{1,\text{Transformed MLE}} &= \frac{1}{2} \log(\hat{\gamma}_2/\hat{\gamma}_1) + \frac{1}{2} [\log(\hat{\gamma}_4/\hat{\gamma}_1) - \log(\hat{\gamma}_3/\hat{\gamma}_1)], \\ \hat{\beta}_{2,\text{Transformed MLE}} &= \frac{1}{2} \log(\hat{\gamma}_3/\hat{\gamma}_1) + \frac{1}{2} [\log(\hat{\gamma}_4/\hat{\gamma}_1) - \log(\hat{\gamma}_2/\hat{\gamma}_1)].\end{aligned}$$

We did not include the DR estimator in comparison because it has a very complex form (see Lemma 2.3).

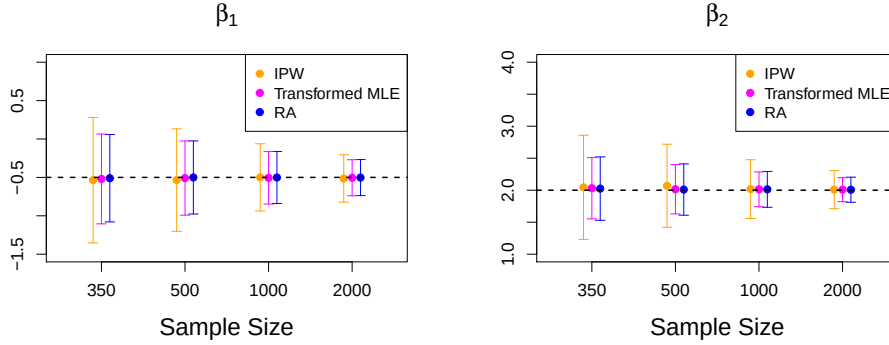


Figure 2.1: Estimation of β in a simulated Cox model for varying sample sizes.

For each of the sample sizes (350, 500, 1000, 2000), we generate 1000 data sets and construct the three estimators. The means and the standard errors are calculated across the 1000 replicates, and we show the 95% confidence interval. The results are summarized in Figure 2.1. We first observe that all three estimators are consistent and have standard errors that appear to roughly decline at the appropriate $\sqrt{n}$ -rate. The IPW estimator (equivalent to the reweighted Cox model) has the highest variance, which is to be expected because it uses the least information and inverse probabilities are known to introduce instability. The multiple imputation / regression adjustment (MI/RA) estimator uses slightly more information than the IPW estimator by assuming a parametric model for $p_{Y,\Delta|X}(y, \delta|x)$. As

the imputation distribution and the complete odds are typically not variation independent, we are forced to fit them together. We see that the RA estimator appears to have similar variance to the transformed MLE, which is to be expected for large enough M (number of imputations).

2.4.2 Binary treatment problem

We perform another simulation under the binary treatment setting from causal inference to show consistency under a new setting. The generating process is described in Appendix A.5.2. For each of the samples sizes (1000, 2000, 5000, 10000), we generate 1000 data sets and construct 6 estimators. We consider two methods for accounting for the missing covariates (IPW- R and RA- R) and three methods for estimating the ATE (IPW- A , RA- A and DR- A). So, there are 6 combinations of one method for missing covariates and one method for the ATE. The left panel of Figure 2.2 displays the results of three estimators formed by using an IPW approach for the missing covariates: (IPW- R , IPW- A), (IPW- R , RA- A), and (IPW- R , DR- A). In this case, we use the bootstrap percentile method for computing the confidence intervals. The right panel of Figure 2.2 provides the results of three estimators using a RA (imputation) approach for the missing covariates: (RA- R , IPW- A), (RA- R , RA- A), and (RA- R , DR- A). We use $M = 50$ imputations and report the 95% confidence intervals. We simply construct the confidence intervals by computing the standard errors across all 1000 generated data sets. Again, we also report the means across the 1000 replicates.

In the left panel of Figure 2.2, we observe some surprising results. First, in finite samples, the (IPW- R , IPW- A) estimator can lead to unstable estimates for some data sets due to the fact that the complete odds are estimated using a density ratio. Although we expect the doubly-robust estimator (IPW- R , DR- A) to have smaller variance than the (IPW- R , IPW- A) estimator, we see that it has larger variance. This is because the unstable part of

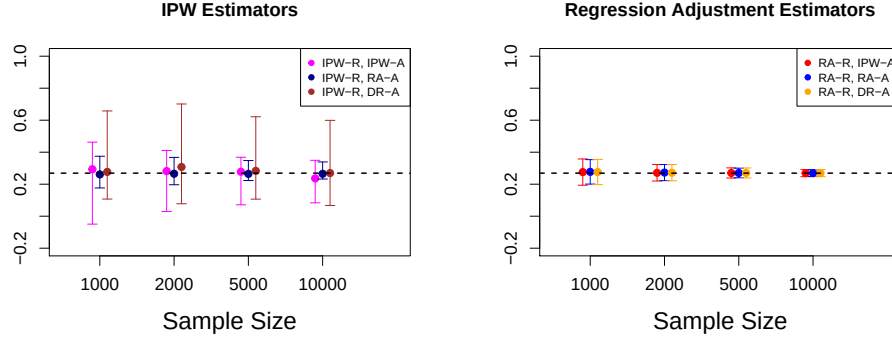


Figure 2.2: Estimation of the ATE in a simulated causal inference problem for varying sample sizes.

IPW- R appears in each of the 3 terms of the DR- A estimator.

In the right panel of Figure 2.2, we see very similar performance across all 3 estimators. These 3 estimators dominate the other 3 estimators in the right panel, which suggests that a multiple imputation approach can provide more stable results in practice. We discuss further sensitivity analysis for this simulation in Appendix A.4.2.

2.5 Data Analysis

We apply our methods on the data from the National Alzheimer’s Coordinating Center’s (NACC) Uniform Data Set¹. This data set contains longitudinal data collected at Alzheimer’s Disease Research Centers, funded by the National Institutes of Health and starting from the year 2005. Participating individuals have varying cognitive statuses from no cognitive impairment to mild impairment to severe impairment, and data is collected on a yearly basis with certain covariates measured at entry into the study.

We consider individuals who display mild cognitive impairment (MCI) and follow their trajectory until they either display evidence of dementia (mild, moderate, severe) or are

¹<https://nacccdata.org/>

censored. These individuals are binned into 5 age groups based on when they first display MCI: 60-64, 65-69, 70-74, 75-79, and 80-84. In each age group, we fit a Cox model, regressing on the binary variables: sex (female or not), race (white or not), education (have graduate degree or not), depression within last 2 years (yes or no), and anxiety (yes or no). Tracking the change in the β across the different age groups allows us to observe the age effect. The first three covariates (sex, race, and education) are generally measured for all individuals at the baseline. There are a few individuals for which the education status is unknown, but because these represent less than 1% of the observations, we exclude these individuals from the analysis for simplicity. The last two covariates (depression and anxiety) are subject to missingness, yielding a total of $2^2 = 4$ covariate missing patterns. There are about 16% participants in our data with complete observations.

We consider three different estimators: complete case (CC) analysis, IPW and RA estimators. The CC analysis is a common approach used by practitioners. In the CC analysis, we remove any individuals with any missing covariates and fit a standard Cox model. For the IPW estimator, we use the CCMV assumption using logistic models for the complete odds. Since the variables are binary, the imputation distribution in the RA estimator is the usual nonparametric model. Point estimates and the 95% confidence intervals (constructed using the nonparametric bootstrap) are provided in Figure 2.3.

First, we observe that the results from a CC analysis are generally not significant. For example, in the plot for β_3 (education), the CC confidence intervals all strongly cover 0. On the other hand, across the plots, the RA approach generally yields narrower confidence intervals while the IPW approach sometimes yields narrower confidence intervals; this is due to the instability introduced by the inverse probability weights. Specifically, for β_3 (education), the RA provides clear significant results for several of the age groups. These plots suggest that often, when applying a CC analysis, we suffer from such a loss in sample

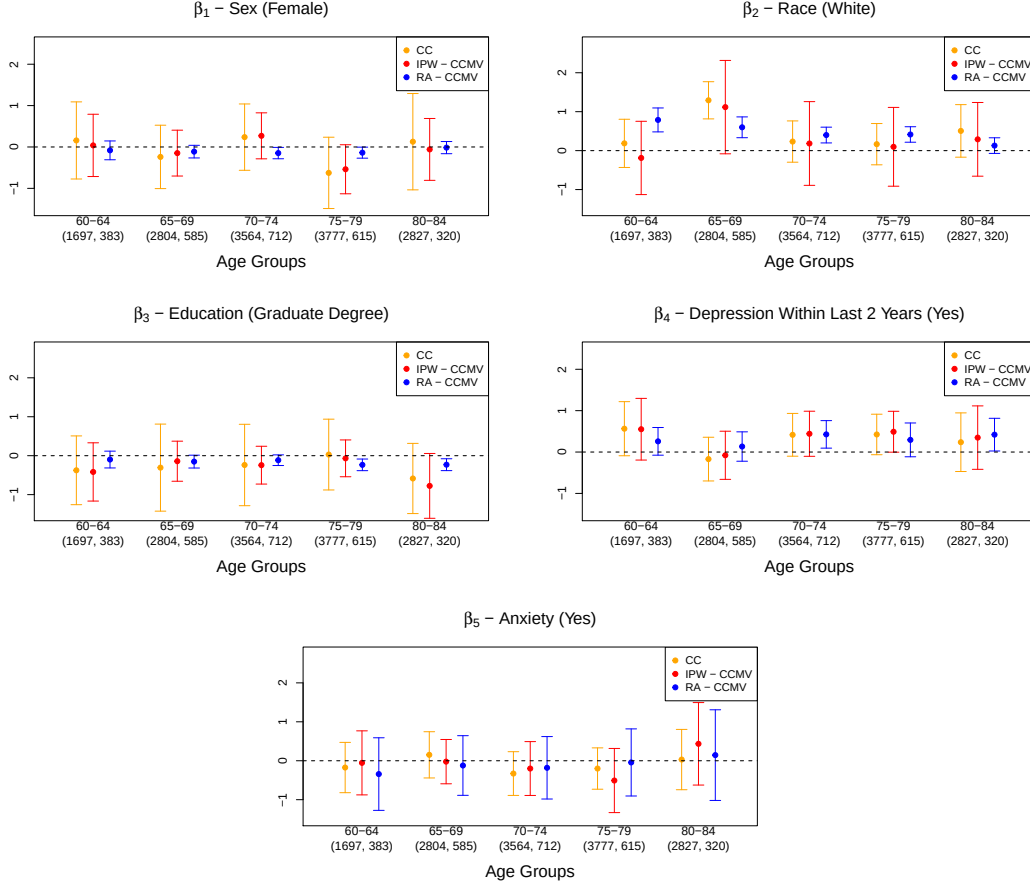


Figure 2.3: Point estimates and the 95% confidence intervals for each β parameter. The total sample size and the number of complete cases is included on the x -axis for each age group.

size that it becomes difficult to draw a significant conclusion.

Moreover, it is known that higher education provides a protection against cognitive impairment (known as the cognitive reserve; Meng and D'Arcy 2012; Thow et al. 2018). The point estimates for β_3 from all three estimation procedures are negative, which is what we would expect. As the RA approach allows one to attain significant negative point estimates, it highlights the potential benefits of accounting for the missing data. We discuss further details for the analysis in Appendix A.6.

2.5.1 Sensitivity analysis

We conduct a sensitivity analysis to investigate the robustness of our conclusion under CCMV. We perturb the complete odds using an exponential tilting parameter ρ such that

$$O_{r,a}(X, W_a) = Q_{r,a}(X_r, W_a)e^{\rho_{\bar{r}}X_{\bar{r}}}$$

and every element in the vector $\rho_{\bar{r}}$ takes the same value ρ . When $\rho = 0$, there is no perturbation, but as $|\rho| \rightarrow \infty$, the scale of the perturbation increases. For the imputation model, we use a sensitivity analysis parameter $\zeta \in [0, 1]$. When $\zeta \rightarrow 1$, the imputation probability moves towards 0, and when $\zeta \rightarrow 0$, the imputation probability moves towards 1. If $\zeta = 1/2$, then the imputation probability remains unchanged (i.e., the case of CCMV). We provide additional discussion of these sensitivity parameters in Appendix A.4. The results are displayed in Figures 2.4 and 2.5. The confidence intervals are shaded in gray. There is some slight wiggles in the upper and lower bands because of the Monte Carlo errors from the bootstrap (we construct each confidence interval using the nonparametric bootstrap).

In Figure 2.4, we display the result of the IPW estimator. The IPW estimator is robust when we perturb the missing data assumption in the sense that only the education effect has a stronger dependency with the sensitivity parameter. Across all $\rho \in [-1, 1]$, the education effect is still negative, which suggests that the cognitive reserve we obtain from the IPW estimator is resilient.

In Figure 2.5, we observe that the RA estimator for β_3 (education effect) is robust to changes in the sensitivity parameter, which again reinforces our findings on the cognitive reserve. On the other hand, we see that the confidence interval width of β_4 (depression) depends on the sensitivity parameter ζ . While the sign of the effect remains the same and the effect (value of β_4) is also stable, the width of confidence intervals do change drastically.

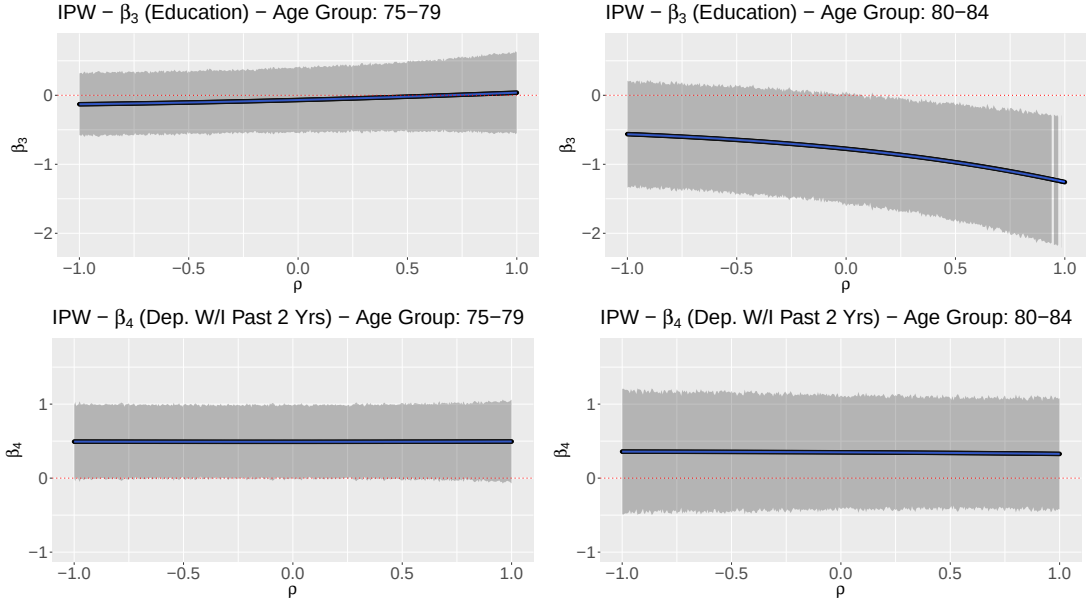


Figure 2.4: Point estimates and the 95% confidence intervals for the IPW estimator (β_3 and β_4) as we vary the sensitivity parameter ρ for the age groups 75 – 79 and 80 – 84.

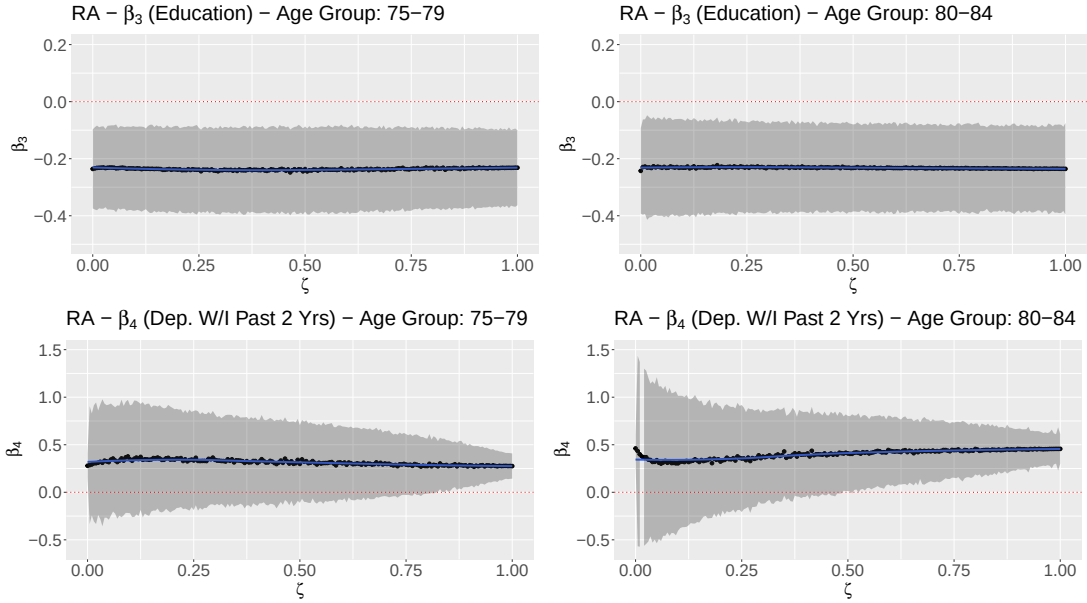


Figure 2.5: Point estimates and the 95% confidence intervals for the RA estimator (β_3 and β_4) as we vary the sensitivity parameter ζ for the age groups 75 – 79 and 80 – 84.

So, we have to be careful when making claims regarding β_4 .

2.6 Discussion

In this chapter, we introduced a process for constructing a multistage estimator for problems with missing not at random covariates and incomplete outcomes. We discussed our method in the context of Cox regression from survival analysis and a missing response problem (in Appendix A.1). Although we did not show the specific details, the missing response problem provides a bridge to the binary treatment problem from causal inference and a more general longitudinal study problem with dropout. So, the arguments we provide for the missing response problem naturally extend to these two problems.

Figure 2.6 shows an interesting tradeoff between model complexity, computation, and efficiency in the special case of the CCMV assumption for the covariates. The RA estimator is the easiest to model since we only need to place a model on $p(x|1_d, w_a, a)$ for all a . But this comes at a cost of computation since we generally need to perform imputation to fill in the missing data. Also, RA does not achieve the (semiparametric) efficiency bound. The IPW estimator is easy to compute because it is a simple reweighting of the cases with completely observed covariates. But we need to model every $Q_{r,a}(X_r, W_a)$, which would require many models. Also, IPW does not achieve the (semiparametric) efficiency bound and

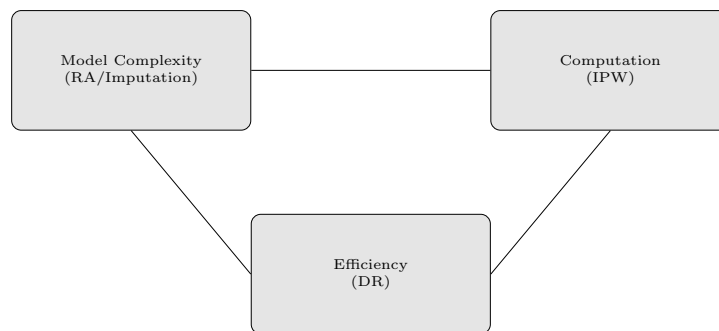


Figure 2.6: The tradeoff between model complexity, computation, and efficiency in the CCMV setting.

may be unstable if the weights are too large. The DR estimator achieves the semiparametric efficiency bound by utilizing components from both the regression adjustment and inverse probability weighted estimator. Due to the augmenting term, imputation must be performed even for the complete cases, so the computational cost is high and we have to place models for every $Q_{r,a}(X_r, W_a)$. Thus, efficiency comes at the cost of complexity in modeling and high computation.

There are several possible areas for future work. One direction is to study the problem of model compatibility issues between complete odds and regression functions. When the assumption on missing covariates is not CCMV, the complete odds and regression function may again not be variationally independent. Thus, extra caution has to be taken when placing models on them. Another direction is to study the use of nonparametric models in complete odds and regression functions. This would lead to a more flexible modeling strategy and avoid the problem of model compatibility, but overcoming the slow convergence rate will be a challenge.

Appendix A: Chapter 2 Appendices

The appendix has several sections to provide additional theoretical results, proofs, details on simulations and data analyses, and further generalizations of work in the main body. In Appendix A.1, we construct a multistage estimator for the missing response problem. The assumptions for the asymptotic theory are provided in Appendix A.2. Since we utilize CCMV as the main missing data assumption in the main body, we generalize this assumption in Appendix A.3 and show how to attain identifiable complete odds and imputation distributions. Further details on the sensitivity analyses and simulations are included in Appendices A.4 and A.5, respectively. We include more data analyses in Appendix A.6. The proofs for any theoretical results are included in Appendix B.8.

The table of contents is provided below.

1. Appendix A.1: Missing response problem.
2. Appendix A.2: Technical assumptions for the asymptotic theory.
3. Appendix A.3: Beyond CCMV.
4. Appendix A.4: Procedure of sensitivity analysis.
5. Appendix A.5: Details of simulations.
6. Appendix A.6: Additional analysis of the NACC data.
7. Appendix B.8: Proofs of 1. identification theories (Appendix A.7.1), 2. efficiency theories (Appendix A.7.2), 3. asymptotic theories (Appendix A.7.3)

A.1 A Multistage Estimator for the Missing Response Problem

We now turn our attention to the missing response problem and show how to construct a multistage estimator in this setting. Note that this procedure can be easily generalized to the

binary treatment effect in causal inference. The parameter of interest is the marginal mean of the outcome variable $\mathbb{E}[Y]$. Let $\pi_A(X) = P(A = 1|X)$ and $m_1(X)$ denote the propensity score and regression function for Y , respectively. We start with the classic DR estimator assuming fully-observed covariates for the missing response problem. The efficient influence function takes the form

$$\frac{1(A=1)Y}{\pi_A(X)} + m_1(X) - \frac{1(A=1)m_1(X)}{\pi_A(X)}.$$

The expectation of the above quantity equals $\mathbb{E}[Y]$ as long as at least one of $\pi_A(X)$ and $m_1(X)$ is correctly specified. Using empirical average of the above quantity, we obtain the DR estimator under the missing response problem.

A.1.1 Multiply-robust property

In this section, we highlight the multiply-robust property of the multistage estimator by demonstrating the construction of one such multiply-robust estimator in the missing response problem.

For this problem, we also define the following regression functions

$$M_{m,r,0}(X_r) = \mathbb{E}[m_1(X)|X_r, R = r, A = 0], \quad (2.15)$$

$$M_{m,r,1}(X_r, Y) = \mathbb{E}[m_1(X)|X_r, R = r, Y, A = 1], \quad (2.16)$$

$$M_{1/\pi_A,r,1}(X_r, Y) = \mathbb{E}\left[\frac{1}{\pi_A(X)} \middle| X_r, R = r, Y, A = 1\right], \quad (2.17)$$

$$M_{m/\pi_A,r,1}(X_r, Y) = \mathbb{E}\left[\frac{m_1(X)}{\pi_A(X)} \middle| X_r, R = r, Y, A = 1\right]. \quad (2.18)$$

Definition 2.4 (Semiparametric submodels). Define the following semiparametric submodels.

- $\mathcal{M}_R(r)$: The semiparametric model such that the parametric model $p(x_{\bar{r}}|x_r, r, w_a, a; \eta_{r,a})$

is correctly specified.

- $\mathcal{M}_O(r)$: The semiparametric model such that the parametric model $O_{r,a}(x, w_a; \alpha_{r,a})$ is correctly specified.

It follows from the previous definition that under model $\mathcal{M}_R(r)$, the regression functions (2.15)-(2.18) are correctly specified. We specify the semiparametric model in this manner to avoid the model congeniality issue described in Section 2.3.2. Note that one could instead define $\mathcal{M}_R(r)$ to be the semiparametric model under which these regression functions are correctly specified.

Theorem 2.3 (Double robustness for missing response under CCMV for given r). *Let $\theta_r = \int y P(dx, r, dy, da)$ be the parameter of interest. Consider the uncentered influence function*

$$\begin{aligned}
 q_r(X, R, Y, A) = & m_1(X)1(A=0)1(R=1_d)Q_{r,0}(X_r) \\
 & + m_1(X)1(A=1)1(R=1_d)Q_{r,1}(X_r, Y) \\
 & + M_{m,r,0}(X_r)1(A=0)[1(R=r) - 1(R=1_d)Q_{r,0}(X_r)] \\
 & + M_{m,r,1}(X_r, Y)1(A=1)[1(R=r) - 1(R=1_d)Q_{r,1}(X_r, Y)] \\
 & + \frac{1(R=1_d)1(A=1)Y}{\pi_A(X)}Q_{r,1}(X_r, Y) \\
 & + M_{1/\pi_A,r,1}(X_r, Y)Y1(A=1)[1(R=r) - 1(R=1_d)Q_{r,1}(X_r, Y)] \\
 & - \frac{1(R=1_d)1(A=1)m_1(X)}{\pi_A(X)}Q_{r,1}(X_r, Y) \\
 & - M_{m/\pi_A,r,1}(X_r, Y)1(A=1)[1(R=r) - 1(R=1_d)Q_{r,1}(X_r, Y)].
 \end{aligned}$$

Under the model $\mathcal{M}_R(r) \cup \mathcal{M}_O(r)$ (i.e., either the complete odds or the regression functions are correct), $\mathbb{E}[q_r(X, R, Y, A)] = \theta_r$.

Theorem 2.3 shows that the estimator for θ_r based on the provided influence function

is doubly robust. One key takeaway is that for a fairly simple problem, the estimation procedure for the doubly robust estimator is rather complex. In particular, there are 12 terms in total. With this in mind, we expect that this doubly robust estimator to be more of a theoretical construct than one that will generally be implemented in practice.

We further develop the theory by considering a doubly robust estimator for θ_r for each r . Merging this together, we are able to estimate

$$\theta = \int y P(dx, dr, dy, da) = \sum_r \theta_r.$$

We see that in the following corollary, this yields a multiply-robust estimator for $\theta = \mathbb{E}[Y]$.

Corollary 2.1 (Multiply-robustness for missing response under CCMV). *Define the parameter of interest*

$$\theta = \sum_r \theta_r = \sum_r \int y P(dx, r, dy, da) = \int y P(dx, dr, dy, da).$$

The uncentered influence function $\sum_r q_r(X, R, Y, A)$ yields a consistent estimator of $\theta = \sum_r \theta_r$ under the model $\bigcap_r (\mathcal{M}_R(r) \cup \mathcal{M}_O(r))$.

Corollary 2.1 provides a strong result in that the estimator is $2^{|\mathcal{R}|-1}$ -robust, where $|\mathcal{R}|$ is the number of missing covariate patterns. As shown in Theorem 2.3, for an estimator of θ_r to be consistent, it is sufficient for at least one of $O_{r,a}(X, W_a)$ or the distribution $p(x_{\bar{r}}|x_r, w_a, a)$ to be correctly specified. In other words, we have two chances to obtain a correct estimator for each missing covariate pattern $r \neq 1_d$.

Since the missing response and binary treatment problems are related and well-studied problems, we omit the construction of our multistage estimator for the binary treatment problem. The included arguments can easily be extended to the binary treatment problem

and the problem of longitudinal data with monotone missingness in the response.

A.1.2 Semiparametric efficiency

We now show that the multiply-robust estimator developed in Theorem 2.3 and Corollary 2.1 is the semiparametric efficient estimator that attains the minimum possible variance in the model $\bigcap_r (\mathcal{M}_R(r) \cap \mathcal{M}_O(r))$, where all the regression functions and complete odds are correctly specified.

Theorem 2.4 (EIF for missing response under CCMV for X). *For the mean statistical functional*

$$\psi(P) = \int y P(dx, dr, dy, da),$$

the multiply-robust influence function in Corollary 2.1 is the efficient influence function (EIF), and it achieves the semiparametric efficiency bound in the model $\bigcap_r (\mathcal{M}_R(r) \cap \mathcal{M}_O(r))$.

This theorem further shows that the multiply-robust estimator is also full semiparametric efficient.

A.2 Asymptotic Theory

In this section, we discuss the asymptotic theory for our multistage estimators. We direct our attention to estimating the population quantity

$$\theta = \mathbb{E}[h(X, W, A)] = \sum_{a \in \mathcal{A}} \mathbb{E}[f_a(X, W_a) 1(A = a)],$$

where is an ODF as described in Definition 2.1. We will show that the estimation methods that we have previously outlined lead to consistent, asymptotically normal estimators under appropriate conditions. This theory will be useful for constructing confidence intervals for data analysis and demonstrating consistency. If h is an estimating equation, then the

asymptotic normality of the solution to the finite-sample version of the estimating equation follows under standard regularity conditions for Z -estimators.

A.2.1 Assumptions for estimating f_a

Fix an observed outcome pattern a . These assumptions hold for estimating the function f_a .

(A1) There exists τ_a^* in the interior of the parameter space T_a such that

$$f_a(x, w_a; \tau_a^*) = f_a(x, w_a).$$

We assume that there exists $\hat{\tau}_a$, which is a regular asymptotically linear estimator such that

$$\hat{\tau}_a - \tau_a^* = \frac{1}{n} \sum_{i=1}^n \phi_{r,a}(X_i, R_i, W_{i,A_i}, A_i) + o_p(1/\sqrt{n}),$$

and the following convergence in distribution holds

$$\hat{\tau}_a - \tau_a^* \xrightarrow{d} N(0, \sigma_{f_a}^2)$$

for some $\sigma_{f_a}^2 > 0$.

(A2) Assume that f_a is uniformly bounded for all values of its parameter such that

$$\sup_{x, w_a, \tau_a} |f_a(x, w_a; \tau_a)| < M$$

for some constant $M > 0$.

(A3) Let $\int (f_a(x, w_a; \hat{\tau}_a) - f_a(x, w_a; \tau_a^*))^2 P(dx, dr, dw_a, da) = o_p(1)$.

(A4) Let $f_a(x, r, w_a; \tau_a)$ have a nonzero first derivative with respect to τ_a , and assume that

$$\int |\nabla_{\tau_a} f_a(x, w_a; \tau_a)| P(dx, dr, dw_a, da)$$

is finite in all of its entries.

The assumption (A1) says that the f_a functions belong to a parametric class and can be estimated at $\sqrt{n}$ -consistency. We elect to use a parametric assumption to make the analysis easier. Assumption (A2) is a bounded assumption, which is reasonable if the input variables have a compact support. Assumption (A3) describes the $L_2(P)$ convergence of the estimated f_a functions. Finally, Assumption (A4) is a technical assumption to ensure that asymptotic normality can be achieved and allows for application of the Dominated Convergence Theorem.

A.2.2 The IPW estimator

Assumptions

We now introduce the following 4 assumptions for the IPW theorem below. Each assumption holds for each pair of fixed covariate pattern r and observed-outcome pattern a .

(B1) For all missing covariate patterns r and observed outcome patterns a , there exist constants C_1 and C_2 such that

$$0 < C_1 \leq O_{r,a}(x, w_a; \alpha_{r,a}) \leq C_2 < 1.$$

(B2) There exists $\alpha_{r,a}^*$ in the interior of the parameter space $A_{r,a}$ such that

$$O_{r,a}(x, w_a; \alpha_{r,a}^*) = \frac{P(R = r|x, a, w_a)}{P(R = 1_a|x, a, w_a)}.$$

We assume that there exists $\widehat{\alpha}_{r,a}$, which is a regular asymptotically linear estimator

$$\widehat{\alpha}_{r,a} - \alpha_{r,a}^* \xrightarrow{d} N(0, \sigma_{\text{odds},r,a}^2)$$

for some $\sigma_{\text{odds},r,a}^2 > 0$.

(B3) The following class of functions is Donsker:

$$\{O_{r,a}(x, w_a; \alpha_{r,a} : \alpha_{r,a} \in A_{r,a}\}.$$

(B4) Assume that $O_{r,a}(x, w_a; \alpha_{r,a})$ have a nonzero first derivative with respect to $\alpha_{r,a}$. Also, let

$$\int |\nabla_{\alpha_{r,a}} O_{r,a}(x, w_a; \alpha_{r,a})| P(dx, dr, dw_a, da)$$

be finite in all of its entries, and assume that

$$\int (O_{r,a}(x, w_a; \widehat{\alpha}_{r,a}) - O_{r,a}(x, w_a; \alpha_{r,a}^*))^2 P(dx, dr, dw_a, da) = o_p(1).$$

Assumption (B1) is to ensure that the complete odds are sufficiently bounded away from 0, which is true if the variables belong to a compact support. Assumption (B2) says that the complete odds belong to a parametric model, and there exists an estimator that is $\sqrt{n}$ -consistent. This is reasonable for most common parametric models under consideration. Assumption (B3) is a Donsker condition, which is also reasonable if the parametric model is sufficiently smooth. Assumption (B4) ensures that the estimated regression function $O_{r,a}(x, w_a; \widehat{\alpha}_{r,a})$ converges in the $L_2(P)$ norm to the true regression function $O_{r,a}(x, w_a; \alpha_{r,a}^*)$. The finite first derivative ensures that we can apply the Dominated Convergence Theorem to simplify the expression in the proof.

Asymptotic theory

We continue our goal of estimating θ by using an inverse probability weighting approach. First, using Lemma 2.1, the parameter of interest writes as

$$\theta = \sum_a \sum_r \mathbb{E}[f_a(X, W_a) O_{r,a}(X, W_a) 1(R = 1_a) 1(A = a)],$$

where $O_{r,a}(X, W_a) = P(R = r|X, W_a, A = a)/P(R = 1_a|X, W_a, A = a)$ is a complete odds.

Theorem 2.5 (Inverse probability weighting). *Suppose the assumptions (A1)-(A4) and (B1)-(B4) hold. Then, the estimator*

$$\begin{aligned} \hat{\theta}_{IPW} &= \sum_r \sum_a \hat{\theta}_{IPW,r,a} \\ &= \sum_r \sum_a \left(\frac{1}{n} \sum_{i=1}^n f_a(X_i, W_{i,a}) O_{r,a}(X_i, W_{i,a}) 1(R_i = 1_a) 1(A_i = a) \right) \end{aligned}$$

is consistent, and for some $\sigma_{IPW}^2 > 0$, we have the following convergence in distribution

$$\sqrt{n}(\hat{\theta}_{IPW} - \theta) \xrightarrow{d} N(0, \sigma_{IPW}^2).$$

A.2.3 The RA estimator

Assumptions

We introduce the following 3 assumptions for the RA estimator.

(C1) For all covariate patterns r and observed-outcome patterns a , there exists $\eta_{r,a}^*$ in the interior of the parameter space $\mathcal{H}_{r,a}$ such that

$$P_{ex}(x_{\bar{r}}|x_r, r, w_a, a) = P_{ex}(x_{\bar{r}}|x_r, r, w_a, a; \eta_{r,a}^*).$$

Let $\widehat{P}_{ex}(x_{\bar{r}}|x_r, r, w_a, a) \equiv P_{ex}(x_{\bar{r}}|x_r, r, w_a, a; \widehat{\eta}_{r,a})$. We assume that $\widehat{\eta}_{r,a}$ is a regular asymptotically linear (RAL) estimator satisfying

$$\widehat{\eta}_{r,a} - \eta_{r,a}^* = \frac{1}{n} \sum_{i=1}^n \psi_{r,a}(X_i, R_i, W_{i,A_i}, A_i) + o_p(1/\sqrt{n}).$$

Moreover, $\sqrt{n}(\widehat{\eta}_{r,a} - \eta_{r,a}^*) \xrightarrow{d} N(0, \sigma_{\text{RA},r,a}^2)$ for some $\sigma_{\text{RA},r,a}^2 > 0$.

(C2) For a fixed covariate pattern r and observed-outcome pattern a , define

$$\begin{aligned} m_{r,a}(x_r, w_a; \eta_{r,a}) &= \mathbb{E}[f_a(X, W_a)|X_r = x_r, R = r, W_a = w_a, A = a] \\ &= \int f_a(x, w_a) P_{ex}(dx_{\bar{r}}|x_r, r, w_a, a; \eta_{r,a}). \end{aligned}$$

The following class of regression functions is Donsker:

$$\{m_{r,a}(x_r, w_a; \eta_{r,a}) : \eta_{r,a} \in \mathcal{H}_{r,a}\}.$$

(C3) For all covariate patterns r and observed-outcome patterns a , assume that $m_{r,a}(x_r, w_a; \eta_{r,a})$ has a nonzero first derivative with respect to $\eta_{r,a}$. Also, let

$$\int |\nabla_{\eta_{r,a}} m_{r,a}(x_r, w_a; \eta_{r,a})| P_{obs}(dx_r, r, dw_a, a)$$

be finite in all of its entries, and assume that

$$\int (m_{r,a}(x_r, w_a; \widehat{\eta}_{r,a}) - m_{r,a}(x_r, w_a; \eta_{r,a}^*))^2 P_{obs}(dx_r, r, dw_a, a) = o_p(1).$$

These are fairly mild assumptions. Assumption (C1) says that the extrapolation distri-

bution belongs to a parametric family and that the estimator $\hat{\eta}_{r,a}$ admits an asymptotically linear expansion. Provided the parametric model holds, the $\sqrt{n}$ -consistency of $\hat{\beta}_{r,a}$ is a reasonable assumption since $\eta_{r,a}^*$ is often estimated using M -estimators. Assumptions (C2) and (C3) are similar to assumptions (B3) and (B4) made in the IPW estimation theorem.

A.2.4 Asymptotic theory

We can estimate θ using a regression adjustment approach. Using Lemma 2.2, the parameter writes as

$$\begin{aligned}\theta &= \mathbb{E}[h(X, W, A)] \\ &= \sum_a \mathbb{E}[f_a(X, W_a)1(A = a)] \\ &= \sum_a \sum_r \mathbb{E}[m_{r,a}(X_r, W_a)1(R = r)1(A = a)] \\ &= \sum_a \sum_r \theta_{RA,r,a}.\end{aligned}$$

Thus, the main strategy will be to show asymptotic results for every fixed r and a pair.

Theorem 2.6 (Regression adjustment). *Suppose the assumptions (A1)-(A4) and (C1)-(C3) hold. Then, the estimator*

$$\begin{aligned}\hat{\theta}_{RA} &= \sum_r \sum_a \hat{\theta}_{RA,r,a} \\ &= \sum_r \sum_a \left(\frac{1}{n} \sum_{i=1}^n m_{r,a}(X_r, W_a; \hat{\eta}_{r,a}) 1(R_i = r) 1(A_i = a) \right)\end{aligned}$$

is consistent, and for some σ_{RA}^2 , we have the following convergence in distribution

$$\sqrt{n}(\hat{\theta}_{RA} - \theta) \xrightarrow{d} N(0, \sigma_{RA}^2).$$

Theorem 2.6 establishes that a population quantity expressible in the form (A.2) can be estimated via a regression adjustment approach. The asymptotic normality of the provided estimator can be used to construct a confidence interval for inference. Although it is possible to construct a sandwich estimator, we again recommend using the bootstrap to estimate the variance because writing a closed form expression may not be so elegant.

A.3 Beyond CCMV

We extend the work in the main body of the paper to see the analysis under an identifying restriction that is not CCMV. We consider the case with 2 covariate variables (X_1, X_2) subject to missingness for a total of $2^2 = 4$ possible patterns. For simplicity, let $v = (w_a, a)$. The CCMV assumption (Definition 2.3) is equivalent to:

$$\begin{aligned} p(x_1|x_2, R = 01, v) &= p(x_1|x_2, R = 11, v), \\ p(x_2|x_1, R = 10, v) &= p(x_2|x_1, R = 11, v), \\ p(x_1, x_2|R = 00, v) &= p(x_1, x_2|R = 11, v). \end{aligned} \tag{2.19}$$

The last equality can be relaxed by other assumptions. Here we consider the following identifying restriction:

$$\begin{aligned} p(x_1|x_2, R = 01, v) &= p(x_1|x_2, R = 11, v), \\ p(x_2|x_1, R = 10, v) &= p(x_2|x_1, R = 11, v), \\ p(x_1, x_2|R = 00, v) &= \kappa_1 \cdot p(x_1, x_2|R = 01, v) + \kappa_2 \cdot p(x_1, x_2|R = 10, v) + \kappa_3 \cdot p(x_1, x_2|R = 11, v) \end{aligned} \tag{2.20}$$

for nonnegative constants κ_1, κ_2 , and κ_3 satisfying $\kappa_1 + \kappa_2 + \kappa_3 = 1$. When $\kappa_3 = 1$, equation (2.20) reduces to the CCMV in equation (2.19). If there exist variables that are always observed, then the above identifying restrictions can be generalized to allow κ_1, κ_2 , and κ_3

to be real-valued functions of the observed variables. Under suitable choice of $\kappa_1, \kappa_2, \kappa_3$ (as a function of observed variables), the last restriction in equation (2.20) becomes

$$p(x_1, x_2 | R = 00, v) = p(x_1, x_2 | R \in \{01, 10, 11\}, v),$$

which shows connections to the available-case missing value assumption in (Molenberghs et al., 1998).

A.3.1 IPW estimators

We now show that the (complete) odds admits a nice identifiable representation. Let $V = (W_A, A)$. The two odds

$$\frac{P(R = 10 | X, V)}{P(R = 11 | X, V)} \quad \text{and} \quad \frac{P(R = 01 | X, V)}{P(R = 11 | X, V)}$$

are identifiable via the same formulas under CCMV. We now decompose the odds $P(R = 00 | X, V) / P(R \in \{01, 10, 11\} | X, V)$ via

$$\begin{aligned} & \kappa_1 \cdot \frac{p(x_1, x_2 | R = 01, v) P(R = 00 | v)}{p(x_1, x_2, R \in \{01, 10, 11\} | v)} + \kappa_2 \cdot \frac{p(x_1, x_2 | R = 10, v) P(R = 00 | v)}{p(x_1, x_2, R \in \{01, 10, 11\} | v)} \\ & + \kappa_3 \cdot \frac{p(x_1, x_2 | R = 11, v) P(R = 00 | v)}{p(x_1, x_2, R \in \{01, 10, 11\} | v)}. \end{aligned}$$

The first term can be decomposed by

$$\begin{aligned} & \kappa_1 \cdot \frac{p(x_1, x_2 | R = 01, v) P(R = 00 | v)}{p(x_1, x_2, R \in \{01, 10, 11\} | v)} \\ & = \kappa_1 \cdot \frac{p(x_1 | x_2, R = 01, v) p(x_2 | R = 01, v) P(R = 00 | v)}{p(x_2 | x_1, R = 11, v) p(x_1, R = 10 | v) + p(x_1 | x_2, R = 11 | v) p(x_2, R = 01 | v) + p(x_1, x_2, R = 11 | v)} \end{aligned}$$

$$\begin{aligned}
&= \kappa_1 \cdot \frac{p(x_2|R=01, v)P(R=00|v)}{\frac{p(x_2|x_1, R=11, v)p(x_1, R=10|v)}{p(x_1|x_2, R=01, v)} + p(x_2, R=01|v) + p(x_2, R=11|v)} \\
&= \kappa_1 \cdot \frac{p(x_2|R=01, v)P(R=00|v)}{\frac{p(x_1, x_2, R=11|v)p(x_1, R=10|v)p(x_2, R=01|v)}{p(x_1, x_2, R=01|v)p(x_1, R=11|v)} + p(x_2, R=01|v) + p(x_2, R=11|v)} \\
&= \kappa_1 \cdot \frac{p(x_2|R=01, v)P(R=00|v)}{\frac{P(R=11|x_2, v)}{P(R=01|x_2, v)} \cdot \frac{P(R=10|x_1, v)}{P(R=11|x_1, v)} \cdot p(x_2, R=01|v) + p(x_2, R=01|v) + p(x_2, R=11|v)} \\
&= \kappa_1 \cdot \frac{P(R=00|v)/P(R=01|v)}{\frac{P(R=11|x_2, v)}{P(R=01|x_2, v)} \cdot \frac{P(R=10|x_1, v)}{P(R=11|x_1, v)} + \frac{P(R=11|x_2, v)}{P(R=01|x_2, v)} + 1},
\end{aligned}$$

which is identifiable.

A similar argument for the second term yields

$$\kappa_2 \cdot \frac{P(R=00|v)/P(R=10|v)}{\frac{P(R=11|x_1, v)}{P(R=10|x_1, v)} \cdot \frac{P(R=01|x_2, v)}{P(R=11|x_2, v)} + \frac{P(R=11|x_1, v)}{P(R=10|x_1, v)} + 1}.$$

The third term yields

$$\begin{aligned}
&\kappa_3 \cdot \frac{p(x_1, x_2|R=11, v)P(R=00|v)}{p(x_1, x_2, R \in \{01, 10, 11\}|v)} \\
&= \kappa_3 \cdot \frac{P(R=00|v)/P(R=11|v)}{\frac{P(R=10|x_1, v)}{P(R=11|x_1, v)} + \frac{P(R=01|x_2, v)}{P(R=11|x_2, v)} + 1},
\end{aligned}$$

which is identifiable.

Thus, this implies that $P(R=00|x, v)/P(R=11|x, v)$ is identifiable. We can model the odds $P(R=r|x_r, v)/P(R=11|x_r, v)$, and then, plug them into the formulas above to obtain

an identifiable expression for $P(R = 00|x, v)/P(R = 11|x, v)$.

A.3.2 RA estimator (imputation)

First, as the identifying restriction is expressed in terms of the extrapolation distribution, the regression adjustment estimator is fairly straightforward. Since the extrapolation distribution $p(x_1, x_2|R = 00)$ is a mixture of three identifiable distributions, we can again utilize a Monte Carlo approximation to construct the regression adjustment estimator. This approximation is essentially the same as under CCMV except we sample from this mixture distribution for pattern $R = 00$.

A.4 Sensitivity Analysis

We present details on the sensitivity analyses utilized in the paper.

A.4.1 Sensitivity analysis for the odds under logistic regression

We have showed with Proposition 2.2 that under CCMV, the odds satisfies

$$O_{r,a}(X, W_a) := \frac{P(R = r|X, W_a, a)}{P(R = 1_d|X, W_a, a)} \stackrel{\text{CCMV}}{=} \frac{P(R = r|X_r, W_a, a)}{P(R = 1_d|X_r, W_a, a)} =: Q_{r,a}(X_r, W_a).$$

Thus, the complete odds depend only on the variables observed under pattern $R = r$. We can introduce a sensitivity parameter to perturb this CCMV assumption using an exponential tilting method with logistic regression (Kim and Yu, 2011; Shao and Wang, 2016; Zhao et al., 2017). This results in

$$O_{r,a}(X, W_a) = Q_{r,a}(X_r, W_a)e^{\rho_{\bar{r}}X_{\bar{r}}}, \quad (2.21)$$

where $\rho_{\bar{r}}$ is a sensitivity parameter that consists of a vector of real numbers (one for each missing covariate under pattern r). In practice, we often let every element of $\rho_{\bar{r}}$ to be the

same for simplicity.

We now consider the special case of logistic regression. Under CCMV with an assumed logistic regression model, the complete odds satisfies

$$\log O_{r,a}(X, W_a; \alpha_{r,a}) = \log Q_{r,a}(X_r, W_a; \alpha_{r,a}) = \alpha_{r,a}^\top X_r.$$

Under the sensitivity model (equation (2.21)), this becomes

$$\log O_{r,a}(X, W_a; \alpha_{r,a}, \rho_{\bar{r}}) = \alpha_{r,a}^\top X_r + \rho_{\bar{r}}^\top X_{\bar{r}}.$$

Note that the first term on the RHS is the traditional term obtained by fitting a logistic regression model, and the second term is the tilting factor to push the missing data assumption away from CCMV.

Varying the elements of $\rho_{\bar{r}}$ will modify the missing data assumption and change the weights. This allows one to determine the robustness of the estimator to changes in this sensitivity parameter. This sensitivity parameter has a nice interpretation because it is the coefficient in a linear model, and the range can be specified based on one's belief about the relation of impact of the missing variables to the impact of the observed variables. For example, if the coefficient of the X_1 variable is 1 and if X_2 is subject to missingness, then the sensitivity parameter corresponding to X_2 can be chosen to be within the interval $[-1, 1]$ by assuming that X_2 has at most the same impact in magnitude as X_1 .

A.4.2 Sensitivity analysis for binary imputation via rejection sampling

We also present an approach for sensitivity analysis under imputation model. We first consider the case where the variable to be imputed X is binary. Suppose that under an identifying restriction (such as the CCMV), we will impute the variable X by drawing $X \sim$

Binomial(p), where p is the probability parameter from a particular imputation model.

To perturb the imputation model, we consider the following procedure:

- Draw $X \sim \text{Bernoulli}(p)$.
- Consider a sensitivity parameter $\zeta \in [0, 1]$. We iterate the following steps until a draw is accepted:
 - If $X = 1$, with probability ζ , redraw $X \sim \text{Bernoulli}(p)$. Otherwise, accept the draw.
 - If $X = 0$, with probability $1 - \zeta$, redraw $X \sim \text{Bernoulli}(p)$. Otherwise, accept the draw.

We denote this new random variable by Y . We can see when $\zeta = 1$, we force Y to be deterministically 0, and when $\zeta = 0$, we force Y to be deterministically 1. When $\zeta = 1/2$, we resample exactly half of the time, so we obtain $Y = X$. Namely, when $\zeta = 1/2$, we are under the unperturbed model. Every other value of ζ interpolates between these extremes. Thus, the parameter ζ can be viewed as a preference toward $X = 0$.

It turns out that the above procedure admits a simpler expression. Given that the final output is a binary number, the accepted value Y follows from a Bernoulli random variable. Since Y is binary, $\mathbb{E}[Y]$ is the probability of success. On the initial iteration, we draw 1 and accept it with probability $p(1 - \zeta)$. Otherwise, we redraw with probability $p\zeta + (1 - p)(1 - \zeta)$ and return to the same state. This yields the following equation

$$\mathbb{E}[Y] = p(1 - \zeta) + p\zeta\mathbb{E}[Y] + (1 - p)(1 - \zeta)\mathbb{E}[Y],$$

so we have
$$\mathbb{E}[Y] = \frac{p(1 - \zeta)}{1 - p\zeta - (1 - p)(1 - \zeta)} = \frac{p(1 - \zeta)}{p + \zeta - 2p\zeta}.$$

Thus, we actually do not need to perform the rejection sampling procedure. Instead, we can

simply draw

$$Y \sim \text{Bernoulli}\left(\frac{p(1-\zeta)}{p+\zeta-2p\zeta}\right).$$

A.4.3 Sensitivity analysis for continuous imputation via Gaussian tilting

Now, we discuss the case of sensitivity analysis on imputing a continuous random variable. Suppose that without perturbations, we impute a continuous random variable X from a PDF $p(x)$. We propose the following modified approach:

- Draw $X \sim p(x)$.
- Consider a sensitivity parameter $\xi \in [0, \infty)$. First, we sample $X \sim p(x)$. We iterate the following steps until a draw is accepted:
 - With probability $\exp(-\xi X^2)$, accept $X \sim p(x)$.
 - Otherwise, redraw X .

First, note that as $\xi \rightarrow \infty$, we almost always resample, and as $\xi \rightarrow 0$, we almost never resample. We denote this new random variable by Y . The distribution of Y is given by

$$P(Y \leq y) = P(X \leq y \mid \text{accept } X) = P(X \leq y \mid U \leq \exp(-\xi X^2))$$

for a random variable $U \sim \text{Uniform}(0, 1)$ independent of X . We see that

$$\begin{aligned} P(Y \leq y) &= P(X \leq y \mid U \leq \exp(-\xi X^2)) \\ &\propto P(X \leq y, U \leq \exp(-\xi X^2)) \\ &= P(U \leq \exp(-\xi X^2) \mid X \leq y) \cdot P(X \leq y) \\ &= \int_{-\infty}^y \frac{P(U \leq \exp(-\xi x^2) \mid x)}{P(X \leq y)} \cdot p(x) \, dx \cdot P(X \leq y) \\ &= \int_{-\infty}^y P(U \leq \exp(-\xi x^2) \mid x) \cdot p(x) \, dx \end{aligned}$$

$$= \int_{-\infty}^y \exp(-\xi x^2) p(x) dx.$$

Therefore, we see that this resampling method is equivalent to multiplying the original PDF $p(x)$ by $\exp(-\xi x^2)$ and renormalizing. Since the multiplication factor $\exp(-\xi x^2)$ is an exponential, this action can be considered a form of *Gaussian tilting*.

$p(x)$ is Gaussian. Suppose that the original imputation distribution is Gaussian such that $X \sim N(\mu, \sigma^2)$. A simple completing the square argument shows that

$$-\frac{1}{2\sigma^2}(x - \mu)^2 - \xi x^2 = -\left(\frac{1}{2\sigma^2} + \xi\right) \left(x - \frac{\mu}{2\sigma^2\xi + 1}\right)^2 + \left(\frac{\mu^2}{2\sigma^2 + 4\sigma^4\xi} - \frac{\mu^2}{2\sigma^2}\right).$$

Therefore, Gaussian tilting by $\exp(-\xi x^2)$ results in a random variable

$$Y \sim N\left(\frac{\mu}{2\sigma^2\xi + 1}, \frac{\sigma^2}{2\sigma^2\xi + 1}\right).$$

As $\xi \rightarrow \infty$, the distribution for Y degenerates to a point mass of 1 at 0. As $\xi \rightarrow 0$, we approach the original distribution $p(x)$, which is what we expect as we almost never resample.

$p(x)$ is Exponential. Suppose $p(x)$ is Exponential(λ) such that $p(x) = \lambda \exp(-\lambda x)$. Then, Gaussian tilting by $\exp(-\xi x^2)$ results in a truncated Gaussian that is formed by taking the PDF over the positive support of a Gaussian with mean $-\lambda/(2\xi)$ and variance $1/\xi$. As $\xi \rightarrow \infty$, again, the distribution for Y degenerates to a point mass of 1 at 0.

A.5 Details on Survival Analysis Simulation and additional simulation studies

A.5.1 Details on Cox model simulation

Here we provide details on the Cox model simulation. Note that the data generating process is described in the main paper.

Estimation of the imputation model

In our simulated Cox model problem, there are two covariates $X = (X_1, X_2) \in \{0, 1\}^2$. In the complete case, the observed-data distribution can be decomposed into the following terms:

$$p(x, R = 11, y, \delta) = \underbrace{p(x, R = 11)}_{\text{Empirical}} \cdot \underbrace{\frac{\overbrace{P(R = 11|x, y, \delta)}^{\text{Propensity score}}}{P(R = 11|x)}}_{\text{Integration}} \cdot \underbrace{p(y, \delta|x)}_{\text{EM}}.$$

Each of the above four components in the RHS is estimated as follows:

- **Empirical.** This component can be directly estimated from the data using empirical proportion.
- **Propensity score.** This is estimated by combining estimators of the complete odds and use equation (2.5).
- **Integration.** The term $P(R = 11|x)$ can be rewritten as follows

$$P(R = 11|x) = \sum_{\delta=0}^1 \int_0^\infty P(R = 11|x, y, \delta) p(y, \delta|x) dy.$$

In practice, this integral does not have an analytic form, so we use numerical integration in R to circumvent the problem.

- **EM.** The estimation of this component is more involved. We use an EM-like method to estimate it. We assume a parametric model $p_{Y,\Delta|X}(y, \delta|x) = [\nu_1 \exp(x^\top \beta)]^\delta \nu_2^{1-\delta} e^{-y(\nu_1 \exp(x^\top \beta) + \nu_2)}$, this model is derived from our design of the simulation. Fitting $p(y, \delta|x)$ is a simple maximum likelihood procedure when the covariates are fully observed. Consider a specific $\tilde{x}$. The model becomes

$$p_{Y,\Delta|X}(y, \delta|\tilde{x}) = [\nu_1 \exp(\tilde{x}^\top \beta)]^\delta \nu_2^{1-\delta} e^{-y(\nu_1 \exp(\tilde{x}^\top \beta) + \nu_2)}$$

$$= \gamma_{\tilde{x}}^{\delta} \nu_2^{1-\delta} e^{-y(\gamma_{\tilde{x}} + \nu_2)}$$

for $\gamma_x = \nu_1 \exp(x^\top \beta)$. The ML estimators are

$$\hat{\gamma}_{\tilde{x}} = \frac{\sum_{i=1}^n \delta_i 1(x_i = \tilde{x})}{\sum_{i=1}^n y_i 1(x_i = \tilde{x})} \quad \text{and} \quad \hat{\nu}_2 = \frac{\sum_{i=1}^n (1(x_i = \tilde{x}) - \delta_i 1(x_i = \tilde{x}))}{\sum_{i=1}^n y_i 1(x_i = \tilde{x})}.$$

In practice, we do not know the covariates for all of the individuals. So, we fit the models using an EM algorithm. We replace the indicators $1(x_i = \tilde{x})$ with weights w_i .

Suppose $\tilde{x} = (1, 1)$. The weights are as follows:

$$w_i = \begin{cases} 1 & \text{if } x_1 = x_2 = 1 \\ 0 & \text{if } x_1 = 0 \text{ or } x_2 = 0 \\ \hat{P}(X_2 = 1 | X_1 = 1, R = 11, y, \delta) & \text{if } x_1 = 1 \text{ and } x_2 \text{ missing} \\ \hat{P}(X_1 = 1 | X_2 = 1, R = 11, y, \delta) & \text{if } x_2 = 1 \text{ and } x_1 \text{ missing} \end{cases}.$$

The weights depend on the fitted parameters. For the EM algorithm, we iterate between updating w_i (E-step) and computing the MLEs (M-step).

With the above procedure, we are able to estimate the distribution $p(x, R = 11, y, \delta)$, which describes all imputation models under the CCMV assumption.

Sensitivity analysis

We include the following sensitivity analysis as part of our simulations. We fix a sample size of $n = 1000$ with $M = 50$ imputations. For the IPW approach, we vary the exponential tilting sensitivity parameter ρ in the interval $[-3, 3]$. For the imputation/RA approach, we change the binary sensitivity analysis parameter ζ in the interval $[0, 1]$. For each corresponding

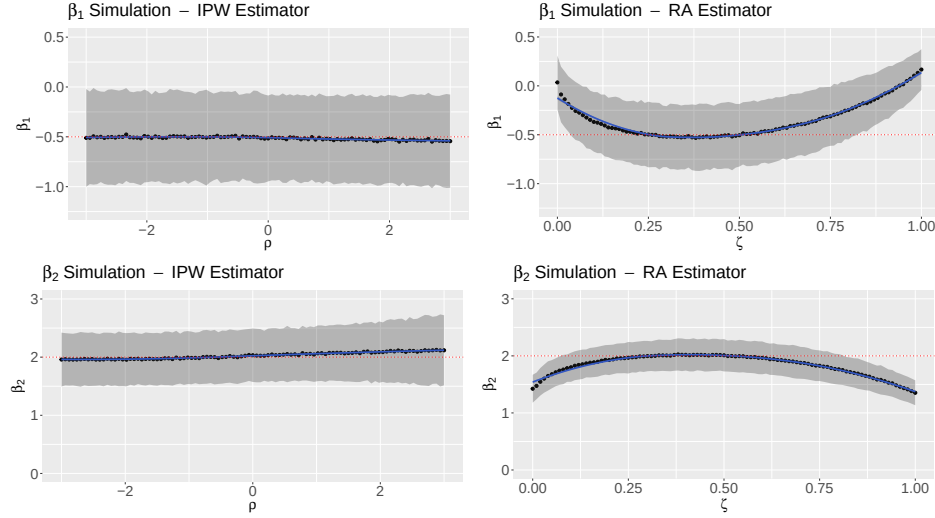


Figure 2.7: Sensitivity analysis for the Cox model simulation with both the IPW and RA estimators.

value of the sensitivity parameter, we compute the IPW and RA estimators, respectively. The true value of the parameter is indicated with the red line.

First, we observe that the IPW estimator appears to be fairly stable as we vary the sensitivity parameter, and the size of the confidence interval does not change drastically. Therefore, we conclude that the β_1 parameter is significant even as we perturb the estimator. On the other hand, the point estimates of the RA estimator do appear to change slightly as we vary the imputation model. The β_2 parameter is large enough that any of these perturbations do not result in a change in conclusion. However, the β_1 parameter is relatively close in magnitude to 0, so for extreme values of the ζ parameter, the results are not significant. This suggests that the choice of identifying restriction is critical, and a sensitivity analysis is crucial for a sound data analysis. We note that the extreme values of the ζ parameter generally imply near almost sure deterministic imputation for both covariates, which generally will be unlikely to be true.

A.5.2 Details on binary treatment simulation

Data generating process

We generate the data as follows:

1. We sample (R, Y, A) as discrete random variables as follows: 2

- $P(R = 11, Y = 0, A = 0) = 1/16$
- $P(R = 11, Y = 0, A = 1) = 1/12$
- $P(R = 11, Y = 1, A = 0) = 1/16$
- $P(R = 11, Y = 1, A = 1) = 1/24$
- $P(R = 10, Y = 0, A = 0) = 1/12$
- $P(R = 10, Y = 0, A = 1) = 1/8$
- $P(R = 10, Y = 1, A = 0) = 1/24$
- $P(R = 10, Y = 1, A = 1) = 1/12$
- $P(R = 01, Y = 0, A = 0) = 1/16$
- $P(R = 01, Y = 0, A = 1) = 1/24$
- $P(R = 01, Y = 1, A = 0) = 1/16$
- $P(R = 01, Y = 1, A = 1) = 1/12$
- $P(R = 00, Y = 0, A = 0) = 1/36$
- $P(R = 00, Y = 0, A = 1) = 1/24$
- $P(R = 00, Y = 1, A = 0) = 1/24$
- $P(R = 00, Y = 1, A = 1) = 1/18$

2. We sample $(X_1, X_2)|R, Y, A$ via the following distributions:

- $X_1, X_2 | R = 11, Y = 0, A = 0 \sim N \left(\begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 11, Y = 0, A = 1 \sim N \left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 11, Y = 1, A = 0 \sim N \left(\begin{bmatrix} 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \right)$
- $X_1, X_2 | R = 11, Y = 1, A = 1 \sim N \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 10, Y = 0, A = 0 \sim N \left(\begin{bmatrix} 2 \\ 19/5 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 10, Y = 0, A = 1 \sim N \left(\begin{bmatrix} 2 \\ 16/5 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 10, Y = 1, A = 0 \sim N \left(\begin{bmatrix} 1 \\ 3/2 \end{bmatrix}, \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \right)$
- $X_1, X_2 | R = 10, Y = 1, A = 1 \sim N \left(\begin{bmatrix} 3 \\ 11/5 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 01, Y = 0, A = 0 \sim N \left(\begin{bmatrix} 13/5 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 01, Y = 0, A = 1 \sim N \left(\begin{bmatrix} 14/5 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{bmatrix} \right)$

- $X_1, X_2 | R = 01, Y = 1, A = 0 \sim N \left(\begin{bmatrix} 3.125 \\ 2.5 \end{bmatrix}, \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \right)$
- $X_1, X_2 | R = 01, Y = 1, A = 1 \sim N \left(\begin{bmatrix} 1.9 \\ 1.5 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 00, Y = 0, A = 0 \sim N \left(\begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 00, Y = 0, A = 1 \sim N \left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{bmatrix} \right)$
- $X_1, X_2 | R = 00, Y = 1, A = 0 \sim N \left(\begin{bmatrix} 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \right)$
- $X_1, X_2 | R = 00, Y = 1, A = 1 \sim N \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix} \right)$

The generative model for $(X_1, X_2) | R, Y, A$ is constructed in such a way to ensure that the CCMV assumption holds.

Details on the estimation

Note that in the binary treatment problem, we have two estimators; one estimator to account for the missing covariates (denoted as IPW-R or RA-R) and the other estimator for the ATE (denoted as IPW-A, RA-A, DR-A).

Final estimators. For ease of simulation, we can consider the following 6 estimators for $\mathbb{E}[Y(1)]$. Note that here we describe the linear functional form of the estimator. We take the empirical average of them to form the final estimate.

1. *IPW-R, IPW-A.*

$$\frac{1(R = 11)1(A = 1)Y}{P(A = 1|X)} \sum_r Q_{r,1}(X_r, Y)$$

2. *IPW-R, RA-A.*

$$\frac{1(R = 11)m_1(X)}{P(A = 1|X)} \sum_r \sum_a 1(A = a)Q_{r,a}(X_r, Y)$$

3. *IPW-R, DR-A.*

$$\begin{aligned} & \frac{1(R = 11)1(A = 1)(Y - m_1(X))}{P(A = 1|X)} \sum_r Q_{r,1}(X_r, Y) \\ & + \frac{1(R = 11)m_1(X)}{P(A = 1|X)} \sum_r \sum_a 1(A = a)Q_{r,a}(X_r, Y) \end{aligned}$$

4. *RA-R, IPW-A.*

$$1(A = 1)Y \sum_r 1(R = r)M_{1/\pi_A, r, 1}(X_r, Y)$$

5. *RA-R, RA-A.*

$$\sum_r \sum_a 1(R = r)1(A = a)M_{m, r, a}(X_r, Y)$$

6. *RA-R, DR-A.*

$$\begin{aligned} & 1(A = 1)Y \sum_r 1(R = r)M_{m, r, 1}(X_r, Y) \\ & + \sum_r \sum_a 1(R = r)1(A = a)M_{m, r, a}(X_r, Y) \\ & - 1(A = 1) \sum_r 1(R = r)M_{m/\pi_A, r, 1}(X_r, Y) \end{aligned}$$

As a reminder, the notation m_1 , $M_{m, r, 0}$, $M_{m, r, 1}$, $M_{1/\pi_A, r, 1}$, and $M_{m/\pi_A, r, 1}$ refer to the

following regression functions:

$$\begin{aligned}
m_1(X) &= \mathbb{E}[Y|A = 1, X], \\
M_{m,r,0}(X_r) &= \mathbb{E}[m_1(X)|X_r, R = r, A = 0], \\
M_{m,r,1}(X_r, Y) &= \mathbb{E}[m_1(X)|X_r, R = r, Y, A = 1], \\
M_{1/\pi_A,r,1}(X_r, Y) &= \mathbb{E}\left[\frac{1}{\pi_A(X)} \middle| X_r, R = r, Y, A = 1\right], \\
M_{m/\pi_A,r,1}(X_r, Y) &= \mathbb{E}\left[\frac{m_1(X)}{\pi_A(X)} \middle| X_r, R = r, Y, A = 1\right].
\end{aligned}$$

Similar estimators can be constructed for $\mathbb{E}[Y(0)]$. Merging estimators for $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ together, we can obtain 6 estimators similar to the ones above for the ATE. We choose not to construct the multiply-robust estimator for simplicity. As seen in Theorem 2.3, there will be numerous terms, so the multiply-robust estimator will generally be considered a theoretical construct.

Here we describe how to estimate the components in the above estimators.

IPW-R component. The above construction implies that the complete odds has the following form:

$$\begin{aligned}
Q_{r,a}(X_r, Y) &= \frac{P(R = r|X_r, Y, A)}{P(R = 11|X_r, Y, A)} \\
&= \frac{p(x_r|r, y, a)p(r, y, a)}{p(x_r|11, y, a)p(11, y, a)}.
\end{aligned}$$

In the special case where covariances are assumed to be the same for a fixed y and a , one may simplify the complete odds accordingly

$$Q_{10,a}(x_1, y) = \frac{p(x_1|10, y, a)p(r, y, a)}{p(x_1|11, y, a)p(11, y, a)}$$

$$\begin{aligned}
&= \frac{\exp(-(x_1 - \mu_{1,10})^2/(2\sigma^2)) \cdot p(10, y, a)}{\exp(-(x_1 - \mu_{1,11})^2/(2\sigma^2)) \cdot p(11, y, a)} \\
&= \exp\left(\frac{\mu_{1,10} - \mu_{1,11}}{\sigma^2}x_1 + \frac{\mu_{1,11}^2 - \mu_{1,10}^2}{2\sigma^2}\right) \cdot \frac{p(10, y, a)}{p(11, y, a)},
\end{aligned}$$

$$\begin{aligned}
Q_{01,a}(x_2, y) &= \frac{p(x_2|01, y, a)p(r, y, a)}{p(x_2|11, y, a)p(11, y, a)} \\
&= \frac{\exp(-(x_2 - \mu_{2,01})^2/(2\sigma^2)) \cdot p(01, y, a)}{\exp(-(x_2 - \mu_{2,11})^2/(2\sigma^2)) \cdot p(11, y, a)} \\
&= \exp\left(\frac{\mu_{2,01} - \mu_{2,11}}{\sigma^2}x_2 + \frac{\mu_{2,11}^2 - \mu_{2,01}^2}{2\sigma^2}\right) \cdot \frac{p(01, y, a)}{p(11, y, a)}, \\
Q_{00,a}(y) &= \frac{p(00, y, a)}{p(11, y, a)}.
\end{aligned}$$

The probabilities $p(r, y, a)$ can be estimated by the empirical frequencies, and the means and variances are estimated using the sample versions. Finally, the propensity score of $R = 11$ is

$$\pi_R(x, y, a) = P(R = 11|x, y, a) = \frac{1}{1 + Q_{10,a}(x_1, y) + Q_{01,a}(x_1, y) + Q_{00,a}(y)}.$$

RA-R component. We estimate the regression functions $M_{m,r,0}$, $M_{m,r,1}$, $M_{1/\pi_A,r,1}$, and $M_{m/\pi_A,r,1}$ using a Monte Carlo approach, described in Section 2.3.2. Specifically, in the simulation procedure, we make the assumption that the distribution $p(x_1, x_2|r, y, a)$ is a multivariate Gaussian and that the missing mechanism for the covariates satisfies the CCMV assumption. This implies that the conditional distributions $p(x_1|x_2, r, y, a)$ and $p(x_2|x_1, r, y, a)$ are conditional Gaussians. Therefore, we first can estimate the distributions $p(x_1|R = 10, r, y, a)$, $p(x_2|R = 01, r, y, a)$, and $p(x_1, x_2|R = 11, r, y, a)$ with a maximum likelihood procedure directly (plug in the sample means and variances/covariances as parameters). Then, under the CCMV assumption, we can construct the estimates of the distribution $p(x_1, x_2|r, y, a)$ in the

following way:

$$\begin{aligned}\widehat{p}(x_1, x_2 | R = 00, y, a) &= \widehat{p}(x_1, x_2 | R = 11, y, a), \\ \widehat{p}(x_1, x_2 | R = 01, y, a) &= \widehat{p}(x_1 | x_2, R = 11, y, a) \widehat{p}(x_2 | R = 01, y, a), \\ \widehat{p}(x_1, x_2 | R = 10, y, a) &= \widehat{p}(x_2 | x_1, R = 11, y, a) \widehat{p}(x_1 | R = 10, y, a).\end{aligned}$$

This further implies that we can use the estimated distributions

$$\begin{aligned}\widehat{p}(x_1, x_2 | R = 00, y, a) &= \widehat{p}(x_1, x_2 | R = 11, y, a), \\ \widehat{p}(x_1 | x_2, R = 01, y, a) &= \widehat{p}(x_1 | x_2, R = 11, y, a), \\ \widehat{p}(x_2 | x_1, R = 10, y, a) &= \widehat{p}(x_2 | x_1, R = 11, y, a)\end{aligned}$$

to complete our data. We use the above extrapolation densities as an estimated imputation model. We then sample the missing covariates $M = 50$ times to generate multiple complete data sets $\widetilde{D}^{(j)} = \{(\widetilde{X}_i^{(j)}, R_i, Y_i, A_i)\}_{i=1}^n$, where $\widetilde{X}_i^{(j)} = (X_{i,R_i}, \widetilde{X}_{i,R_i}^{(j)})$ refers to the j th imputed missing covariates appended to the observed covariates corresponding to the i th observation.

We pool all $\widetilde{D}^{(j)}$ data sets together to form one large data set $\check{D}$. This enables us to take the empirical average of the RA- R estimators directly using all of the completed data $\check{D}$. For example, for the (RA- R , IPW- A) estimator, we have

$$\begin{aligned}& \frac{1}{n} \sum_{i=1}^n \left[1(A_i = 1) Y_i \sum_r 1(R_i = r) \widehat{M}_{1/\pi_{A,r,1}}(X_{R_i,i}, Y_i) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \left[1(A_i = 1) Y_i \cdot \frac{1}{M} \sum_{j=1}^M \frac{1}{\widehat{\pi}_A(\widetilde{X}_i^{(j)})} \right] \\ &= \frac{1}{Mn} \sum_{i=1}^n \sum_{j=1}^M \left[1(A_i = 1) Y_i \cdot \frac{1}{\widehat{\pi}_A(\widetilde{X}_i^{(j)})} \right].\end{aligned}$$

Similar logic implies that

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \left[\sum_r \sum_a 1(R_i = r) 1(A_i = a) \widehat{M}_{m,r,a}(X_{R_i,i}, Y_i) \right] \\
&= \frac{1}{n} \sum_{i=1}^n \left[\sum_a 1(A_i = a) \frac{1}{M} \sum_{j=1}^M \widehat{m}_1(\widetilde{X}_i^{(j)}) \right] \\
&= \frac{1}{Mn} \sum_{i=1}^n \sum_{j=1}^M \widehat{m}_1(\widetilde{X}_i^{(j)}), \\
& \frac{1}{n} \sum_{i=1}^n \left[1(A_i = 1) \sum_r 1(R_i = r) \widehat{M}_{m/\pi_A,r,a}(X_{R_i,i}, Y_i) \right] \\
&= \frac{1}{n} \sum_{i=1}^n \left[1(A_i = 1) \frac{1}{M} \sum_{j=1}^M \frac{\widehat{m}_1(\widetilde{X}_i^{(j)})}{\widehat{\pi}_A(\widetilde{X}_i^{(j)})} \right] \\
&= \frac{1}{Mn} \sum_{i=1}^n \sum_{j=1}^M 1(A_i = 1) \frac{\widehat{m}_1(\widetilde{X}_i^{(j)})}{\widehat{\pi}_A(\widetilde{X}_i^{(j)})}.
\end{aligned}$$

With these equations, we can construct the (RA-R, RA-A) and (RA-R, DR-A) estimators.

IPW-A component. The propensity score can be calculated via

$$\pi_A(x) = P(A = 1|x_1, x_2) = \frac{\sum_r \sum_{y=0}^1 p(x_1, x_2|r, y, A = 1)p(r, y, A = 1)}{\sum_r \sum_{y=0}^1 \sum_{a=0}^1 p(x_1, x_2|r, y, a)p(r, y, a)}.$$

We describe how to estimate the distribution $\widehat{p}(x_1, x_2|r, y, a)$ in the previous discussion on the **RA-R component**.

RA-A component. Since Y is binary, the regression function can be calculated via

$$\begin{aligned}
m_a(x) &= \mathbb{E}[Y|A = a, x_1, x_2] = P(Y = 1|A = a, x_1, x_2) \\
&= \frac{\sum_r p(x_1, x_2|r, Y = 1, A = a)p(r, Y = 1, A = a)}{\sum_r \sum_{y=0}^1 p(x_1, x_2|r, y, a)p(r, y, a)}.
\end{aligned}$$

We describe how to estimate the distribution $\hat{p}(x_1, x_2 | r, y, a)$ in the previous discussion on the **RA-R component**.

Sensitivity analysis

We perform a sensitivity analysis for this simulation by varying the sensitivity parameters $\rho \in [-2, 0.5]$ and $\xi \in [0, 4]$ based on Appendices A.4.1 and A.4.3. For each sensitivity parameter, we generate 1000 data sets with $n = 1000$ and perturb the 6 estimators. We use $M = 50$ imputations for the RA estimator. In keeping with what was done in the primary analysis, the confidence intervals for the IPW- R estimators are constructed by pooling the results across all 1000 iterations and then taking the 0.025% and 97.5% quantiles. This was done to reduce the effect of outliers (created by unstable inverse probability weights) on the standard error estimate. The confidence intervals for the RA- R estimators are generated by calculating the standard error across all 1000 iterations. The true ATE is plotted in red.

First, we observe that when ξ reaches 4, we are very close to deterministic imputation, so often in most practical applications, ξ does not need to belong to a large interval. The estimator where we use RA in both stages (RA- R , RA- A) performs the best and appears fairly robust to changes in ξ . The (RA- R , IPW- A) estimator also appears to have minimal bias as we vary ξ , but the variance increases. The IPW estimators are fairly sensitive to changes in ρ , and in particular, the variance can severely increase for moderate changes in ρ . This sensitivity analysis suggests that RA tends to provide the most stable estimates when the assumptions are perturbed. Intuitively, this makes sense because a perturbation in the missing data assumption changes the complete odds, and therefore, can drastically inflate the inverse probability weights in some scenarios.

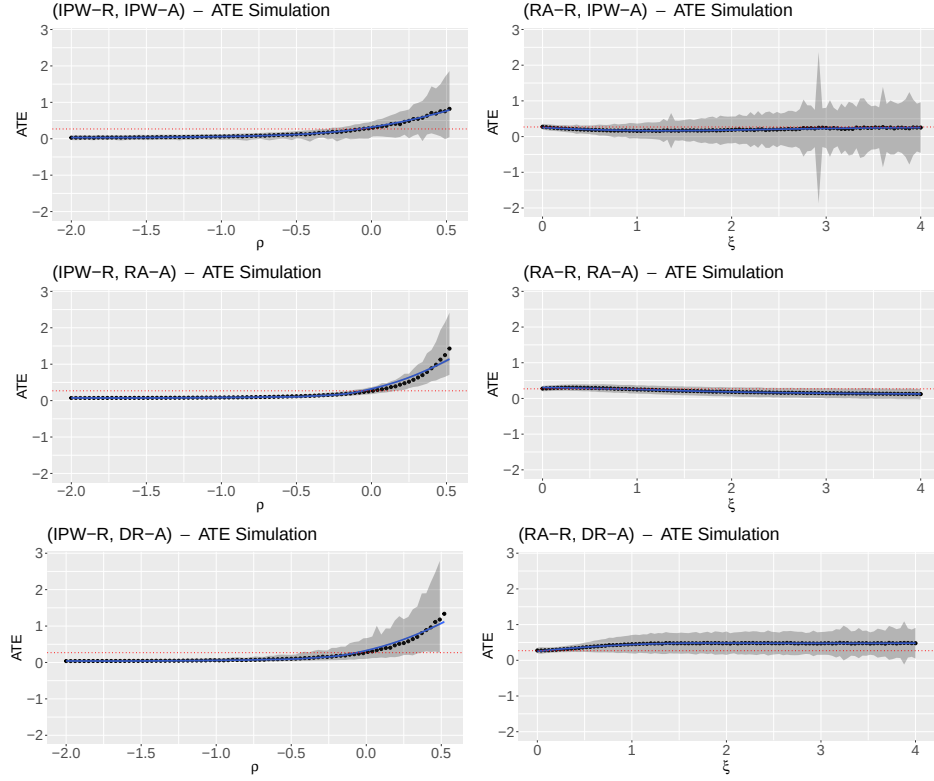


Figure 2.8: Sensitivity analysis for the ATE simulation with both the IPW and RA estimators.

A.6 Additional Analysis of the NACC Data

We performed sensitivity analysis under the IPW and RA settings by perturbing both the IPW and RA estimators with sensitivity parameters $\rho \in [-1, 1]$ and $\zeta \in [0, 1]$ (since the covariates in the NACC data are binary). The details for sensitivity analysis are outlined in Appendices A.4.1 and A.4.2.

We include the results of all sensitivity analysis of the depression and anxiety in this section. Figures 2.9 and 2.10 show that for three age groups (60 – 64, 65 – 69, and 70 – 74), the point estimates given by the IPW and RA estimators are generally fairly stable even as we vary the sensitivity parameters. Moreover, the confidence intervals for the age groups

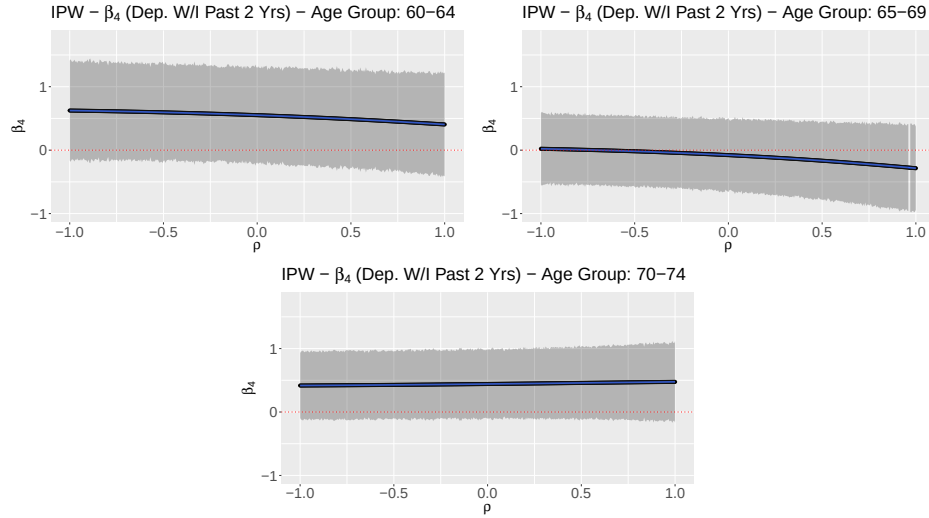


Figure 2.9: Sensitivity analysis results for the IPW estimator for β_4 .

60 – 64 and 70 – 74 for both the IPW and RA estimators either exhibit significance or have endpoints close to 0 across much of the sensitivity parameters' ranges. This preliminary analysis suggests that at least for some age groups, recent depression is associated with shorter time to dementia. There is evidence in the literature to support this claim as well (Kitching, 2015).

Figures 2.11 and 2.12 display the results for anxiety. We see across all age groups, as we vary the respective sensitive parameters ρ and ζ , the point estimates for β_5 are all close to 0 with all of the confidence intervals covering 0. This suggests some evidence that the self-reported anxiety is not associated with shorter time to dementia.

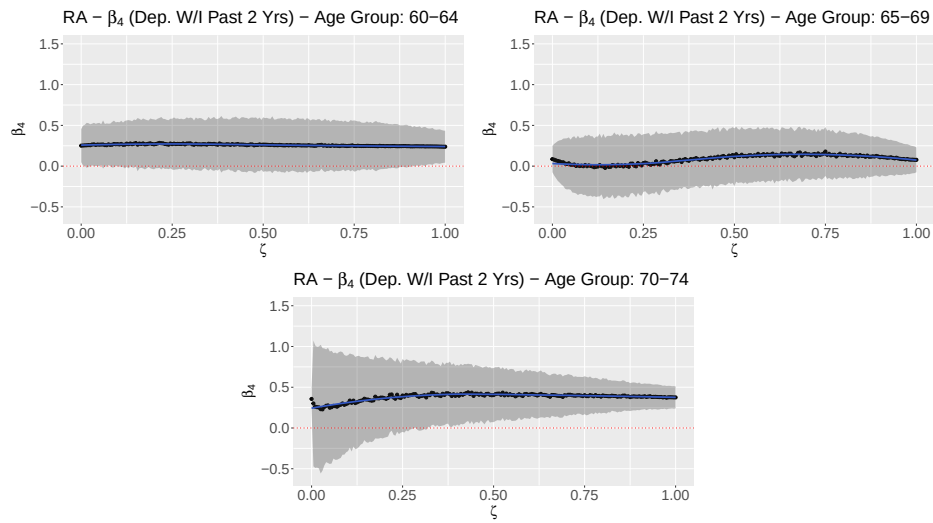


Figure 2.10: Sensitivity analysis results for the RA estimator for β_4 .

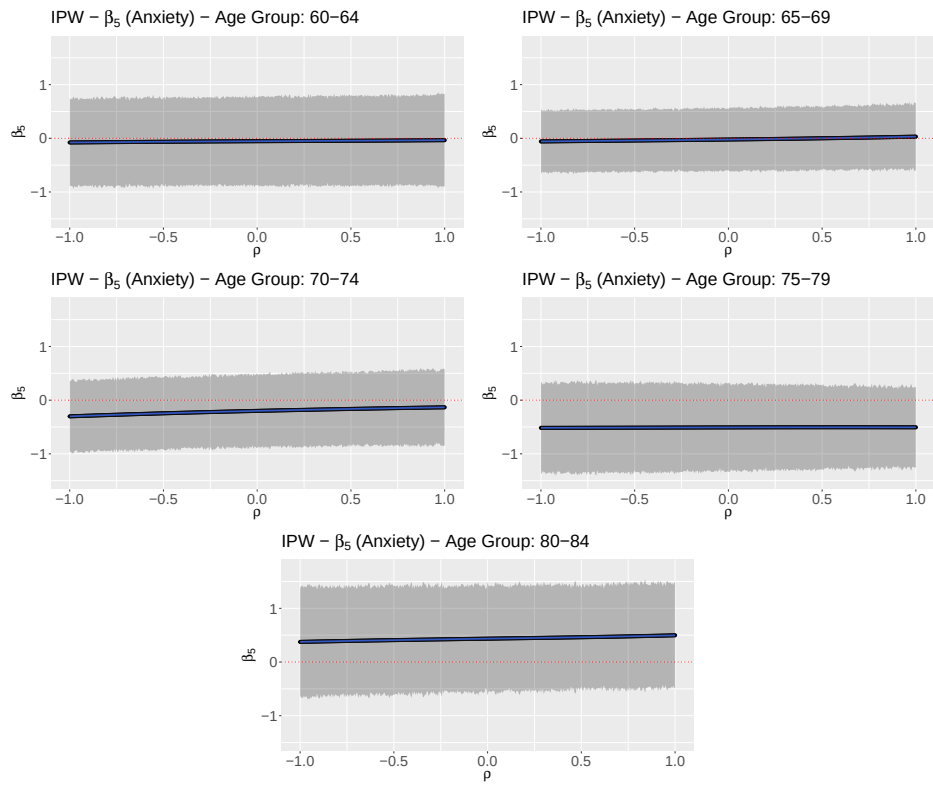


Figure 2.11: Sensitivity analysis results for the IPW estimator for β_5 .

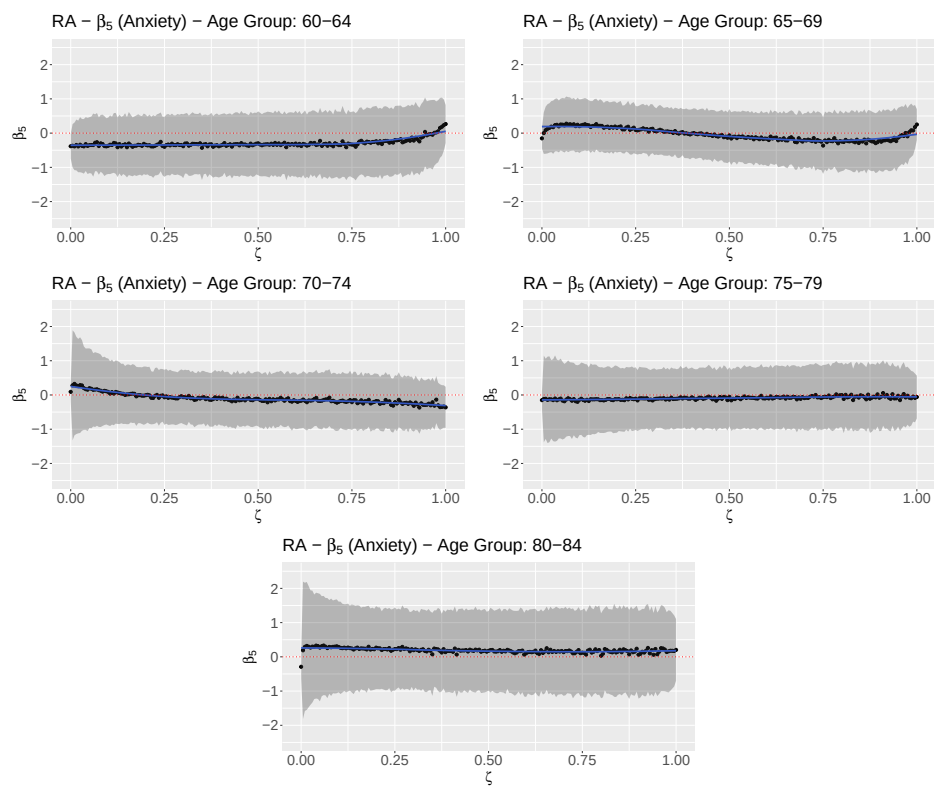


Figure 2.12: Sensitivity analysis results for the RA estimator for β_5 .

A.7 Proofs

The proofs of the results in the paper are provided here. We separate the proofs into three sections by topic: 1. estimator construction and identification, 2. multiple robustness and efficiency theory, and 3. asymptotic theory.

A.7.1 Estimator construction and identification

This section contains proofs of Lemma 2.1, Proposition 2.1, Proposition 2.2, Theorem 2.1, Lemma 2.2.

of Lemma 2.1. Let the observed outcome pattern a be fixed. We have

$$\begin{aligned}
& \mathbb{E}[f_a(X, W_a)1(A = a)1(R = r)] \\
&= \int f_a(x, w_a)1(a' = a)1(r' = r) p(x, r', w_a, a') dx dr' dw_a da' \\
&= \int f_a(x, w_a)1(a' = a) p(x, r, w_a, a') dx dw_a da' \\
&= \int f_a(x, w_a)1(a' = a) \frac{p(x, r, w_a, a')}{p(x, 1_d, w_a, a')} p(x, 1_d, w_a, a') dx dw_a da' \\
&= \int f_a(x, w_a)1(a' = a) \frac{P(R = r|x, w_a, a')}{P(R = 1_d|x, w_a, a')} \\
&\quad \times 1(s = 1_d) p(x, s, w_a, a') dx ds dw_a da' \\
&= \mathbb{E}[f_a(X, W_a)1(A = a)O_{r,a}(X, W_a)1(R = 1_d)],
\end{aligned}$$

as desired.

Finally, note that summing over all patterns r yields the second decomposition

$$\mathbb{E}[f_a(X, W_a)1(A = a)] = \sum_r \mathbb{E}[f_a(X, W_a)1(A = a)O_{r,a}(X, W_a)1(R = 1_d)].$$

□

of Proposition 2.1. Fix an observed outcome pattern a . We start with the forward direction. Suppose the complete odds $O_{r,a}(X, W_a)$ is identifiable for all r . Then,

$$\sum_r O_{r,a}(X, W_a) = \sum_r \frac{P(R = r|X, W_a, A = a)}{P(R = 1_d|X, W_a, A = a)} = \frac{1}{P(R = 1_d|X, W_a, A = a)}$$

is identifiable. The desired selection probabilities can be recovered because

$$P(R = r|X, W_a, A = a) = P(R = 1_d|X, W_a, A = a) \cdot O_{r,a}(X, W_a)$$

for any covariate missingness pattern r .

For the converse, an identifiable set of selection probabilities $P(R = r|X, W_a, A = a)$ for all r immediately implies that the complete odds can be formed and is identifiable since the complete odds is the exactly the ratio of two selection probabilities. $\square$

of Proposition 2.2. This result is best seen by showing that the CCMV formulation in (2.6) is equivalent to the CCMV formulation in (2.7), so we will proceed with a series of equivalences. We have

$$\begin{aligned} p(x_{\bar{r}}|x_r, r, w_a, a) &= p(x_{\bar{r}}|x_r, R = 1_d, w_a, a) \\ \iff \frac{p(x, r, w_a, a)}{p(x_r, r, w_a, a)} &= \frac{p(x, 1_d, w_a, a)}{p(x_r, 1_d, w_a, a)} \\ \iff \frac{p(x, r, w_a, a)}{p(x, 1_d, w_a, a)} &= \frac{p(x_r, r, w_a, a)}{p(x_r, 1_d, w_a, a)} \\ \iff \frac{P(R = r|X, W_a, a)}{P(R = 1_d|X, W_a, a)} &= \frac{P(R = r|X_r, W_a, a)}{P(R = 1_d|X_r, W_a, a)} \\ \iff O_{r,a}(X, W_a) &= Q_{r,a}(X_r, W_a). \end{aligned}$$

Thus, the CCMV missing data assumption on the covariates is equivalent through the pattern mixture model and complete odds formulations. $\square$

of *Theorem 2.1*. We wish to show that under the CCMV assumption, the full-covariate distribution $p(x, r, w_a, a)$ is nonparametrically identified. First, we factorize the target distribution using a pattern mixture model approach

$$p(x, r, w_a, a) = p(x_{\bar{r}}|x_r, r, w_a, a)p(x_r, w_a, a).$$

Note that the distribution $p(x_r, r, w_a, a)$ is identifiable from the observed data, so the only part we need to handle is the extrapolation density $p(x_{\bar{r}}|x_r, r, w_a, a)$. By the CCMV assumption, we have the equality

$$p(x_{\bar{r}}|x_r, r, w_a, a) \stackrel{\text{CCMV}}{=} p(x_{\bar{r}}|x_r, R = 1_d, w_a, a).$$

Since the distribution $p(x, R = 1_d, w_a, a)$ is identifiable from the observed data (consider only the observations with fully-observed covariates and observed outcome pattern $A = a$), the conditional distribution $p(x_{\bar{r}}|x_r, R = 1_d, w_a, a)$ is identifiable. This means that the extrapolation density is equated to an identifiable quantity.

The distribution $p(x, r, w_a, a)$ derived from the CCMV identifying restriction does not conflict with the observed data because after marginalizing out the unobserved variable $x_{\bar{r}}$, we obtain $p(x_r, w_a, a)$, so we have nonparametric identification.

□

of *Lemma 2.2*. The expectation rewrites as

$$\begin{aligned} \mathbb{E}[f_a(X, W_a)1(A = a)1(R = r)] &= \int f_a(x, w_a)1(a' = a)1(r' = r) p(x, r, w_a, a') dx dr' dw_a da' \\ &= \int f_a(x, w_a) p(x, r, w_a, a) dx dw_a \\ &= \int f_a(x, w_a) p(x_{\bar{r}}|x_r, r, w_a, a) dx_{\bar{r}} p(x_r, r, w_a, a) dx_r dw_a. \end{aligned}$$

Recall that $m_{r,a}(x_r, w_a) = \int f_a(x, w_a) p(x_{\bar{r}}|x_r, r, w_a, a) dx_{\bar{r}}$ is the regression function. Substituting it into the previous expression yields

$$\int m_{r,a}(x_r, w_a) p(x_r, r, w_a, a) dx_r dw_a = \mathbb{E}[m_{r,a}(x_r, w_a) 1(R = r) 1(A = a)].$$

Finally, summing over all patterns r yields the second result

$$\mathbb{E}[f_a(X, W_a) 1(A = a)] = \sum_r \mathbb{E}[m_{r,a}(x_r, w_a) 1(R = r) 1(A = a)].$$

□

A.7.2 Multiple robustness and efficiency theory

This section contains proofs of Lemma 2.3, Theorem 2.2, Theorem 2.3, Corollary 2.1, Theorem 2.4.

of Lemma 2.3. Fix the observed outcome pattern a , and suppose the CCMV assumption holds for X . We will show that $\mathbb{E}[g_r(X, W_a)] = \mathbb{E}[f_a(X, W_a) 1(A = a) 1(R = r)]$ provided either $O_{r,a}(X, W_a)$ or $m_{r,a}(X_r, W_a)$ is correctly specified. By Lemmas 2.1 and 2.2, the terms (2.12) and (2.13) are known to have expected value equal to $\mathbb{E}[f_a(X, W_a) 1(A = a) 1(R = r)]$ when the complete odds $O_{r,a}(X, W_a)$ and the regression functions $m_{r,a}(X_r, W_a)$ are correctly specified, respectively.

Therefore, we focus our attention on term (2.14). There are two cases.

CASE 1. In the first case, we will assume that the complete odds $O_{r,a}(X, W_a)$ is correctly specified in that it equals $P(R = r|X, W_a, A = a)/P(R = 1_d|X, W_a, A = a)$. Since term (2.12) has the desired expectation when the complete odds are correctly specified, it suffices

to show that term (2.13) minus term (2.14) has mean 0. The difference of the terms is

$$m_{r,a}(X_r, W_a)1(R = r)1(A = a) - m_{r,a}(X_r, W_a)1(A = a)1(R = 1_d)O_{r,a}(X, W_a).$$

Taking the expectation yields

$$\begin{aligned} & \int m_{r,a}(x_r, w_a)p(x_r, r, w_a, a) dx_r dw_a \\ & - \int m_{r,a}(x_r, w_a)1(a' = a)1(s = 1_d)O_{r,a}(x, w_a) p(x, s, w_a, a') dx ds dw_a da' \\ & = \int m_{r,a}(x_r, w_a)p(x_r, r, w_a, a) dx_r dw_a \\ & - \int m_{r,a}(x_r, w_a)O_{r,a}(x, w_a) p(x, 1_d, w_a, a) dx dw_a. \end{aligned}$$

Since $O_{r,a}(X, W_a)$ is assumed to properly specified, this further simplifies to

$$\begin{aligned} & \int m_{r,a}(x_r, w_a)p(x_r, r, w_a, a) dx_r dw_a - \int m_{r,a}(x_r, w_a)p(x, r, w_a, a) dx dw_a \\ & = \int m_{r,a}(x_r, w_a)p(x_r, r, w_a, a) dx_r dw_a - \int m_{r,a}(x_r, w_a)p(x_r, r, w_a, a) dx_r dw_a \\ & = 0. \end{aligned}$$

Therefore, $\mathbb{E}[g(X, W_a)] = \mathbb{E}[f_a(X, W_a)1(A = a)]$ when $O_{r,a}(X, W_a)$ is correctly specified.

CASE 2. We now consider the second case. Suppose that the regression function $m_{r,a}(X_r, W_a)$ is correctly specified. Since term (2.13) has the desired expectation under this assumption, it suffices to show that term (2.12) minus term (2.14) has mean 0.

Consider the expectation of (2.12). Using the law of total expectation, we have

$$\mathbb{E}[f_a(X, W_a)1(A = a)1(R = 1_d)O_{r,a}(X, W_a)]$$

$$\begin{aligned}
&= \mathbb{E}[\mathbb{E}[f_a(X, W_a)O_{r,a}(X, W_a)|X_r, R = 1_d, W_a, A = a]|R = 1_d, A = a] \\
&\quad \cdot P(R = 1_d, A = a) \\
&\stackrel{\text{CCMV}}{=} \mathbb{E}[\mathbb{E}[f_a(X, W_a)Q_{r,a}(X_r, W_a)|X_r, R = r, W_a, A = a]|R = 1_d, A = a] \\
&\quad \cdot P(R = 1_d, A = a) \\
&= \mathbb{E}[\mathbb{E}[f_a(X, W_a)|X_r, R = r, W_a, A = a] \cdot Q_{r,a}(X_r, W_a)|R = 1_d, A = a] \\
&\quad \cdot P(R = 1_d, A = a).
\end{aligned}$$

Since the regression function $m_{r,a}(X_r, W_a)$ is assumed to be correctly specified, this further simplifies as

$$\begin{aligned}
&\mathbb{E}[m_{r,a}(X_r, W_a) \cdot Q_{r,a}(X_r, W_a)|R = 1_d, A = a] \cdot P(R = 1_d, A = a) \\
&= \mathbb{E}[m_{r,a}(X_r, W_a)Q_{r,a}(X_r, W_a)1(A = a)1(R = 1_d)] \\
&\stackrel{\text{CCMV}}{=} \mathbb{E}[m_{r,a}(X_r, W_a)O_{r,a}(X_r, W_a)1(A = a)1(R = 1_d)].
\end{aligned}$$

This expression is exactly equal to the expected value of term (2.14), so it follows that term (2.12) minus term (2.14) has mean 0.

Since we showed the expectation is correct in these two cases, we can conclude this form is doubly robust. $\square$

of *Theorem 2.2*. The heart of the proof lies with Lemma 2.3. We will argue that the population estimating equation has mean 0. Consider the multiple robust population estimating equation with

$$\begin{aligned}
U_r(\beta) &= \mathbb{E} \left[\Delta \left(X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \right) Q_{r,1}(X_r, Y) 1(R = 1_d) \right] \\
&+ \mathbb{E} \left[\mathbb{E} \left[X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \middle| X_r, R, Y, \Delta \right] \cdot 1(R = r) \cdot 1(\Delta = 1) \right]
\end{aligned}$$

$$- \mathbb{E} \left[\mathbb{E} \left[X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \middle| X_r, R, Y, \Delta \right] \cdot 1(\Delta = 1) \cdot 1(R = 1_d) \cdot Q_{r,1}(X_r, Y) \right],$$

$$\begin{aligned} s_r^{(1)}(\beta) &= \sum_{\delta} \mathbb{E}[1(Y \geq t) X e^{\beta^\top X} Q_{r,1}(X_r, Y) 1(\Delta = \delta) 1(R = 1_d)] \\ &\quad + \sum_{\delta} \mathbb{E}[m_{r,\delta}^{(1)}(X_r, Y; t, \beta) 1(R = r) 1(\Delta = \delta)] \\ &\quad - \sum_{\delta} \mathbb{E} \left[m_{r,\delta}^{(1)}(X_r, Y; t, \beta) \cdot 1(\Delta = 1) \cdot 1(R = 1_d) \cdot Q_{r,1}(X_r, Y) \right], \end{aligned}$$

$$\begin{aligned} s_r^{(0)}(\beta) &= \sum_{\delta} \mathbb{E}[1(Y \geq t) e^{\beta^\top X} Q_{r,1}(X_r, Y) 1(\Delta = \delta) 1(R = 1_d)] \\ &\quad + \sum_{\delta} \mathbb{E}[m_{r,\delta}^{(0)}(X_r, Y; t, \beta) 1(R = r) 1(\Delta = \delta)] \\ &\quad - \sum_{\delta} \mathbb{E} \left[m_{r,\delta}^{(0)}(X_r, Y; t, \beta) \cdot 1(\Delta = 1) \cdot 1(R = 1_d) \cdot Q_{r,1}(X_r, Y) \right]. \end{aligned}$$

From Lemma 2.3, the three above terms have expectation equivalent to

$$\begin{aligned} &\mathbb{E} \left[\Delta \left(X - \frac{s^{(1)}(Y; \beta)}{s^{(0)}(Y; \beta)} \right) 1(R = r) \right], \\ &\mathbb{E}[1(Y \geq t) X e^{\beta^\top X} 1(R = r)], \\ &\mathbb{E}[1(Y \geq t) e^{\beta^\top X} 1(R = r)], \end{aligned}$$

respectively, provided at least one of the regression functions or the complete odds $Q_{r,1}(X_r, Y)$ is correctly specified. Thus, $\sum_r U_r(\beta)$ has mean 0, and the near-solution $\widehat{\beta}_{\text{mr}}$ to the finite-sample estimating equation is consistent under standard regularity conditions. $\square$

of Theorem 2.3. We will rewrite it in the following form to exploit the results from the previous subsections

$$1(A = 1) \cdot \left(\frac{Y - m_1(X)}{\pi_A(X)} + m_1(X) \right) + 1(A = 0) \cdot m_1(X).$$

Since $\mathcal{A} = \{0, 1\}$, we have two functions f_0 and f_1 to consider. The previous display suggests taking

$$f_0(X) = m_1(X), \tag{2.22}$$

$$f_1(X, Y) = \frac{Y}{\pi_A(X)} - \frac{m_1(X)}{\pi_A(X)} + m_1(X). \tag{2.23}$$

We now apply Lemma 2.3 to both functions. For each term of $f_0(X)$ and $f_1(X, Y)$ that is a function of X , Lemma 2.3 yields 3 terms. As such, we obtain a total of 12 terms. Applying Lemma 2.3 to (2.22) yields

$$\begin{aligned} & m_1(X)1(A = 0)1(R = 1_d)Q_{r,0}(X_r) + M_{m,r,0}(X_r)1(R = r)1(A = 0) \\ & - M_{m,r,0}(X_r)1(A = 0)1(R = 1_d)Q_{r,0}(X_r). \end{aligned}$$

We obtain the following 9 terms from applying Lemma 2.3 to (2.23):

$$\begin{aligned} & m_1(X)1(A = 1)1(R = 1_d)Q_{r,1}(X_r, Y) + M_{m,r,1}(X_r, Y)1(R = r)1(A = 1) \\ & - M_{m,r,1}(X_r, Y)1(A = 1)1(R = 1_d)Q_{r,1}(X_r, Y) \\ & + \frac{Y}{\pi_A(X)}1(A = 1)1(R = 1_d)Q_{r,1}(X_r, Y) + Y M_{1/\pi_A,r,1}(X_r, Y)1(R = r)1(A = 1) \\ & - Y M_{1/\pi_A,r,1}(X_r, Y)1(A = 1)1(R = 1_d)Q_{r,1}(X_r, Y) \\ & - \frac{m_1(X)}{\pi_A(X)}1(A = 1)1(R = 1_d)Q_{r,1}(X_r, Y) - M_{m/\pi_A,r,1}(X_r, Y)1(R = r)1(A = 1) \end{aligned}$$

$$+ M_{m/\pi_{A,r,1}}(X_r, Y)1(A = 1)1(R = 1_d)Q_{r,1}(X_r, Y).$$

The double robustness follows from the proof of Lemma 2.3. $\square$

of Corollary 2.1. Compute the uncentered doubly robust influence function $q_r(X, R, Y, A)$ for each missing pattern r using Theorem 2.3. By construction, each influence function is consistent for θ_r provided either $Q_{r,a}(X_r, W_a)$ and $p(x_{\bar{r}}|x_r, w_a, a)$ are correctly specified. Each estimator for a given missing covariate pattern r is doubly robust from Theorem 2.3. The doubly robustness for each pattern implies that the final estimator is $2^{|\mathcal{R}|-1}$ -robust. $\square$

of Theorem 2.4. Let P_0 denotes the true distribution. We solve for efficient influence function using the semiparametric theory outlined in van der Vaart (1998). By Lemma 2.1 and under CCMV, the parameter of interest $\psi(P_0)$ can be written as the following statistical functional

$$\begin{aligned}\theta^{(0)} &= \int y P_0(dx, dr, dy, da) \\ &= \int \frac{1(a = 1)y}{P_0(A = 1|x)} \sum_r Q_{r,1}^{(0)}(x_r, y)1(r' = 1_d) P_0(dx, dr', dy, da),\end{aligned}$$

where $Q_{r,1}^{(0)}(x_r, y)$ denotes the complete odds $Q_{r,1}(x_r, y)$ under P_0 .

For ease of typesetting, in our abbreviations, we will adopt the convention that R comes before A . In other words, the expression $p(x, 1_d, y, 1)$ should be taken to be equivalent to $p(x, R = 1_d, y, A = 1)$.

Define $p_0(x, r', y, 1)$ and $p_0(x, r', 0)$ to be the true model, and consider a first order perturbation of the model

$$p_t(x, r', y, 1) = p_0(x, r', y, 1)[1 + t \cdot g_1(x, r', y, 1)],$$

$$p_t(x, r', 0) = p_0(x, r', 0)[1 + t \cdot g_0(x, r', 0)]$$

for functions g_1 and g_0 satisfying $\mathbb{E}_{p_0}[1(A=1)g_1(X, R', Y, A)] = \mathbb{E}_{p_0}[1(A=0)g_0(X, R', A)] = 0$. Then, the efficient influence function is the mean-zero function $\text{EIF}(x, r', y, a) = 1(A=1)\text{EIF}_1(x, r', y) + 1(A=0)\text{EIF}_0(x, r')$ satisfying

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{\theta^{(t)} - \theta^{(0)}}{t} &= \int \text{EIF}_1(x, r', y) p_0(x, r', y, 1) g(x, r', y, 1) \, dx dr' dy \\ &\quad + \int \text{EIF}_0(x, r') p_0(x, r', 0) g(x, r', 0) \, dx dr'. \end{aligned}$$

There are three terms to be perturbed: $p(x, r', y, 1)$, $Q_{r,1}(x_r, y)$, and $P(A=1|x)$. It suffices to perturb each one up to first order and hold the remaining two unchanged. Thus, we divide our task into 3 parts and proceed as follows:

Part (I). From perturbing $p(x, r', y, 1)$, a direct calculation shows that

$$\frac{1(A=1)Y}{P(A=1|X)} \sum_r Q_{r,1}(X_r, Y) 1(R=1_d) \quad (2.24)$$

contributes to the EIF.

Part (II). We now perturb the odds $Q_{r,1}(x_r, y)$. Up to first order, we have

$$\begin{aligned} Q_{r,1}^{(t)}(x_r, y) &= \frac{p_t(x_r, r, y, a)}{p_t(x_r, 1_d, y, a)} \\ &= \frac{p_0(x_r, r, y, a) + t \int p_0(x'_r, x_r, r, y, a) g_1(x'_r, x_r, r, y, a) \, dx'_r}{p_0(x_r, 1_d, y, a) + t \int p_0(x'_r, x_r, 1_d, y, a) g_1(x'_r, x_r, 1_d, y, a) \, dx'_r} + O(t^2) \\ &= Q_{r,1}^{(0)}(x_r, y) + t \left(\frac{\int p_0(x'_r, x_r, r, y, a) g_1(x'_r, x_r, r, y, a) \, dx'_r}{p_0(x_r, 1_d, y, a)} \right. \\ &\quad \left. - \frac{p_0(x_r, r, y, a) \int p_0(x'_r, x_r, 1_d, y, a) g_1(x'_r, x_r, 1_d, y, a) \, dx'_r}{p(x_r, 1_d, y, a) p(x_r, 1_d, y, a)} \right) + O(t^2). \end{aligned}$$

Thus, there are 2 additional terms that arise from this perturbation. The first term can

be found by simplifying the following expression

$$\begin{aligned}
& \int \frac{1(a=1)y}{P_0(A=1|x)} \cdot \sum_r \frac{\int p_0(x'_r, x_r, r, y, a) g_1(x'_r, x_r, r, y, a) dx'_r}{p_0(x_r, 1_d, y, a)} \cdot 1(r' = 1_d) P_0(dx, dr', dy, da) \\
&= \int \frac{1(a=1)y}{P_0(A=1|x)} \cdot \sum_r 1(s=r) \cdot p_0(x_r | x_r, 1_d, y, a) dx_r \\
&\quad p_0(x'_r, x_r, s, y, a) g_1(x'_r, x_r, s, y, a) dx'_r dx_r ds dy da.
\end{aligned}$$

So, the first term is

$$1(A=1)Y \sum_r 1(R=r) M_{1/\pi_A, r, 1}(X_r, Y), \quad (2.25)$$

where the notations $M_{f,r,a}$ is defined in equations (2.15)-(2.18).

The second term can be found by simplifying

$$\begin{aligned}
& - \int \frac{1(a=1)y}{P_0(A=1|x)} \cdot \sum_r \frac{p_0(x_r, r, y, a) \int p_0(x'_r, x_r, 1_d, y, a) g_1(x'_r, x_r, 1_d, y, a) dx'_r}{p(x_r, 1_d, y, a) p(x_r, 1_d, y, a)} \\
&\quad \cdot 1(r' = 1_d) P_0(dx, dr', dy, da) \\
&= - \int \frac{1(a=1)y}{P_0(A=1|x)} \cdot \sum_r \frac{Q_{r,1}^{(0)}(x_r, y) \int 1(s=1_d) p_0(x'_r, x_r, s, y, a) g_1(x'_r, x_r, s, y, a) dx'_r ds}{p(x_r, 1_d, y, a)} \\
&\quad \cdot 1(r' = 1_d) P_0(dx, dr', dy, da) \\
&= - \int \frac{1(a=1)y}{P_0(A=1|x)} \cdot \sum_r Q_{r,1}^{(0)}(x_r, y) \int 1(s=1_d) p_0(x'_r, x_r, s, y, a) g_1(x'_r, x_r, s, y, a) dx'_r ds \\
&\quad \cdot p(x_r | x_r, 1_d, y, a) dx_r dx_r dy da \\
&= - \int \frac{1(a=1)y}{P_0(A=1|x)} \cdot \sum_r Q_{r,1}^{(0)}(x_r, y) \cdot 1(s=1_d) \cdot p(x_r | x_r, 1_d, y, a) dx_r \\
&\quad \cdot p_0(x'_r, x_r, s, y, a) g_1(x'_r, x_r, s, y, a) dx'_r dx_r ds dy da.
\end{aligned}$$

So, the second term is

$$-1(A=1)Y1(R=1_d) \sum_r Q_{r,1}(X_r, Y) M_{1/\pi_A, r, 1}(X_r, Y). \quad (2.26)$$

Part (III). This part requires the most tedious calculations. We start by rewriting $P(A=1|x)$ as an identifiable expression under CCMV

$$\begin{aligned} & \frac{1}{P(A=1|x)} \\ &= \frac{p(x)}{p(x, A=1)} \\ &= \frac{\sum_{s \neq 1_d} \left[\int \overbrace{Q_{s,1}(x_s, y')}^{(i)} \overbrace{p(x, 1_d, y', 1)}^{(ii)} dy' + \overbrace{Q_{s,0}(x_s)}^{(iii)} \overbrace{p(x, 1_d, 0)}^{(iv)} \right] + \overbrace{p(x, 1_d)}^{(vi)}}{\sum_{s \neq 1_d} \left[\int \underbrace{Q_{s,1}(x_s, y')}_{(v)} p(x, 1_d, y', 1) dy' \right] + p(x, 1_d, 1)}. \end{aligned} \quad (2.27)$$

Our goal is to perturb the above expression and find the first order approximation. We consider each term in the above expression, and we expand up to first order:

$$\begin{aligned} (i) : \quad Q_{s,1}^{(t)}(x_s, y') &= \frac{p_t(x_s, s, y', 1)}{p_t(x_s, 1_d, y', 1)} \\ &= \frac{p_0(x_s, s, y', 1) + t \int p_0(x'_s, x_s, s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s}{p_0(x_s, 1_d, y', 1) + t \int p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s} + O(t^2) \\ &= Q_{s,1}^{(0)}(x_s, y') + t \left(\frac{\int p_0(x'_s, x_s, s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s}{p_0(x_s, 1_d, y', 1)} \right. \\ &\quad \left. - \frac{p_0(x_s, s, y', 1) \int p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s}{p(x_s, 1_d, y', 1) p(x_s, 1_d, y', 1)} \right) + O(t^2). \end{aligned}$$

$$(ii) : \quad p_t(x, 1_d, y', 1) = p_0(x, 1_d, y', 1) + t p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) + O(t^2).$$

$$\begin{aligned} (iii) : \quad Q_{s,0}^{(t)}(x_s) &= \frac{p_t(x_s, s, 0)}{p_t(x_s, 1_d, 0)} \\ &= \frac{p_0(x_s, s, 0) + t \int p_0(x'_s, x_s, s, 0) g_0(x'_s, x_s, s, 0) dx'_s}{p_0(x_s, 1_d, 0) + t \int p_0(x'_s, x_s, 1_d, 0) g_0(x'_s, x_s, 1_d, 0) dx'_s} + O(t^2) \\ &= Q_{s,0}^{(0)}(x_s) + t \left(\frac{\int p_0(x'_s, x_s, s, 0) g_0(x'_s, x_s, s, 0) dx'_s}{p_0(x_s, 1_d, 0)} \right. \\ &\quad \left. - \frac{p_0(x_s, s, 0) \int p_0(x'_s, x_s, 1_d, 0) g_0(x'_s, x_s, 1_d, 0) dx'_s}{p(x_s, 1_d, 0) p(x_s, 1_d, 0)} \right) + O(t^2). \end{aligned}$$

$$\begin{aligned} (iv) : \quad p_t(x, 1_d, 0) &= \int p_t(x, 1_d, 0) dy' \\ &= p_0(x, 1_d, 0) + t \int p_0(x, 1_d, 0) g_0(x, 1_d, 0) dy' + O(t^2). \end{aligned}$$

$$\begin{aligned} (v) : \quad p_t(x, 1_d, 1) &= \int p_t(x, 1_d, y', 1) dy' \\ &= p_0(x, 1_d, 1) + t \int p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy' + O(t^2). \end{aligned}$$

$$(vi) : \quad p_t(x, 1_d) = \int p_t(x, 1_d, y', 1) dy' + p_t(x, 1_d, 0)$$

$$= p_0(x, 1_d) + t \int p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy' da' + t p_0(x, 1_d, 0) g_0(x, 1_d, 0) + O(t^2).$$

Thus, the denominator of equation (2.27), $p(x, A = 1)$, takes the following form (up to first order):

$$\begin{aligned} A_0 + A_1 t &= p_0(x, 1) + t \int p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy' \\ &\quad + t \sum_{s \neq 1_d} \int Q_{s,1}(x_s, y') p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy' \\ &\quad + t \sum_{s \neq 1_d} \int \frac{p_0(x, 1_d, y', 1)}{p(x_s, 1_d, y', 1)} p(x'_s, x_s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s dy' \\ &\quad - t \sum_{s \neq 1_d} \int \frac{p_0(x, 1_d, y', 1) p_0(x_s, s, y', 1)}{p(x_s, 1_d, y', 1) p(x_s, 1_d, y', 1)} p(x'_s, x_s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s dy' \\ &= p_0(x, 1) + t \underbrace{\int p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy'}_{(a)} \\ &\quad + t \underbrace{\sum_{s \neq 1_d} \int Q_{s,1}(x_s, y') p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy'}_{(b)} \\ &\quad + t \underbrace{\sum_{s \neq 1_d} \int p_0(x_s | x_s, 1_d, y', 1) p_0(x'_s, x_s, s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s dy'}_{(c)} \\ &\quad - t \underbrace{\sum_{s \neq 1_d} \int p_0(x_s | x_s, 1_d, y', 1) Q_{s,1}^{(0)}(x_s, y') p(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s dy'}_{(d)} \end{aligned}$$

The numerator of equation (2.27), $p(x)$, takes the following form (up to first order):

$$\begin{aligned} a_0 + a_1 t &= p_0(x) + t \int p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy' da' \\ &\quad + t p_0(x, 1_d, 0) g_0(x, 1_d, 0) \end{aligned}$$

$$\begin{aligned}
& + t \sum_{s \neq 1_d} \int Q_{s,1}(x_s, y') p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy' \\
& + t \sum_{s \neq 1_d} \int \frac{p_0(x, 1_d, y', 1)}{p(x_s, 1_d, y', 1)} p(x'_s, x_s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s dy' \\
& - t \sum_{s \neq 1_d} \int \frac{p_0(x, 1_d, y', 1) p_0(x_s, s, y', 1)}{p_0(x_s, 1_d, y', 1) p_0(x_s, 1_d, y', 1)} p_0(x'_s, x_s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s dy' \\
& + t \sum_{s \neq 1_d} Q_{s,0}(x_s) p_0(x, 1_d, 0) g_0(x, 1_d, 0) \\
& + t \sum_{s \neq 1_d} \int \frac{p_0(x, 1_d, 0)}{p_0(x_s, 1_d, 0)} p_0(x'_s, x_s, 0) g_0(x'_s, x_s, s, 0) dx'_s \\
& - t \sum_{s \neq 1_d} \int \frac{p_0(x, 1_d, 0) p_0(x_s, s, 0)}{p_0(x_s, 1_d, 0) p_0(x_s, 1_d, 0)} p_0(x'_s, x_s, 0) g_0(x'_s, x_s, s, 0) dx'_s \\
& = p_0(x) + t \underbrace{\int p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy'}_{(e)} \\
& + t \underbrace{p_0(x, 1_d, 0) g_0(x, 1_d, 0)}_{(f)} \\
& + t \underbrace{\sum_{s \neq 1_d} \int Q_{s,1}(x_s, y') p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy'}_{(g)} \\
& + t \underbrace{\sum_{s \neq 1_d} \int p_0(x_s | x_s, 1_d, y', 1) p_0(x'_s, x_s, s, y', 1) g_1(x'_s, x_s, s, y', 1) dx'_s dy'}_{(h)} \\
& - t \underbrace{\sum_{s \neq 1_d} \int p_0(x_s | x_s, 1_d, y', 1) Q_{s,1}^{(0)}(x_s, y') p(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s dy'}_{(i)} \\
& + t \underbrace{\sum_{s \neq 1_d} Q_{s,0}(x_s) p_0(x, 1_d, 0) g_0(x, 1_d, 0)}_{(j)} \\
& + t \underbrace{\sum_{s \neq 1_d} \int p_0(x_s | x_s, 1_d, 0) p_0(x'_s, x_s, 0) g_0(x'_s, x_s, s, 0) dx'_s}_{(k)}
\end{aligned}$$

$$\underbrace{-t \sum_{s \neq 1_d} \int p_0(x_{\bar{s}}|x_s, 1_d, 0) Q_{s,0}^{(0)}(x_s) p_0(x'_{\bar{s}}, x_s, 1_d, 0) g_0(x'_{\bar{s}}, x_s, 1_d, 0) dx'_{\bar{s}}}_{(l)}$$

A fraction of the form $\frac{a_0 + a_1 t + O(t^2)}{A_0 + A_1 t + O(t^2)}$ has the following first order expansion

$$\frac{a_0}{A_0} + \left(\frac{a_1}{A_0} - \frac{a_0 A_1}{A_0^2} \right) t + O(t^2).$$

Thus, the perturbation of $1/P(A = 1|x)$ yields 12 terms, based on terms (a)-(l). We now derive each individually.

Term (a). The contribution from (a) gives

$$\begin{aligned} & -t \int \frac{1(r' = 1_d) 1(a = 1) y p_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{P_0(A = 1|x) p_0(x, 1)} \\ & \quad \times p_0(x, 1_d, y', 1) g_1(x, 1_d, y', 1) dy' dx dr' dy da \\ & = -t \int \frac{1(r' = 1_d) 1(a = 1) y p_0(x, 1_d, y, 1) \sum_r Q_{r,1}(x_r, y)}{P_0(A = 1|x) p_0(x, 1)} \\ & \quad \times p_0(x, r', y', a) g_1(x, r', y', a) dy dx dr' dy' da \\ & = -t \int \frac{1(r' = 1_d) 1(a = 1) y p(y|x, 1)}{P_0(A = 1|x)} \\ & \quad \times p_0(x, r', y', a) g_1(x, r', y', a) dy dx dr' dy' da \\ & = -t \int \frac{1(r' = 1_d) 1(a = 1) m_1(X)}{P_0(A = 1|x)} \\ & \quad \times p_0(x, r', y', a) g_1(x, r', y', a) dx dr' dy' da. \end{aligned}$$

So, (a) contributes

$$-1(R = 1_d) 1(A = 1) \frac{m_1(X)}{\pi_A(X)}.$$

Term (b). The contribution from (b) gives

$$\begin{aligned}
& -t \int \frac{1(r' = 1_d)1(a = 1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{P_0(A = 1|x)p_0(x, 1)} \\
& \quad \times \sum_{s \neq 1_d} Q_{s,1}(x_s, y')p_0(x, 1_d, y', 1)g_1(x, 1_d, y', 1) dy' dx dr' dy da \\
& = -t \int \frac{1(r' = 1_d)1(a = 1)yp(y|x, 1)}{P_0(A = 1|x)} \\
& \quad \times \sum_{s \neq 1_d} Q_{s,1}(x_s, y')p_0(x, r', y', a)g_1(x, r', y', a) dy dx dr' dy' da
\end{aligned}$$

So, (b) contributes

$$-1(R = 1_d)1(A = 1) \frac{m_1(X)}{\pi_A(X)} \sum_{r \neq 1_d} Q_{r,1}(X_r, Y).$$

Term (c). The contribution from (c) gives

$$\begin{aligned}
& -t \int \frac{1(r' = 1_d)1(a = 1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{P_0(A = 1|x)p_0(x, 1)} \\
& \quad \times \sum_{s \neq 1_d} p_0(x_{\bar{s}}|x_s, 1_d, y', 1)p_0(x'_{\bar{s}}, x_s, s, y', 1)g_1(x'_{\bar{s}}, x_s, s, y', 1) dx'_{\bar{s}} dy' dx_{\bar{s}} dx_s dr' dy da \\
& = -t \int \frac{1(a = 1)m_1(X)}{P_0(A = 1|x)} \\
& \quad \times \sum_{s \neq 1_d} p_0(x_{\bar{s}}|x_s, 1_d, y', 1)p_0(x'_{\bar{s}}, x_s, s, y', a)g_1(x'_{\bar{s}}, x_s, s, y', a) dx_{\bar{s}} dx'_{\bar{s}} dx_s dy' da \\
& = -t \int 1(a = 1) \\
& \quad \times \sum_{s \neq 1_d} M_{m/\pi_A, s, 1}(x_s, y')p_0(x'_{\bar{s}}, x_s, s, y', a)g_1(x'_{\bar{s}}, x_s, s, y', a) dx_{\bar{s}} dx'_{\bar{s}} dx_s dy' da
\end{aligned}$$

$$\begin{aligned}
&= -t \int 1(a=1) \\
&\quad \times \sum_{s \neq 1_d} 1(r''=s) M_{m/\pi_A, s, 1}(x_s, y') p_0(x'_s, x_s, r'', y', a) g_1(x'_s, x_s, r'', y', a) dx_s dx'_s dx_s dr'' dy' da
\end{aligned}$$

So, (c) contributes

$$-1(A=1) \sum_{r \neq 1_d} 1(R=r) M_{m/\pi_A, r, 1}(X_r, Y).$$

Term (d). The contribution from (d) gives

$$\begin{aligned}
&t \int \frac{1(r'=1_d) 1(a=1) y p_0(x, r', y, a) \sum_r Q_{r, 1}(x_r, y)}{P_0(A=1|x) p_0(x, 1)} \\
&\quad \times \sum_{s \neq 1_d} Q_{s, 1}(x_s, y') p_0(x_s | x_s, 1_d, y', 1) p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s dy' dx_s dx_s dr' dy da \\
&= t \int \frac{1(a=1) m_1(X)}{P_0(A=1|x)} \\
&\quad \times \sum_{s \neq 1_d} Q_{s, 1}(x_s, y') p_0(x_s | x_s, 1_d, y', 1) p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx_s dx'_s dx_s dy' da \\
&= t \int 1(a=1) \\
&\quad \times \sum_{s \neq 1_d} Q_{s, 1}(x_s, y') M_{m/\pi_A, s, 1}(x_s, y') p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s dx_s dy' da \\
&= t \int 1(a=1) \\
&\quad \times \sum_{s \neq 1_d} 1(r''=s) Q_{s, 1}(x_s, y') M_{m/\pi_A, s, 1}(x_s, y') p_0(x'_s, x_s, r'', y', 1) g_1(x'_s, x_s, r'', y', 1) dx'_s dx_s dr'' dy' da
\end{aligned}$$

So, (d) contributes

$$1(A=1) \sum_{r \neq 1_d} 1(R=1_d) M_{m/\pi_A, r, 1}(X_r, Y) Q_{r, 1}(X_r, Y).$$

Term (e). The contribution from (e) gives

$$\begin{aligned}
& t \int \frac{1(r' = 1_d)1(a = 1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\
& \quad \times p_0(x, 1_d, y', 1)g_1(x, 1_d, y', 1) dy' dx dr' dy da \\
& = t \int \frac{1(r' = 1_d)1(a = 1)yp_0(x, 1_d, y, 1) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\
& \quad \times p_0(x, r', y', a)g_1(x, r', y', a) dy dx dr' dy' da \\
& = t \int 1(r' = 1_d)1(a = 1)yp(y|x, 1) \\
& \quad \times p_0(x, r', y', a)g_1(x, r', y', a) dy dx dr' dy' da \\
& = t \int 1(r' = 1_d)1(a = 1)m_1(X) \\
& \quad \times p_0(x, r', y', a)g_1(x, r', y', a) dx dr' dy' da.
\end{aligned}$$

So, (e) contributes

$$1(R = 1_d)1(A = 1)m_1(X).$$

Term (f). The contribution from (f) gives

$$\begin{aligned}
& t \int \frac{1(r' = 1_d)1(a = 1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\
& \quad \times p_0(x, 1_d, 0)g_0(x, 1_d, 0) dx dr' dy da \\
& = t \int \frac{1(r' = 1_d)1(a = 1)1(a' = 0)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\
& \quad \times p_0(x, r', a')g_0(x, r', a') da dy dx dr' da' \\
& = t \int \frac{1(r' = 1_d)1(a = 1)1(a' = 0)yp_0(x, y, a)}{p_0(x, 1)} \\
& \quad \times p_0(x, r', a')g_0(x, r', a') da dy dx dr' da' \\
& = t \int 1(r' = 1_d)1(a' = 0)m_1(X)
\end{aligned}$$

$$\times p_0(x, r', a') g_0(x, r', a') \, da' dy \, dx dr' da'$$

So, (f) contributes

$$1(R = 1_d)1(A = 0)m_1(X).$$

Term (g). The contribution from (g) gives

$$\begin{aligned} & t \int \frac{1(r' = 1_d)1(a = 1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y) \cdot \sum_{s \neq 1_d} Q_{s,1}(x_s, y')}{p_0(x, 1)} \\ & \quad \times p_0(x, 1_d, y', 1)g_1(x, 1_d, y', 1) \, dy' \, dx dr' dy da \\ & = t \int 1(r' = 1_d)1(a = 1)m_1(X) \cdot \sum_{s \neq 1_d} Q_{s,1}(x_s, y') \\ & \quad \times p_0(x, r', y', a)g_1(x, r', y', a) \, dx dr' dy' da \end{aligned}$$

So, (g) contributes

$$1(R = 1_d)1(A = 1)m_1(X) \sum_{r \neq 1_d} Q_{r,1}(X_r, Y).$$

Term (h). The contribution from (h) gives

$$\begin{aligned} & t \int \frac{1(r' = 1_d)1(a = 1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\ & \quad \times \sum_{s \neq 1_d} p_0(x_{\bar{s}}|x_s, 1_d, y', 1)p_0(x'_{\bar{s}}, x_s, s, y', 1)g_1(x'_{\bar{s}}, x_s, s, y', 1) \, dx'_{\bar{s}} dy' \, dx_{\bar{s}} dx_s dr' dy da \\ & = t \int 1(a = 1)m_1(X) \\ & \quad \times \sum_{s \neq 1_d} p_0(x_{\bar{s}}|x_s, 1_d, y', 1)p_0(x'_{\bar{s}}, x_s, s, y', a)g_1(x'_{\bar{s}}, x_s, s, y', a) \, dx_{\bar{s}} \, dx'_{\bar{s}} dx_s dy' da \end{aligned}$$

$$\begin{aligned}
&= t \int 1(a=1) \\
&\quad \times \sum_{s \neq 1_d} M_{m,s,1}(x_s, y') p_0(x'_s, x_s, s, y', a) g_1(x'_s, x_s, s, y', a) dx_s dx'_s dy' da \\
&= t \int 1(a=1) \\
&\quad \times \sum_{s \neq 1_d} 1(r''=s) M_{m,s,1}(x_s, y') p_0(x'_s, x_s, r'', y', a) g_1(x'_s, x_s, r'', y', a) dx_s dx'_s dr'' dy' da
\end{aligned}$$

So, (h) contributes

$$1(A=1) \sum_{r \neq 1_d} 1(R=r) M_{m,r,1}(X_r, Y).$$

Term (i). The contribution from (i) gives

$$\begin{aligned}
&-t \int \frac{1(r'=1_d) 1(a=1) y p_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\
&\quad \times \sum_{s \neq 1_d} Q_{s,1}(x_s, y') p_0(x_s | x_s, 1_d, y', 1) p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s dy' dx_s dx_s dr' dy da \\
&= -t \int 1(a=1) m_1(X) \\
&\quad \times \sum_{s \neq 1_d} Q_{s,1}(x_s, y') p_0(x_s | x_s, 1_d, y', 1) p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx_s dx'_s dx_s dy' da \\
&= -t \int 1(a=1) \\
&\quad \times \sum_{s \neq 1_d} Q_{s,1}(x_s, y') M_{m,s,1}(x_s, y') p_0(x'_s, x_s, 1_d, y', 1) g_1(x'_s, x_s, 1_d, y', 1) dx'_s dx_s dy' da \\
&= -t \int 1(a=1) \\
&\quad \times \sum_{s \neq 1_d} 1(r''=s) Q_{s,1}(x_s, y') M_{m,s,1}(x_s, y') p_0(x'_s, x_s, r'', y', 1) g_1(x'_s, x_s, r'', y', 1) dx'_s dx_s dr'' dy' da
\end{aligned}$$

So, (i) contributes

$$-1(A=1) \sum_{r \neq 1_d} 1(R=1_d) M_{m,r,1}(X_r, Y) Q_{r,1}(X_r, Y).$$

Term (j). The contribution from (j) gives

$$\begin{aligned} & t \int \frac{1(r'=1_d)1(a=1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y) \cdot \sum_{s \neq 1_d} Q_{s,0}(x_s)}{p_0(x, 1)} \\ & \quad \times p_0(x, 1_d, 0)g_0(x, 1_d, 0) \, dx dr' dy da \\ & = t \int 1(r'=1_d)1(a'=0)m_1(X) \cdot \sum_{s \neq 1_d} Q_{s,0}(x_s) \\ & \quad \times p_0(x, r', a')g_0(x, r', a') \, dx dr' da' \end{aligned}$$

So, (j) contributes

$$1(R=1_d)1(A=0)m_1(X) \sum_{r \neq 1_d} Q_{r,0}(X_r).$$

Term (k). The contribution from (k) gives

$$\begin{aligned} & t \int \frac{1(r'=1_d)1(a=1)yp_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\ & \quad \times \sum_{s \neq 1_d} \int p_0(x_{\bar{s}}|x_s, 1_d, 0)p_0(x'_{\bar{s}}, x_s, 0)g_0(x'_{\bar{s}}, x_s, s, 0) \, dx'_{\bar{s}} \, dx_{\bar{s}} dx_s dr' dy da \\ & = t \int 1(a'=-)m_1(X) \\ & \quad \times \sum_{s \neq 1_d} p_0(x_{\bar{s}}|x_s, 1_d, a')p_0(x'_{\bar{s}}, x_s, a')g_0(x'_{\bar{s}}, x_s, s, a') \, dx_{\bar{s}} \, dx'_{\bar{s}} dx_s da' \\ & = t \int 1(a'=0) \end{aligned}$$

$$\times \sum_{s \neq 1_d} 1(r'' = s) M_{m,s,0}(x_s) p_0(x'_s, x_s, r'', a') g_0(x'_s, x_s, r'', a') dx'_s dx_s dr'' da'$$

So, (k) contributes

$$1(A = 0) \sum_{r \neq 1_d} 1(R = r) M_{m,r,0}(X_r).$$

Term (l). The contribution from (l) gives

$$\begin{aligned} & -t \int \frac{1(r' = 1_d) 1(a = 1) y p_0(x, r', y, a) \sum_r Q_{r,1}(x_r, y)}{p_0(x, 1)} \\ & \quad \times t \sum_{s \neq 1_d} \int p_0(x_s | x_s, 1_d, 0) Q_{s,0}^{(0)}(x_s) p_0(x'_s, x_s, 1_d, 0) g_0(x'_s, x_s, 1_d, 0) dx'_s dx_s dr' dy da \\ & -t \int 1(a' = 0) 1(r' = 1_d) m_1(X) \\ & \quad \times t \sum_{s \neq 1_d} \int p_0(x_s | x_s, 1_d, 0) Q_{s,0}^{(0)}(x_s) p_0(x'_s, x_s, r', a') g_0(x'_s, x_s, r', a') dx'_s dx_s dr' da' \\ & -t \int 1(a' = 0) 1(r' = 1_d) \\ & \quad \times t \sum_{s \neq 1_d} \int M_{m,s,0}(x_s) Q_{s,0}^{(0)}(x_s) p_0(x'_s, x_s, r', a') g_0(x'_s, x_s, r', a') dx'_s dx_s dr' da' \end{aligned}$$

So, (l) contributes

$$-1(A = 0) \sum_{r \neq 1_d} 1(R = 1_d) M_{m,r,0}(X_r) Q_{r,0}(X_r).$$

Merging Terms. Adding all of the terms together from all three parts yields the following:

$$\frac{1(A = 1)Y}{P(A = 1|X)} \sum_r Q_{r,1}(X_r, Y) 1(R = 1_d) + 1(A = 1)Y \sum_r 1(R = r) M_{1/\pi_A, r, 1}(X_r, Y)$$

$$\begin{aligned}
& -1(A=1)Y1(R=1_d)\sum_r Q_{r,1}(X_r, Y)M_{1/\pi_A, r, 1}(X_r, Y) \\
& -1(R=1_d)1(A=1)\frac{m_1(X)}{\pi_A(X)} - 1(R=1_d)1(A=1)\frac{m_1(X)}{\pi_A(X)}\sum_{r \neq 1_d} Q_{r,1}(X_r, Y) \\
& -1(A=1)\sum_{r \neq 1_d} 1(R=r)M_{m/\pi_A, 1}(X_r, Y) + 1(A=1)\sum_{r \neq 1_d} 1(R=1_d)M_{m/\pi_A, 1}(X_r, Y)Q_{r,1}(X_r, Y) \\
& + 1(R=1_d)1(A=1)m_1(X) + 1(R=1_d)1(A=0)m_1(X) \\
& + 1(R=1_d)1(A=1)m_1(X)\sum_{r \neq 1_d} Q_{r,1}(X_r, Y) \\
& + 1(A=1)\sum_{r \neq 1_d} 1(R=r)M_{m, r, 1}(X_r, Y) - 1(A=1)\sum_{r \neq 1_d} 1(R=1_d)M_{m, r, 1}(X_r, Y)Q_{r,1}(X_r, Y) \\
& + 1(R=1_d)1(A=0)m_1(X)\sum_{r \neq 1_d} Q_{r,0}(X_r) \\
& + 1(A=0)\sum_{r \neq 1_d} 1(R=r)M_{m, r, 0}(X_r) - 1(A=0)\sum_{r \neq 1_d} 1(R=1_d)M_{m, r, 0}(X_r)Q_{r,0}(X_r) \\
& = \frac{1(A=1)Y}{P(A=1|X)}\sum_r Q_{r,1}(X_r, Y)1(R=1_d) + 1(A=1)Y\sum_r 1(R=r)M_{1/\pi_A, r, 1}(X_r, Y) \\
& - 1(A=1)Y1(R=1_d)\sum_r Q_{r,1}(X_r, Y)M_{1/\pi_A, r, 1}(X_r, Y) \\
& - 1(R=1_d)1(A=1)\frac{m_1(X)}{\pi_A(X)}\sum_r Q_{r,1}(X_r, Y) \\
& - 1(A=1)\sum_{r \neq 1_d} 1(R=r)M_{m/\pi_A, 1}(X_r, Y) + 1(A=1)\sum_{r \neq 1_d} 1(R=1_d)M_{m/\pi_A, 1}(X_r, Y)Q_{r,1}(X_r, Y) \\
& + 1(R=1_d)1(A=0)m_1(X)\sum_r Q_{r,0}(X_r) + 1(R=1_d)1(A=1)m_1(X)\sum_r Q_{r,1}(X_r, Y) \\
& + 1(A=1)\sum_{r \neq 1_d} 1(R=r)M_{m, r, 1}(X_r, Y) - 1(A=1)\sum_{r \neq 1_d} 1(R=1_d)M_{m, r, 1}(X_r, Y)Q_{r,1}(X_r, Y) \\
& + 1(A=0)\sum_{r \neq 1_d} 1(R=r)M_{m, r, 0}(X_r) - 1(A=0)\sum_{r \neq 1_d} 1(R=1_d)M_{m, r, 0}(X_r)Q_{r,0}(X_r) \\
& = q(X, R, Y, A).
\end{aligned}$$

This has mean equivalent to θ , so the efficient influence function is $\text{EIF}(X, R, Y, A) = q(X, R, Y, A) - \theta$. This shows that the multiply-robust estimator for the missing response problem achieves the semiparametric efficiency bound.

□

A.7.3 Asymptotic theory

This section contains proof of Theorem 2.5 and Theorem 2.6. For the ease of derivation and referencing the literature (e.g. Chapter 19 of [van der Vaart 1998](#)), we introduce notations from the empirical process theory. We denote P as the true probability measure that generates our data and P_n be the empirical measure. Namely, in the case of observing IID random elements $Z_1, \dots, Z_n \in \mathcal{Z}$, $P(z)$ is the underlying distribution that generates $Z_1, \dots, Z_n$ and $P_n(z) = \frac{1}{n} \sum_{i=1}^n I(Z_i \leq z)$ is the corresponding empirical distribution. For any function $f : \mathcal{Z} \rightarrow \mathbb{R}$, we denote

$$Pf = \int f(z)P(dz) = \mathbb{E}(f(Z)), \quad P_nf = \int f(z)P_n(dz) = \frac{1}{n} \sum_{i=1}^n f(Z_i).$$

of Theorem 2.5. Following Lemma 2.1, we rewrite the population quantity of interest as

$$\begin{aligned} \theta &= \mathbb{E}[h(X, W, A)] \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}[f_a(X, W_a)1(A = a)] \\ &= \sum_{a \in \mathcal{A}} \sum_r \mathbb{E}[f_a(X, W_a)1(A = a)O_{r,a}(X, W_a)1(R = 1_d)] \\ &= \sum_{a \in \mathcal{A}} \sum_r \theta_{\text{IPW}, r, a}. \end{aligned}$$

Since we need to estimate f_a and $O_{r,a}$, our estimator is

$$\begin{aligned}\hat{\theta}_{\text{IPW}} &= \sum_{a \in \mathcal{A}} \sum_r \left(\frac{1}{n} \sum_{i=1}^n \hat{f}_a(X_i, W_{i,a}) 1(A_i = a) \hat{O}_{r,a}(X_i, W_{i,a}) 1(R_i = 1_d) \right) \\ &= \sum_{a \in \mathcal{A}} \sum_r \hat{\theta}_{\text{IPW},r,a}.\end{aligned}$$

We will show that $\hat{\theta}_{\text{IPW},r,a}$, as defined above, is an asymptotically linear estimator for $\theta_{\text{IPW},r,a}$.

For ease of typesetting, define

$$\begin{aligned}q_{r,a}(X, R, W_a, A) &:= f_a(X, W_a) 1(A = a) O_{r,a}(X, W_a) 1(R = 1_d), \\ \hat{q}_{r,a}(X, R, W_a, A) &:= \hat{f}_a(X, W_a) 1(A = a) \hat{O}_{r,a}(X, W_a) 1(R = 1_d).\end{aligned}$$

Using the empirical process notations, we have

$$\begin{aligned}\hat{\theta}_{\text{IPW},r,a} - \theta_{\text{IPW},r,a} &= P_n \hat{q}_{r,a} - P_0 q_{r,a} \\ &= \underbrace{(P_n - P_0) q_{r,a}}_{(I)} + \underbrace{P_0 (\hat{q}_{r,a} - q_{r,a})}_{(II)} + \underbrace{(P_n - P_0) (\hat{q}_{r,a} - q_{r,a})}_{(III)}.\end{aligned}$$

Therefore, there are 3 terms we need to consider to show asymptotic normality.

Term (I). Expanding this term, we obtain

$$(P_n - P_0) q_{r,a} = \frac{1}{n} \sum_{i=1}^n (q(X_i, R_i, W_{i,a}, A_i) - \mathbb{E}[q(X, R, W_a, A)]),$$

which is immediately asymptotically linear.

Term (II). Collect the parameters τ_a and $\eta_{r,a}$ into a single parameter $\kappa_{r,a}$ for simplicity. As the first derivative of $q_{r,a}(x, r, w_a, a; \kappa_{r,a})$ exists and is nonzero (assumptions (A4) and

(B4)), Term (II) equals

$$\begin{aligned} P_0(\widehat{q}_{r,a} - q_{r,a}) &= \int [q_{r,a}(x, r, w_a, a; \widehat{\kappa}_{r,a}) - q_{r,a}(x, r, w_a, a; \kappa_{r,a}^*)] P(dx, dr, dw_a, da) \\ &= (\widehat{\kappa}_{r,a} - \kappa_{r,a}^*)^\top \left(\int \nabla_{\kappa_{r,a}} q_{r,a}(x, r, w_a, a; \kappa_{r,a}) P(dx, dr, dw_a, da) \right) \\ &\quad + o_p(1/\sqrt{n}). \end{aligned}$$

The interchange of the derivative and the integral follows from the Dominated Convergence Theorem. Since we assume the $\widehat{\kappa}_{r,a}$ admits an asymptotically linear expansion (assumptions (A1) and (B2)), this further simplifies to

$$\frac{1}{n} \sum_{i=1}^n \varphi_{r,a}(X_i, R_i, W_{i,a}, A_i)^\top \left(\int \nabla_{\kappa_{r,a}} q_{r,a}(x, r, w_a, a; \kappa_{r,a}) P(dx, dr, dw_a, da) \right) + o_p(1/\sqrt{n}),$$

where $\varphi_{r,a}$ is the influence function for $\widehat{\kappa}_{r,a}$.

Term (III). For the last term, we show that is negligible up to first order. We apply Lemma 19.24 from (van der Vaart, 1998). We verify the following two conditions:

1. $\mathcal{Q} := \{q_{r,a}(x, r, w_a; \tau, \eta) : \tau, \eta\}$ is a P -Donsker class:

Since $\{f_a(x, w_a; \tau) : \tau\}$ and $\{O_{r,a}(x, w_a; \eta) : \eta\}$ are uniformly bounded P -Donsker classes, it follows that $\mathcal{Q}$ is also P -Donsker.

The class of functions formed by the pairwise products of functions from two uniformly bounded P -Donsker classes is P -Donsker (this permanence property can be seen using a Lipschitz transformation argument; see Example 2.10.8 from (van der Vaart and Wellner, 1996)).

2. $q_{r,a}(x, r, w_a, a; \widehat{\tau}_a, \widehat{\eta}_{r,a})$ converges to some $q_{r,a}(x, r, w_a, a; \tau_a^*, \eta_{r,a}^*)$ in $L_2(P)$:

We have

$$\begin{aligned}
& \int [q_{r,a}(x, s, w_a, a'; \widehat{\tau}_a, \widehat{\eta}_{r,a}) - q_{r,a}(x, s, w_a, a'; \tau_a^*, \eta_{r,a}^*)]^2 P(dx, ds, dw_a, da') \\
&= \int 1(a' = a) 1(s = 1_d) \left[f_a(x, w_a; \widehat{\tau}_a) O_{r,a}(x, w_a; \widehat{\eta}_{r,a}) \right. \\
&\quad \left. - f_a(x, w_a; \tau_a^*) O_{r,a}(x, w_a; \eta_{r,a}^*) \right]^2 P(dx, ds, dw_a, da') \\
&\leq \int [f_a(x, w_a; \widehat{\tau}_a) O_{r,a}(x, w_a; \widehat{\eta}_{r,a}) - f_a(x, w_a; \tau_a^*) O_{r,a}(x, w_a; \eta_{r,a}^*)]^2 \\
&\quad P(dx, ds, dw_a, da').
\end{aligned}$$

We can upper bound the $L_2(P)$ norm via

$$\begin{aligned}
& \|f_a(x, w_a; \widehat{\tau}_a) O_{r,a}(x, w_a; \widehat{\eta}_{r,a}) - f_a(x, w_a; \tau_a^*) O_{r,a}(x, w_a; \eta_{r,a}^*)\|_2 \\
&\leq \|f_a(x, w_a; \widehat{\tau}_a) [O_{r,a}(x, w_a; \widehat{\eta}_{r,a}) - O_{r,a}(x, w_a; \eta_{r,a}^*)]\|_2 \\
&\quad + \|[f_a(x, w_a; \widehat{\tau}_a) - f_a(x, w_a; \tau_a^*)] O_{r,a}(x, w_a; \eta_{r,a}^*)\|_2 \\
&\leq \|f_a(x, w_a; \widehat{\tau}_a)\|_\infty \cdot \|O_{r,a}(x, w_a; \widehat{\eta}_{r,a}) - O_{r,a}(x, w_a; \eta_{r,a}^*)\|_2 \\
&\quad + \|f_a(x, w_a; \widehat{\tau}_a) - f_a(x, w_a; \tau_a^*)\|_2 \cdot \|O_{r,a}(x, w_a; \eta_{r,a}^*)\|_\infty
\end{aligned}$$

by Minkowski's and Holder's inequalities. By assumptions (A2), (A3), (B1), and (B4),

we have $L_2(P)$ convergence of $f_a(x, w_a; \widehat{\tau}_a) O_{r,a}(x, w_a; \widehat{\eta}_{r,a})$ to the function $f_a(x, w_a; \tau_a^*) O_{r,a}(x, w_a; \eta_{r,a}^*)$.

Therefore, the sufficient conditions of the Lemma 19.24 of (van der Vaart, 1998) are satisfied, and it follows that Term (III) is $o_p(1/\sqrt{n})$.

Merging terms. Finally, putting everything together, we have

$$\sqrt{n}(\widehat{\theta}_{r,a} - \theta_{r,a})$$

$$\begin{aligned}
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n (q(X, R, W_a, A) - \mathbb{E}[q(X, R, W_a, A)]) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi_{r,a}(X_i, R_i, W_{i,a}, A_i)^\top \left(\int \nabla_{\kappa_{r,a}} q_{r,a}(x, r, w_a, a; \kappa_{r,a}) P(dx, dr, dw_a, da) \right) \\
&\quad + o_p(1)
\end{aligned}$$

for some influence function $\varphi_{r,a}$ for $\widehat{\kappa}_{r,a}$ (this follows from the asymptotic linearity assumptions (A1) and (B2)). Next, we note that $\widehat{\theta}_{\text{IPW},r,a}$ is generally correlated with this $\widehat{\theta}_{\text{IPW},r',a'}$ for some $r \neq r'$ and $a \neq a'$. Therefore, they do not typically converge jointly to a degenerate distribution. As we have provided the asymptotic linear forms for each of the estimators marginally, we have joint convergence of the $\{\sqrt{n}(\widehat{\theta}_{\text{IPW},r,a} - \theta_{\text{IPW},r,a})\}_{r,a}$ to a zero-mean multivariate Gaussian by the Central Limit Theorem. The asymptotic normality of $\widehat{\theta}_{\text{IPW}}$ follows from the Continuous Mapping Theorem. Consistency follows from asymptotic normality since we have convergence at the standard $\sqrt{n}$ -rate. $\square$

of Theorem 2.6. Using Lemma 2.2, we can decompose θ using a regression adjustment method as follows

$$\begin{aligned}
\theta &= \mathbb{E}[h(X, W, A)] \\
&= \sum_a \mathbb{E}[f_a(X, W_a)1(A = a)] \\
&= \sum_a \sum_r \mathbb{E}[m_{r,a}(X_r, W_a)1(R = r)1(A = a)] \\
&= \sum_a \sum_r \theta_{\text{RA},r,a},
\end{aligned}$$

where $m_{r,a}(X_r, W_a) = E[f_a(X, W_a)|X_r = r, W_a = w_a, A = a]$ is a regression function. Here $\theta_{\text{RA},r,a} = \int \mathbb{E}[f_a(X, W_a)|X_r = x_r, R = r, W_a = w_a, A = a] P_{\text{obs}}(dx_r, r, dw_a, a)$.

To start, we fix a covariate pattern r and an observed outcome pattern a . Define

$$\begin{aligned}\widehat{\theta}_{\text{RA},r,a} &= \int f_a(x, w_a; \widehat{\tau}_a) \widehat{P}_{ex}(dx_{\bar{r}}|x_r, r, w_a, a) \widehat{P}_{obs}(dx_r, r, dw_a, a), \\ \widetilde{\theta}_{\text{RA},r,a} &= \int f_a(x, w_a; \tau_a^*) \widehat{P}_{ex}(dx_{\bar{r}}|x_r, r, w_a, a) \widehat{P}_{obs}(dx_r, r, dw_a, a), \\ \theta_{\text{RA},r,a} &= \int f_a(x, w_a; \tau_a^*) P_{ex}(dx_{\bar{r}}|x_r, r, w_a, a) P_{obs}(dx_r, r, dw_a, a).\end{aligned}$$

Then, we can decompose the difference via

$$\sqrt{n}(\widehat{\theta}_{\text{RA},r,a} - \theta_{\text{RA},r,a}) = \underbrace{\sqrt{n}[\widehat{\theta}_{\text{RA},r,a} - \widetilde{\theta}_{\text{RA},r,a}]}_{(I)} + \underbrace{\sqrt{n}[\widetilde{\theta}_{\text{RA},r,a} - \theta_{\text{RA},r,a}]}_{(II)}.$$

Term (I). First, we start with Term (I), which rewrites as

$$\begin{aligned}& \sqrt{n} \int [f_a(x, w_a; \widehat{\tau}_a) - f_a(x, w_a; \tau_a^*)] \widehat{P}_{ex}(dx_{\bar{r}}|x_r, w_a, a) \widehat{P}_{obs}(dx_r, r, dw_a, a) \\ &= \sqrt{n}(\widehat{\tau}_a - \tau_a^*)^\top \left(\int \nabla_{\tau_a} f_a(x, w_a; \tau_a) \widehat{P}(dx, r, dw_a, a) \right) + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_a(X_i, R_i, W_{i,A_i}, A_i) \left(\int \nabla_{\tau_a} f_a(x, w_a; \tau_a) \widehat{P}(dx, r, dw_a, a) \right) + o_p(1)\end{aligned}$$

by the parametric assumption of (A1).

Term (II). Now, we consider Term (II), which requires slightly more work. Expanding it and grouping terms, we obtain the following decomposition.

$$\begin{aligned}& \sqrt{n}(\widetilde{\theta}_{\text{RA},r,a} - \theta_{\text{RA},r,a}) \\ &= \sqrt{n} \int \int f_a(x, w_a; \tau_a^*) \widehat{P}_{ex}(dx_{\bar{r}}|x_r, r, w_a, a) \widehat{P}_{obs}(dx_r, r, dw_a, a) \\ &\quad - \sqrt{n} \int \int f_a(x, w_a; \tau_a^*) P_{ex}(dx_{\bar{r}}|x_r, r, w_a, a) P_{obs}(dx_r, r, dw_a, a)\end{aligned}$$

$$\begin{aligned}
&= \underbrace{\sqrt{n} \int \int f_a(x, w_a; \tau_a^*) \hat{P}_{ex}(dx_{\bar{r}}|x_r, r, w_a, a) (\hat{P}_{obs}(dx_r, r, dw_a, a) - P_{obs}(dx_r, r, dw_a, a))}_{(A)} \\
&\quad - \underbrace{\sqrt{n} \int \int f_a(x, w_a; \tau_a^*) (\hat{P}_{ex}(dx_{\bar{r}}|x_r, r, w_a, a) - P_{ex}(dx_{\bar{r}}|x_r, r, w_a, a)) P_{obs}(dx_r, r, dw_a, a)}_{(B)}
\end{aligned}$$

Term A of (I). Term A rewrites as

$$\sqrt{n} \int m_{r,a}(x_r, w_a; \hat{\eta}_{r,a}) (\hat{P}_{obs}(dx_r, r, dw_a, a) - P_{obs}(dx_r, r, dw_a, a)),$$

This is equal to

$$\begin{aligned}
&\sqrt{n} \int (m_{r,a}(x_r, w_a; \hat{\eta}_{r,a}) - m_{r,a}(x_r, w_a; \eta_{r,a}^*)) (\hat{P}_{obs}(dx_r, r, dw_a, a) - P_{obs}(dx_r, r, dw_a, a)) \\
&\quad + \sqrt{n} \int m_{r,a}(x_r, w_a; \eta_{r,a}^*) (\hat{P}_{obs}(dx_r, r, dw_a, a) - P_{obs}(dx_r, r, dw_a, a)).
\end{aligned}$$

Observe that the conditions of Lemma 19.24 of (van der Vaart, 1998) are satisfied (namely, $\hat{m}_{r,a}$ belongs to a Donsker class by (C2) and we have $L_2(P)$ convergence to the true regression function by (C3)). Therefore, it follows that the first term in the previous display is $o_p(1)$. Thus, for asymptotic analysis, we can replace $\hat{\eta}_{r,a}$ in Term A with the population version $\eta_{r,a}^*$. This means that Term A

$$\sqrt{n} \int m_{r,a}(x_r, w_a; \hat{\eta}_{r,a}) (\hat{P}_{obs}(dx_r, r, dw_a, a) - P_{obs}(dx_r, r, dw_a, a))$$

is asymptotically equivalent to (up to an $o_p(1)$ term)

$$\sqrt{n} \int m_{r,a}(x_r, w_a; \eta_{r,a}^*) (\hat{P}_{obs}(dx_r, r, dw_a, a) - P_{obs}(dx_r, r, dw_a, a)).$$

Term B of (II). We now consider the term B. It rewrites as

$$\begin{aligned} & \sqrt{n} \int (m_{r,a}(x_r, w_a; \hat{\eta}_{r,a}) - m_{r,a}(x_r, w_a; \eta_{r,a}^*)) P_{obs}(dx_r, r, dw_a, a) \\ &= \sqrt{n}(\hat{\eta}_{r,a} - \eta_{r,a}^*)^\top \left(\int \nabla_{\eta_{r,a}} m_{r,a}(x_r, w_a; \eta_{r,a}) P_{obs}(dx_r, r, dw_a, a) \right) + o_p(1). \end{aligned}$$

Since the first derivative of $m_{r,a}(x_r, w_a; \eta_{r,a})$ is bounded, the interchange of the derivative and the integral follows from the Dominated Convergence Theorem.

Collecting Terms (I) and (II). We combine Terms A and B of Term (II) and add them to Term (I), and it follows that

$$\begin{aligned} & \sqrt{n}(\hat{\theta}_{RA,r,a} - \theta_{RA,r,a}) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_a(X_i, R_i, W_{i,A_i}, A_i) \left(\int \nabla_{\tau_a} f_a(x, w_a; \tau_a) \hat{P}(dx, r, dw_a, a) \right) \\ & \quad + \sqrt{n} \int m_{r,a}(x_r, w_a; \eta_{r,a}^*) (\hat{P}_{obs}(dx_r, r, dw_a, a) - P_{obs}(dx_r, r, dw_a, a)) \\ & \quad + \sqrt{n}(\hat{\eta}_{r,a} - \eta_{r,a}^*)^\top \left(\int \nabla_{\eta_{r,a}} m_{r,a}(x_r, w_a; \eta_{r,a}) P_{obs}(dx_r, r, dw_a, a) \right) \\ & \quad + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_a(X_i, R_i, W_{i,A_i}, A_i) \left(\int \nabla_{\tau_a} f_a(x, w_a; \tau_a) \hat{P}(dx, r, dw_a, a) \right) \\ & \quad + \sqrt{n}(\hat{\eta}_{r,a} - \eta_{r,a}^*)^\top \left(\int \nabla_{\eta_{r,a}} m_{r,a}(x_r, w_a; \eta_{r,a}) P_{obs}(dx_r, r, dw_a, a) \right) \\ & \quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[m_{r,a}(X_{i,R_i}, W_{i,A_i}; \eta_{r,a}^*) \right. \\ & \quad \quad \left. - \mathbb{E}[1(R=r)1(A=a)m_{r,a}(X_R, W_A; \beta_{r,a}^*)] \right] \\ & \quad + o_p(1). \end{aligned}$$

By Slutsky's lemma and the WLLN, this is equivalent up to an $o_p(1)$ term to

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_a(X_i, R_i, W_{i,A_i}, A_i) \cdot \mathbb{E}[1(R=r)1(A=a) \nabla_{\tau_a} f_a(X, W_a; \tau_a)] \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{r,a}(X_i, R_i, W_{i,A_i}, A_i) \cdot \mathbb{E}[1(R=r)1(A=a) \nabla_{\beta_{r,a}} m_{r,a}(X_r, W_a; \eta_{r,a})] \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[m_{r,a}(X_{i,R_i}, R_i, W_{i,A_i}, A_i; \beta_{r,a}^*) \right. \\
& \quad \left. - \mathbb{E}[1(R=r)1(A=a) m_{r,a}(X_R, R, W_A, A; \beta_{r,a}^*)] \right],
\end{aligned}$$

where $\psi_{r,a}$ is the influence function of $\hat{\eta}_{r,a}$ since it is assumed to be an asymptotic linear estimator by Assumption (C1).

Thus, we have showed that $\hat{\theta}_{\text{RA},r,a}$ is an asymptotically linear estimator. Next, note that in standard missing data problems, different subsets of the observed data set are used to estimate the extrapolation distribution $P_{ex}(x_{\bar{r}}|x_r, r, w_a, a)$ for different values of r and a , as this relies on the identifying restrictions used. Therefore, $|\hat{\theta}_{\text{RA},r,a}, \hat{\theta}_{\text{RA},r',a'}| \neq 1$ when $r \neq r'$ or $a \neq a'$, so the joint convergence of all the $\hat{\theta}_{\text{RA},r,a}$ s to a multivariate normal distribution follows since marginally, they all admit an asymptotic linear expansion. Finally, applying the Continuous Mapping Theorem yields the asymptotic normality of $\hat{\theta}_{\text{RA}}$, and consistency follows. $\square$

Chapter 3

MODELING MISSING AT RANDOM NEUROPSYCHOLOGICAL TEST SCORES USING A MIXTURE OF BINOMIAL PRODUCT EXPERTS

1 Introduction

Modeling multivariate discrete data is a common problem across many fields such as social sciences, psychology, and ecology. For instance, in education research, discrete data arises from students' test scores (Maydeu-Olivares and Liu, 2015), and in ecology, discrete data arises when the counts of a given animal population are measured over various areas and time periods (Anderson et al., 2019; Maslen et al., 2023). As the size of data sets continues to expand, missing data becomes increasingly prevalent (Kang, 2013), and it is common to have access to a rich set of discrete variables that is subject to missingness. Both the multivariate discreteness and the missingness alone can create modeling difficulties, but together, they present a unique challenge for statistical inference.

1.1 *The National Alzheimer's Coordinating Center database*

Although the general problem statement is widely important, our work is primarily motivated by the analysis of neuropsychological test scores in the database of the National Alzheimer's Coordinating Center (NACC)¹. The NACC is an NIH/NIA-funded center that maintains the largest longitudinal database for Alzheimer's disease in the United States. The NACC coordinates 33 Alzheimer's Disease and Research Centers (ADRC) in the United States.

¹<https://naccdata.org/>

In the NACC data, we have 41,181 individuals with data collected from 2005 to 2019. In this study, we are particularly interested in information regarding the neuropsychological tests. The neuropsychological tests are a set of examinations measuring an individual's cognitive ability in four different domains: language, attention, executive function, and memory. The scores from these tests are important features for dementia research because they are not based on a clinician's judgment. Similar to a conventional exam, the outcome of a neuropsychological test is a discrete number taking integer values and has a known range. Note that different tests have a different ranges. In our study, we consider eight neuropsychological test scores per individual.

A goal of this chapter is to introduce a simple and interpretable statistical model for modeling the neuropsychological tests. The discrete and correlated nature of the neuropsychological tests presents a challenge for modeling them. As far as we are aware, there is no statistical model for handling these test scores directly, not to mention the additional challenges from missingness of the test scores. To simplify the problem, we focus on a static model that only considers the test scores from each individual's initial visit. While there are many covariates that could be incorporated in the analysis, we choose to include only four demographic variables (age, education, sex, and race) to reduce the complexity of the model. These variables are key demographic variables that are often included in any dementia research.

1.2 Research questions

Because the ranges of exams are generally large (30 to 300), it is infeasible to nonparametrically model the joint distribution of the eight test scores. Moreover, we have demographic variables such as education, age, and sex. It is a nontrivial problem to study the relation between these demographic variables and the test scores. Furthermore, the existence of miss-

ing test scores among certain individuals further compounds the complexity of the overall analysis. In this chapter, our primary focus centers on addressing the following research questions:

1. **Finding a feasible and interpretable model for multiple discrete outcomes with missing values.** As mentioned before, our data contains multiple dependent discrete variables with missing values. We need to design a proper model to model the dependency among discrete variables and deal with the missing data problem. We also need an estimation and inference procedure that quantifies the uncertainty in the model while properly accounting for the missing data in a statistically principled way.
2. **Discovering latent groups of individuals using neuropsychological tests.** Clinicians have developed a set of rules to categorize individuals into different clinical groups (normal cognition, MCI, and dementia). However, this rule is based on clinical judgments and does not involve any neuropsychological information. Clustering has been of interest to the dementia research community ([Alashwal et al., 2019](#)). Thus, an ideal model should be able to cluster individuals into groups using neuropsychological test scores that represent multiple cognitive domains of interest.
3. **Investigating the association between neuropsychological tests and other variables.** It is often of a lot of interest to study how neuropsychological tests are associated with demographic variables or clinical judgments. For instance, there is a hypothesis that people with a higher education are more resilient to cognitive decline as a form of cognitive reserve ([Meng and D’Arcy, 2012](#); [Thow et al., 2018](#)). We also want to use the NACC data to test this hypothesis.

1.3 Literature review

We provide some background on the current research in the Alzheimer’s disease and statistics methodology literature.

Dementia-related research

Multiple outcome variables are common in dementia-related research, but there is no clear widespread solution to the modeling problem. In some previous methods in the Alzheimer’s disease literature, multiple test scores are standardized and averaged into a single holistic score for an individual (Boyle et al., 2018). This turns the multiple outcome problem into a single outcome problem and ignores the dependence structure. Mjørud et al. (2014) examined variables associated with multiple quality-of-life related discrete outcomes and performed univariate regression on each outcome. Dimension reduction is also commonly used to reduce the number of outcome variables (Yesavage et al., 2016; Qiu et al., 2019), but this can lead to loss of interpretability as any downstream analysis is not using the original variables. Missing data is often ignored completely or when accounted for, single imputation methods or an off-the-shelf imputation approach are typically used (Brenowitz et al., 2017; Qiu et al., 2019).

Clustering has also become increasingly valuable in the dementia research community (Alashwal et al., 2019). For example, Escudero et al. (2011) used the K -means algorithm to divide subjects into pathologic and non-pathologic groups to study the early detection of Alzheimer’s disease. Tosto et al. (2016) used K -means on a subset of NACC data to identify subgroups within Alzheimer’s disease patients to understand the heterogeneity of the disease. Several papers also described model-based clustering approaches. For example, De Meyer et al. (2010) clustered biomarker data using a simple two-component Gaussian mixture model for early detection of Alzheimer’s disease. Qiu et al. (2019) used neuropsych-

chological test scores from a smaller data set than ours, imputing missing data using single imputation methods and the *mice* R package (van Buuren and Groothuis-Oudshoorn, 2011a). Then, they applied principal components analysis before finally, using a Gaussian mixture model for clustering. In our work, we choose to model the raw test scores to avoid loss of information and preserve more interpretability. It is crucial to be cautious when employing single imputation methods or off-the-shelf packages, as they can potentially lead to an underestimation of the uncertainty associated with missing data, and the underlying assumptions regarding missing data may remain unclear.

Additionally, the cognitive reserve hypothesis emerged several decades ago when it was observed that after autopsies, some individuals with no dementia symptoms had brains that exhibited advanced Alzheimer’s pathology (Katzman et al., 1989). Researchers hypothesized that some activities in life may provide individuals with a given resilience to cognitive decline, which is known as the cognitive reserve hypothesis. As such, there is deep interest in discovering the association between various covariates and dementia (Zhang et al., 1990; Stern, 2012; Meng and D’Arcy, 2012; Thow et al., 2018).

Methodology research

this chapter sits at the intersection of multiple subfields of statistics including latent variable modeling, missing data, and clustering. The latent class model was first proposed by Goodman (1974) for the purpose of modeling multivariate discrete data. The basic form of the latent class model is of a mixture of multivariate multinomials. When the number of levels of the discrete random variable is large, this can lead to a large number of parameters (Bouveyron et al., 2019). There has been much work studying many aspects of the latent class model including the identifiability (Gyllenberg et al., 1994; Carreira-Perpinan and Renals, 2000; Allman et al., 2009; Gu and Xu, 2020) and incorporation of covariates (Huang

and Bandeen-Roche, 2004; Vermunt, 2010; Ouyang and Xu, 2022). The traditional latent class model has been extended in various forms such as a mixture of Rasch models (Rost, 1990) and a mixture of item response models (McParland and Gormley, 2013).

The mixture of experts models are closely related to the mixture model framework and the latent class model (Jacobs et al., 1991). It generalizes the standard mixture model by allowing the parameters to potentially depend on covariates (Gormley and Frühwirth-Schnatter, 2018). The increased flexibility of the models is associated with a large increase in number of parameters. Many estimation procedures have been explored including those based on an EM approach (Jacobs et al., 1991; Jordan and Jacobs, 1993), an EMM approach (Gormley and Murphy, 2008), or a Bayesian framework (Bishop and Svenskn, 2002; Frühwirth-Schnatter and Kaufmann, 2008).

Clustering has previously been explored in combination with missing data. Lee and Harel (2022) proposed a two stage clustering framework by first clustering multiply imputed data sets individually and then, obtaining a final clustering by clustering the set of cluster centers obtained over all imputed data sets. In previous research on model-based clustering with missing data, Serafini et al. (2020) employed the EM algorithm and Monte Carlo methods to estimate Gaussian mixture models in the presence of missing at random data. Sportisse et al. (2023) addressed model-based clustering under missing-not-at-random assumptions, employing a likelihood-based approach and EM algorithm. We include further comments on these methods in Appendix B.4. These methods addressed the clustering problem but did not use covariates or discuss quantifying the uncertainty of the parameters. Since our method is motivated by the statistical analysis of Alzheimer’s disease, handling both of these is critical to our work.

1.4 Outline

This chapter is organized as follows. Section 2 presents a detailed description of our *mixture of binomial product experts* model. The handling of missing data is discussed in Section 3, where we describe the fitting of the model under a nonmonotone missing at random assumption. In Section 4, we delve into the process of inference, specifically in constructing confidence intervals for the parameters of interest. We outline how to perform clustering in Section 5. In Section 6, we provide simulations results to examine the consistency and the coverage of our proposed estimator. In Section 9, we apply our method to an Alzheimer's disease data set, which motivates our model formulation. Finally, Section 8 concludes the chapter with a discussion of the results. In the Appendix, we include further details on derivations and proofs (Appendices B.1 and B.8), identification theory (Appendix B.2), code implementation (Appendix B.3), remarks on related EM algorithms (Appendix B.4), simulations (Appendix 8), comments on model selection (Appendix B.6), comments on clustering (Appendix B.7), and more comments on the real data analysis (Appendix B.9). We provide a code implementation of our method in an R package.

2 Mixture of Binomial Product Experts

2.1 A latent class model for neuropsychological test scores

Suppose we have n IID observations, indexed by $i = 1, \dots, n$. Let $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,p})^\top \in \mathcal{X} \subset \mathbb{R}^p$ be the covariates and $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,d})^\top \in \mathcal{Y} \subset \mathbb{Z}^d$ be the outcome variables of interest, where $\mathcal{Y}$ is a bounded discrete set. In the present study, $\mathbf{Y}_i$ represents the neuropsychological test scores of individual i for a total of d total tests, but this framework can be applied to other problems as well. In the chapter, for notational convenience, we may elect to drop the index i to simplify notation and refer to the generic random variable that is not associated with a specific individual. The j -th test score/outcome variable Y_j belongs

to the set $\{0, 1, 2, \dots, N_j\}$. We assume that the covariates are always observed and that only the outcomes may be subject to missingness.

We start with a simpler scenario in which the data is completely observed (no absence of any outcome variables). The more complicated setting with missing outcome variables is addressed in Section 3. When we have multiple discrete outcome variables, the nonparametric approach is not feasible due to the curse of dimensionality. To see this, suppose we have four outcome variables, and each outcome variable has ten possible values. The joint distribution has a dimension of 10^4 , surpassing the order of magnitude of moderately sized samples. This motivates to use a parametric model.

To model the dependency among discrete variables, we consider a mixture model, which is based on the idea of the latent class model (LCM). Originally presented in its modern form by Goodman (1974), the latent class model takes the form of the following mixture of multivariate multinomials distribution

$$p(y; \theta) = \sum_{k=1}^K w_k p_k(y; \theta_k) = \sum_{k=1}^K w_k \prod_{j=1}^d p_{k,j}(y_j; \theta_{k,j}), \quad (3.1)$$

where $p_{k,j}$ denotes a multinomial with $\theta_{k,j}$ being its the parameter and the w_k 's are probability weights for each component such that $\sum_k w_k = 1$ and $w_k > 0$ for all k . As $p_{k,j}$ is a multinomial distribution, $\theta_{k,j}$ refers to vector of probabilities (one for each level of Y_j) that sum to 1. There is a popular implementation for fitting these models through the *poLCA* package (Linzer and Lewis, 2011).

Let Z denote the component assignment such that $Z = \ell$ if and only if the associated observation is generated from mixture component ℓ . Namely, $P(Y = y | Z = k; \theta) = \prod_{j=1}^d p_{k,j}(y_j; \theta_{k,j})$. By construction, the latent class model implicitly assumes for $j \neq j'$, the variables Y_j and $Y_{j'}$ are conditionally independent given the latent class Z . This is known

as the local independence assumption and is commonly used in the literature to model dependence between discrete random variables. While they are conditionally independent, the variables Y_j and $Y_{j'}$ are allowed to depend marginally for $j \neq j'$. The local independence assumption makes it convenient to calculate the conditional distributions, which are used in Section 3 when we discuss imputation. Some effort has been made to relax the local independence assumption such as mixtures of log-linear models (Bock, 1994) and hierarchical latent class models (Zhang, 2004; Chen et al., 2012), but these have not been as widely adopted due to the increase in the number of parameters.

Throughout this chapter, we will refer to the data set containing tuples of the form $(\mathbf{X}, \mathbf{Y})$ as the *latent incomplete* (LI) data and the data set containing tuples of the form $(\mathbf{X}, \mathbf{Y}, Z)$ as the *latent complete* (LC) data. In the setting of no missing data, the latent incomplete data is what we typically have access to, but this is often insufficient for direct model fitting due to the unobserved latent variable Z . We will show that one can estimate the model parameters using an EM algorithm with the latent incomplete data. We use this nomenclature to avoid confusion in Section 3 when we encounter the traditional type of missingness with the outcome variables.

There has been previous work to incorporate covariates in the latent class model (Huang and Bandeen-Roche, 2004; Vermunt, 2010). This is related to the *mixture of experts* models that arose from the machine learning literature (Jacobs et al., 1991; Jordan and Jacobs, 1993; Yuksel et al., 2012). These models generalize the classical mixture model formulation by allowing the model parameters w_k and θ_k to depend on covariates. The *simple mixture of experts* model (Gormley and Frühwirth-Schnatter, 2018) takes the following form

$$p(y|x; \boldsymbol{\theta}) = \sum_{k=1}^K w_k(x; \beta_k) p_k(y; \theta_k),$$

where β_k is introduced to allow the weights to depend on covariates. As written, the covariates only adjust the weights placed on each mixture component but do not affect the parameters of the individual component distribution.

To apply these ideas to the neuropsychological data, we combine the mixture of experts and LCM, leading to the following mixture of binomial product experts

$$p(y|x; \beta, \theta) = \sum_{k=1}^K w_k(x; \beta_k) \prod_{j=1}^d \underbrace{\binom{N_j}{y_j} (\theta_{k,j})^{y_j} (1 - \theta_{k,j})^{N_j - y_j}}_{\underbrace{p_{k,j}(y_j; \theta_{k,j})}_{p_k(y; \theta_k)}}, \quad (3.2)$$

where $w_k(x; \beta_k) = \exp(\beta_k^\top \cdot (1, x)) / (\sum_{k' \in [K]} \exp(\beta_{k'}^\top \cdot (1, x)))$ is from the multiple-class logistic regression model. Like in the classical LCM, we utilize the local independence assumption as each mixture decomposes as a product of binomials. However, there is a critical difference between our model and the classical LCM. We use the binomial distribution instead of the multinomial distribution, and this leads to large reduction in the number of parameters. Since test scores are ordered variables, we expect the reduction to a binomial distribution to be reasonable as scores further away from the mean score can be less likely. Similar to the simple mixture of experts model, each mixture weight depends on covariates.

Here we will treat $k = 1$ as the reference group, so β_1 is the vector of all 0 by assumption and does not need to be estimated. For $k \in \{2, 3, \dots, K\}$, $\beta_k \in \mathbb{R}^{p+1}$, and for $k \in [K]$, $\theta_k \in \mathbb{R}^d$, we assume that the w_k utilize the logistic function, but it is possible to use other link functions such as the probit and other monotonic functions (Ouyang and Xu, 2022). In our work, the outcome variables are neuropsychological test scores, and we treat the latent groups as subgroups of the population of varying cognitive ability. The five levels of the CDR score offer a natural choice for the number of mixture components.

In our model construction, we view each mixture component as representing a subgroup

of the population, where the subgroups have varying cognitive ability that is summarized by θ . Therefore, we interpret the θ 's as attributes of population-level groups and do not allow them to depend on covariates. Since the weights $w_k(x; \beta_k)$ depend on covariates, every individual has their own weights associated with each population-level group. These covariates can be interpreted as a way to construct weights on an individual level for each of these population-level subgroups. A side product of this assumption is that this reduces model complexity. If the θ 's depend on covariates, then the number of parameters increases from $O(K(d + p))$ to $O(Kdp)$, which is a dramatic change even for moderately sized d and p and thus another reason we do not allow this as part of our model. Because the outcome variables are test scores, we can also interpret the quantity $N_j \cdot \theta_{k,j}$ as the mean of the j -th test score of the k -th latent group, and this is useful as a summary statistic for a given latent class.

Binomial product distributions have previously arisen in the literature. One such application was for modeling test scores for spatial tasks in child development in [Thomas et al. \(1993\)](#), but they considered a fairly limited setting with no covariates, two components, and two outcome variables. Binomial product and Poisson product distributions have also been used in the statistical ecology literature to model species abundance ([Kéry et al., 2005](#); [Haines, 2016](#); [Brintz et al., 2018](#)). These settings differ from ours because they often treat the total number counts N_j as a random quantity. In item response theory (IRT), the Rasch model posits that the probability of answering each question correctly depends on its difficulty and the individual's ability, assuming a Bernoulli distribution for each question ([Rasch, 1960](#)). However, in our data, only total test scores are available. Thus, assuming uniform difficulty across test questions, we model the scores using Binomial distributions. To our knowledge, this chapter is the first time binomial products has been used to analyze neuropsychological test score data while incorporating covariates.

Note that for the k -th component, $\beta_k \in \mathbb{R}^{p+1}$ and $\theta_k \in \mathbb{R}^d$, there are $(K-1)(p+1)$ parameters associated with the covariates β and Kd parameters associated with θ and a total of $K(p+d+1) - p - 1$ parameters for the whole model. Thus, for fixed covariate and outcome dimensions, the number of parameters grows linearly in K .

2.2 Model fitting

We now describe our procedure of fitting the model. An intuitive approach is to estimate the parameters via the maximum likelihood approach. However, computing the maximum likelihood estimator (MLE) is a nontrivial task because the latent incomplete log-likelihood is not concave (Lemma 3.1). As such, solving the first order conditions is no longer sufficient for determining the global maximizer.

Due to Lemma 3.1 (stated formally in Appendix B.8.1), maximizing the latent incomplete log-likelihood function directly is not straightforward. Here is an interesting insight from the mixture model literature: if we had access to the latent class label, then the maximum likelihood estimator would be easy to construct. This insight motivates us to develop an EM algorithm (Dempster et al., 1977) by augmenting the observed data with the true component label Z to calculate the MLE for (β, θ) . For each $i = 1, \dots, n$, let $Z_i = \ell$ if the i -th observation comes from mixture component ℓ . Such data augmentation allows one to bypass the problem of taking the logarithm of a sum. To see this, let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ and $\mathbf{Y}_{1:n} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ be the covariates and outcome variables of all n individuals, respectively. The latent complete (LC) log-likelihood function writes as

$$\begin{aligned} \ell_{\text{LC},n}(\beta, \theta; \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}, \mathbf{Z}_{1:n}) &= \log \left(\prod_{i=1}^n \prod_{k=1}^K (w_k(\mathbf{X}_i; \beta_k) p_k(\mathbf{Y}_i; \theta_k))^{1(Z_i=k)} \right) \\ &= \sum_{i=1}^n \sum_{k=1}^K 1(Z_i = k) \log w_k(\mathbf{X}_i; \beta_k) + \sum_{i=1}^n \sum_{k=1}^K 1(Z_i = k) \log p_k(\mathbf{Y}_i; \theta_k). \end{aligned} \quad (3.3)$$

In the EM algorithm, we start from an initial guess $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)})$ and iterate between an expectation step (E-step) and a maximization step (M-step) to update our guess until convergence. Let t denote the t -th iteration of the EM algorithm. In the E-step, we compute the expected value of the complete log-likelihood $\ell_{\text{LC}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}, \mathbf{Y}, \mathbf{Z})$ conditional on the observed data $\mathbf{X}$ and $\mathbf{Y}$. The expected value of the latent complete log-likelihood given the observed data forms the Q function in the EM algorithm

$$Q_{\text{LC}}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) := \mathbb{E}[\ell_{\text{LC}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}) | \mathbf{X}, \mathbf{Y}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}].$$

In practice, we do not have access to the true expectation, so we work with the sample analogue $Q_{\text{LC},n}$. The sample analogue can be expressed as

$$\begin{aligned} & Q_{\text{LC},n}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) \\ &:= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log w_k(\mathbf{X}_i; \beta_k) + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log p_k(\mathbf{Y}_i; \theta_k), \end{aligned} \quad (3.4)$$

where

$$\hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) := P(Z = k | \mathbf{X}_i, \mathbf{Y}_i; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) = \frac{w_k(\mathbf{X}_i; \hat{\beta}_k^{(t)}) p_k(\mathbf{Y}_i; \hat{\theta}_k^{(t)})}{\sum_{k'=1}^K w_{k'}(\mathbf{X}_i; \hat{\beta}_{k'}^{(t)}) p_{k'}(\mathbf{Y}_i; \hat{\theta}_{k'}^{(t)})} \quad (3.5)$$

is the weight of observation $(\mathbf{X}_i, \mathbf{Y}_i)$ for the k -th mixture component.

Since the latent-complete log-likelihood function (3.3) decomposes as the sum of a function of $\boldsymbol{\beta}$ and a function of $\boldsymbol{\theta}$, the maximization step of the EM algorithm is separable. Note that since w_k and p_k are logistic regression and binomial product models, respectively, one can apply standard off-the-shelf model fitting procedures after reweighting each observation i by $\hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$. In the maximization step, the new estimate for $\boldsymbol{\theta}$ is updated as follows for

each $k \in [K]$ and $j \in [d]$,

$$\hat{\theta}_{k,j}^{(t+1)} := \frac{\sum_{i=1}^n \frac{Y_{i,j}}{N_j} \hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)}{\sum_{i=1}^n \hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)}, \quad (3.6)$$

which is simply the standard MLE formed by the sample proportion, reweighted by $\hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$.

We update our estimate of β_k by fitting a multi-class logistic regression by reweighting each corresponding observation by $\hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$. This is equivalent to regressing the variable $W^{(t)}$ on the covariates X , where $W_i^{(t)} = (\hat{\pi}_1^{(t)}(\mathbf{X}_i, \mathbf{Y}_i), \dots, \hat{\pi}_K^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)) \in S^{K-1}$ for all i , where S^{K-1} is the $(K-1)$ -th probability simplex. Note that this is not the standard logistic regression because the outcome variable belongs to a probability simplex. This logistic regression can still be fit using gradient descent in standard R packages.

Algorithm 3 Fitting the mixture model via EM

Require: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n, (\hat{\beta}^{(0)}, \hat{\theta}^{(0)})$

- 1: Start with an initial guess $(\hat{\beta}^{(0)}, \hat{\theta}^{(0)})$.
 - 2: Initialize the iteration step $t = 0$.
 - 3: **repeat**
 - 4: Compute the weights $\hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$ for all observations i and mixture components k using (3.5).
 - 5: Form the probability vector $W_i^{(t)} = (\hat{\pi}_1^{(t)}(\mathbf{X}_i, \mathbf{Y}_i), \dots, \hat{\pi}_K^{(t)}(\mathbf{X}_i, \mathbf{Y}_i))$ for each i .
 - 6: Update $\hat{\theta}^{(t+1)}$ using (3.6).
 - 7: Update $\hat{\beta}^{(t+1)}$ by fitting a logistic regression model of $W^{(t)}$ on $\mathbf{X}$.
 - 8: Update $t \leftarrow t + 1$.
 - 9: **until** Convergence
 - 10: Let $\hat{\beta} = \lim_{t \rightarrow \infty} \hat{\beta}^{(t)}$ and $\hat{\theta} = \lim_{t \rightarrow \infty} \hat{\theta}^{(t)}$.
 - 11: **return** $(\hat{\beta}, \hat{\theta})$
-

Algorithm 3 describes the process of fitting the mixture of binomial products experts model in the presence of completely observed covariate and outcome data. For a single initialization point, the EM algorithm is not guaranteed to converge to the global maximizer. In practice, we run Algorithm 3 many times with multiple random initial guesses to explore the parameter space sufficiently and choose the parameter estimate from all of them that

maximizes the log-likelihood.

2.3 Identifiability

We now make some brief comments on the identifiability of our model. Identifiability means that each parameter value corresponds to a unique probability distribution. That is, the mapping from the parameter space to the space of probability distributions is one-to-one. If identifiability does not hold, estimation becomes problematic because a unique MLE may not exist. However, this notion of identifiability may be too strong for many practical purposes. For example, even the common problem of label swapping violates this identifiability definition, so researchers often consider the idea of identifiability up to permutation of the parameters. We now consider the notion of *generic identifiability*, which relaxes the definition of identifiability even further. Generic identifiability means identifiability holds almost everywhere in the parameter space (Allman et al., 2009). More formally, this means that the mapping from the parameter space to the space of probability distributions may fail to be one-to-one only on a set of Lebesgue measure zero. Generic identifiability (holding almost everywhere) can be viewed as an intermediate assumption between two common assumptions: global identifiability (holding everywhere) and local identifiability (holding in a neighborhood of the true parameter). In a practical sense, this means that any such model fit on a given data set is unlikely to be unidentifiable, and we consider this notion to be sufficient for our applied data setting.

Proposition 3.1 (Sufficient conditions for generic identifiability). *Suppose the following conditions hold.*

A1 Each mixture is distinct such that $\theta_k \neq \theta_{k'}$ when $k \neq k'$.

A2 The number of outcome variables d and the number of mixtures K satisfies the bound

$$d \geq 2\lceil \log_{1+\min_j N_j} K \rceil + 1.$$

A3 The logistic regression model $w_k(x; \beta_k)$ for all $k \in [K]$ is identifiable.

Then, the mixture of binomial product experts is generically identifiable up to permutation of the parameters.

We note that the sufficient conditions outlined in Proposition 3.1 are fairly mild. Intuitively, one would not want the number of parameters to be too large when the data dimension d is small. The bound in Assumption A2 ensures that the number of mixture components K remains appropriately controlled relative to d and N_j , as a large K increases model complexity. When there are at least $d = 3$ outcome variables of moderate range, K can still be fairly large. Assumption A3 is reasonable as long as the mixture components are well-separated and the covariates are not collinear. We provide further comments on the assumptions and the proof of this proposition in Appendix B.2.

3 Missingness in the Outcome Variables

Another challenge in the NACC's neuropsychological data is the missingness in the outcomes. To properly describe the missing data problem, we first introduce some additional notation. Let $\mathbf{R}_i = (R_{i,1}, R_{i,2}, \dots, R_{i,d})^\top \in \mathcal{R} \subseteq \{0, 1\}^d$ be the random binary vector that denotes the missing pattern associated with individual i . Element $R_{i,j}$ of the binary vector $\mathbf{R}_i$ is 1 if and only if $Y_{i,j}$ is observed. For a given pattern $r \in \mathcal{R}$, let $\mathbf{Y}_{i,r} = (Y_{i,j} : r_{i,j} = 1)$ denote the observed variables and $\mathbf{Y}_{i,\bar{r}} = (Y_{i,j} : r_{i,j} = 0)$ denote the missing variables. For example, when $d = 4$ and $r = 1001$, then $\mathbf{Y}_{i,r} = (Y_{i,1}, Y_{i,4})$ and $\mathbf{Y}_{i,\bar{r}} = (Y_{i,2}, Y_{i,3})$. Similarly, for observation i with random missing pattern $\mathbf{R}_i$, denote $\mathbf{Y}_{i,\mathbf{R}_i}$ and $\mathbf{Y}_{i,\bar{\mathbf{R}}_i}$ as the observed and missing outcome variables for the i th observation, respectively. We place no restrictions on the set of possible patterns $\mathcal{R}$, allowing the pattern to be nonmonotone. Since $|\mathcal{R}| \leq 2^d$, missingness can easily become an exponentially complex problem in the dimension of the outcome variables.

In contrast to the previous section, we will use the term *fully complete* (FC) to refer to a data set containing IID tuples of the form $(\mathbf{X}, \mathbf{Y}, Z, \mathbf{R})$. We call a data set containing IID observations of the form $(\mathbf{X}, \mathbf{Y}_{\mathbf{R}}, \mathbf{R})$ as the *observed* data. Note that the observed data now has two kinds of incompleteness/missingness: incompleteness in the latent class variable (due to mixture models) and missingness in the outcomes. As the case with the latent complete data, we also do not have access to the fully complete data. We collect all of the observed outcome variables in one tuple $\mathbf{Y}_{1:n, \mathbf{R}_{1:n}} = (\mathbf{Y}_{1, \mathbf{R}_1}, \mathbf{Y}_{2, \mathbf{R}_2}, \dots, \mathbf{Y}_{n, \mathbf{R}_n})$.

3.1 Missing at random and an imputation strategy

Rubin (1976) outlined three types of missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing completely at random assumes that the missingness of the variable is independent of all variables in the data. Missing at random assumes that the missingness of a variable can only depend on observed variables. Missing not at random assumes that the missingness of a variable can depend on the value of the variable subject to missingness.

In the MCAR data, model fitting is straightforward because one can run Algorithm 3 on the complete cases, and the estimates of the parameters will be consistent. However, MCAR assumes that the missingness is irrelevant to observed outcomes, which is clearly false in the NACC data as a possible reason to miss some test scores is that the individual was too sick to finish the test. Therefore, we consider the MAR assumption. Formally, the definition of missing at random (MAR) is as follows.

Definition 3.1 (Missing at random). The outcome variables $\mathbf{Y}$ are missing at random (MAR) if the probability of missingness is dependent only on the variables that are observed for the

given pattern. This assumption writes as

$$P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y}) \stackrel{\text{MAR}}{=} P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y}_r) \quad (3.7)$$

for all $r \in \mathcal{R}$.

Notice that the left-hand side of (3.7) represents the selection probability $P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y})$. This quantity is strictly unidentifiable because it depends on unobserved data, specifically $\mathbf{Y}_{\bar{r}}$. Thus, it cannot be estimated even given infinite observed data. The missing at random assumption makes $P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y})$ identifiable because it equates it to a probability $P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y}_r)$ that can be estimated from the observed data; the variables under the conditioning are all of the variables strictly observed under pattern r . The definition of missing at random implies that the probability of a given missing pattern does not depend on variables that are unobserved under that pattern. A key advantage of this assumption is that we avoid the challenge of modeling the joint distribution between $\mathbf{Y}$ and $\mathbf{R}$ since the missingness mechanism $P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y})$ does not need to be modeled directly; this is known as the *ignorability property* (see the discussion later). Thus, we do not have to deal with making potentially unreasonable modeling assumptions on either the selection model $P(\mathbf{R} = r|x, y)$ or the extrapolation distribution $p(y_{\bar{r}}|y_r, x, \mathbf{R} = r)$.

The MAR assumption is untestable because its validity relies strictly on data that is unobserved, and so it cannot be rejected by the observed data. Our primary reason for selecting this assumption is for modeling. This may lead to easier interpretability for scientists and practitioners since we fit a global model $p(y|x)$ across all missing patterns rather than a local model $p(y|x, R = r)$ for every pattern r . In the Alzheimer's disease literature, having more statistically sound work can be meaningful because of limitations in existing work that we have described previously in Section 1.3. Since we believe that it may be plausible because

these tests are correlated due to underlying cognitive ability, we can use it as a starting point. This can be viewed as a baseline before proceeding with more complex MNAR assumptions. We recognize that MNAR may be more reasonable since missing test scores may be attributed to sickness. However, this is in itself a very open research question because MNAR is a broad class of assumptions, and performing mixture modeling with MNAR is not straightforward; we leave this for future work.

An additional feature of the MAR assumption is that this assumption offers a simple approach to impute the data, which makes the computation of the MLE a lot easier. Before describing the procedure of updating model parameters, we first introduce a multiple imputation procedure in Algorithm 4 that can be used to fill in the missing data. In the algorithm, for notational convenience, when we write a binary vector in the summation or product, this means we iterate over all indices whose elements are nonzero. For instance, when we write “For j in $(1, 0, 0, 1)$,” this is equivalent to “For $j = 1, 4$.” Therefore, if $\mathbf{R}_i = (1, 0, 0, 1)$, then “For j in $\mathbf{R}_i$ ” corresponds to “For $j = 1, 4$ ” as well.

As stated, Algorithm 4 describes how to impute the outcome variables, assuming that a good estimate $(\hat{\beta}, \hat{\theta})$ is available. We discuss how to actually obtain such an estimate in the Section 3.2. The multiple imputation algorithm exploits the fact that the conditional distribution $p(y_{\bar{r}}|y_r, x)$ for any $r \in \mathcal{R}$ remains a mixture model. There are two steps to performing multiple imputation: 1) for every observation i with missing observations, we compute the weights of each component of the mixture distribution $p(y_{\bar{\mathbf{R}}_i}|\mathbf{Y}_{i,\mathbf{R}_i}, \mathbf{X}_i)$, and 2) sample M times from the distribution $p(y_{\bar{\mathbf{R}}_i}|\mathbf{Y}_{i,\mathbf{R}_i}, \mathbf{X}_i)$ for each observation i to form M completed data sets. The derivation of this procedure is provided in Appendix B.1.2.

Algorithm 4 Multiple imputation

Require: $\{(\mathbf{X}_i, \mathbf{Y}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n, (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}), M$ (the number of imputed data sets)

```

1: for  $i = 1, \dots, n$  do
2:   if  $\mathbf{R}_i \neq 1_d$  then
3:     for  $k = 1, \dots, K$  do
4:       Form the weights  $\omega_{i,k} = w_k(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_k) p_{k,\mathbf{R}_i}(\mathbf{Y}_{i,\mathbf{R}_i}; \hat{\boldsymbol{\theta}}_k) / \sum_{k'} w_{k'}(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_{k'}) p_{k',\mathbf{R}_i}(\mathbf{Y}_{i,\mathbf{R}_i}; \hat{\boldsymbol{\theta}}_{k'})$ 
5:     end for
6:   end if
7: end for
8: for  $m = 1, \dots, M$  do
9:   for  $i = 1, \dots, n$  do
10:    if  $\mathbf{R}_i \neq 1_d$  then
11:      Sample  $Z_i^{(m)} \sim \text{Categorical}(K, p = (\omega_{i,1}, \dots, \omega_{i,K}))$ 
12:      for  $j \in \bar{\mathbf{R}}_i$  do
13:        Sample  $\widetilde{Y}_{i,j}^{(m)}$  from  $\text{Binomial}\left(N_j, \hat{\theta}_{Z_i^{(m)},j}\right)$ .
14:      end for
15:      Assign  $\widetilde{\mathbf{Y}}_i^{(m)} := (\mathbf{Y}_{i,\mathbf{R}_i}, \widetilde{\mathbf{Y}}_{i,\bar{\mathbf{R}}_i}^{(m)})$ .
16:    end if
17:  end for
18:  Let  $\widetilde{D}_m = \{(\mathbf{X}_i, \widetilde{\mathbf{Y}}_i^{(m)}, \mathbf{R}_i)\}_{i=1}^n$ .
19: end for
20: return  $\{\widetilde{D}_m\}_{m=1}^M$ 

```

3.2 Model fitting under a missing at random assumption

For each r , we assume that the selection probability $P(\mathbf{R} = r | \mathbf{X}, \mathbf{Y}_r; \gamma_r)$ belongs to a parametric family, indexed by γ_r . We collect these parameters into $\gamma = (\gamma_r)_{r \in \mathcal{R}}$. For simplicity, we write the log-likelihood in terms of the probability model without the n samples. Under the MAR assumption, we can write the observed log-likelihood as

$$\begin{aligned} \ell_{\text{obs}}(\gamma, \beta, \theta; x, y_r, r) &= \log P(\mathbf{R} = r | x, y_r; \gamma) + \log p(y_r | x; \beta, \theta) \\ &= \ell_{\text{obs}}^{(1)}(\gamma; x, y_r, r) + \ell_{\text{obs}}^{(2)}(\beta, \theta; y_r, x). \end{aligned}$$

The missingness mechanism is said to be ignorable because estimation of (β, θ) is separated from the estimation of γ . Model fitting under a missing at random assumption typically occurs using an EM algorithm approach. We can proceed with estimating (β, θ) by conditioning on the observed data $(\mathbf{X}, \mathbf{Y}_{\mathbf{R}})$ and using another EM algorithm approach. The population $Q_{\text{FC},r}$ function writes as follows

$$\begin{aligned} Q_{\text{FC},r}(\beta, \theta | \mathbf{X}, \mathbf{Y}_r; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}) &:= \mathbb{E}[\ell_{\text{LI}}(\beta, \theta; \mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}_r; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}] \\ &= \sum_{y_{\bar{r}}} \ell_{\text{LI}}(\beta, \theta; \mathbf{X}, y_{\bar{r}}, \mathbf{Y}_r) p(y_{\bar{r}} | \mathbf{Y}_r, \mathbf{X}; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}) \end{aligned} \quad (3.8)$$

for every $r \in \mathcal{R}$. One major observation is that the conditional expectation relies on being able to fit estimate $p(y_{\bar{r}} | y_r, x; \beta, \theta)$. We can leverage the consequences of the local independence assumption to obtain an easily computable form of the conditional distribution. Unfortunately, since the latent-incomplete log-likelihood function ℓ_{LI} is not linear in the outcome variables, we are unable to reduce (3.8) to a more simple form. However, we can approximate this expectation stochastically by imputing the missing data $\mathbf{Y}_{\bar{r}}$ for every missing pattern r using the distribution $p(y_{\bar{r}} | y_r, x; \hat{\beta}^{(t)}, \hat{\theta}^{(t)})$.

To overcome estimating the conditional expectation of factorial terms, we approximate (3.8) stochastically with a Monte Carlo procedure. For large enough M , we expect that

$$\begin{aligned} Q_{\text{FC},r}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}_r; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) &\approx Q_{\text{FC},r,n}^{(M)}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n, \mathbf{R}_{1:n}}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) \\ &:= \frac{1}{Mn} \sum_{m=1}^M \sum_{i=1}^n \log \left(\sum_{k=1}^K w_k(\mathbf{X}_i; \beta_k) p_k(\tilde{\mathbf{Y}}_i^{(m;t)}; \theta_k) \right) \cdot 1(\mathbf{R}_i = r), \end{aligned}$$

where $\tilde{\mathbf{Y}}_i^{(m;t)} = (\mathbf{Y}_{i, \mathbf{R}_i}, \tilde{\mathbf{Y}}_{i, \bar{\mathbf{R}}_i}^{(m;t)})$ denotes the m th imputed data for the i th observation on the iteration step t given the observed variables. The choice of number of imputations M is important, and in practice, we use $M = 20$ to balance both computational time and good estimation performance. We discuss this in more detail with our simulations in Section 6 and Appendix B.5.2.

In the maximization step, we compute the MLE of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ using the completed data after the multiple imputation. Since we have completed data, the MLE can be found using the original EM algorithm, outlined in Algorithm 3. We summarize the procedure in the following algorithm. Throughout this chapter, we will use the notation $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ (without superscripts relating to t) to denote the point estimate obtained by Algorithm 5 after convergence is achieved. In practice, we assume that convergence is achieved if the L_2 norm of the difference between the current and old parameter estimates falls within a prespecified tolerance level (we choose $\epsilon = 10^{-4}$).

Algorithm 5 is a nested EM procedure because we have two types of missingness: missingness in the outcome variables and missingness in the latent class labels. We embed an EM algorithm for the latent class model fitting (this serves as the M-step in the outer EM algorithm) within an outer EM algorithm for handling the missing at random data. Because we do not obtain a closed-form expression for the conditional expectations in the E-step but rather perform a stochastic approximation, Algorithm 5 is a Monte Carlo EM algorithm

Algorithm 5 Fitting the mixture model with MAR outcomes via a nested EM procedure

Require: $\{(\mathbf{X}_i, \mathbf{Y}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n, (\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}), M$

- 1: Start with an initial guess $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)})$.
 - 2: Initialize the iteration step $t = 0$.
 - 3: **repeat**
 - 4: Let $\tilde{D}_t = \left\{ \left(\mathbf{X}_i, \tilde{\mathbf{Y}}_i^{(m;t)} \right) \right\}_{i=1, \dots, n; m=1, \dots, M}$ be the stacked imputed data sets obtained from Algorithm 4.
 - 5: Obtain $(\hat{\boldsymbol{\beta}}^{(t+1)}, \hat{\boldsymbol{\theta}}^{(t+1)})$ using Algorithm 3 on the imputed data set $\tilde{D}_t$ and starting point $(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)})$.
 - 6: Update $t \leftarrow t + 1$.
 - 7: **until** Convergence
 - 8: Let $\hat{\boldsymbol{\beta}} = \lim_{t \rightarrow \infty} \hat{\boldsymbol{\beta}}^{(t)}$ and $\hat{\boldsymbol{\theta}} = \lim_{t \rightarrow \infty} \hat{\boldsymbol{\theta}}^{(t)}$.
 - 9: **return** $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$
-

(Levine and Casella, 2001). We use the notation $\tilde{\mathbf{Y}}_i^{(m;t)}$ to denote the m th imputed outcome variables using the model parameterized by $(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)})$. When the parameter is understood, we omit the t index. Note that when there is no missingness in the outcome variables, Algorithm 5 reduces to Algorithm 3 because the multiple imputation step is bypassed. Again, in practice, we run Algorithm 5 with multiple random initial points to ensure we explore the parameter space and converge to the global maximizer. We pick the point estimate that maximizes the observed log-likelihood, defined as

$$\ell_{\text{obs},n} := \ell_{\text{obs},n}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_{1:n}, \mathbf{Y}_{1:n, \mathbf{R}_{1:n}}) := \sum_{i=1}^n \log \left(\sum_{k=1}^K w_k(\mathbf{X}_i) p_{k, \mathbf{R}_i}(\mathbf{Y}_{i, \mathbf{R}_i}) \right), \quad (3.9)$$

where $p_{k,r}(y_r) = \prod_{j \in r} p_{k,j}(y_j)$ for any $r \in \mathcal{R}$. Note that when there is no missing data, this reduces to the latent incomplete log-likelihood function $\ell_{\text{LI},n}$. Since we are working under a parametric model and estimating parameters using maximum likelihood, the estimators converge at $\sqrt{n}$ -rate and are asymptotically efficient. We provide theoretical justification for these methods in Appendix B.1.

4 Inference

In this section, we discuss how to perform inference on the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. There has been some previous work in obtaining asymptotic variance estimators for multiple imputations estimators (Wang and Robins, 1998b; Robins and Wang, 2000), but this can be analytically challenging in our setting. The bootstrap is a widely adopted procedure for estimating the sampling distribution of an estimator and obtaining asymptotically valid confidence intervals (Efron, 1979) and has been previously used with the EM algorithm (Celeux and Diebolt, 1987). O’Hagan et al. (2019) have also explored the use of the bootstrap with a Gaussian mixture model. We provide theoretical justification for the bootstrap in our setting in Appendix B.8.3.

Bootstrapping works by resampling from the original data set (which is equivalent to sampling from empirical distribution), and with large enough sample size and under sufficient regularity conditions, this mimics drawing samples from the true generating distribution. We bootstrap the data B times (where B is sufficiently large) including the missing data, and the estimation procedure is run on each of the B bootstrapped data sets.

Since our EM algorithm requires many initialization points in practice to properly explore the parameter space, there is a question of how to initialize the bootstrap procedure. We follow the recommendation outlined in Chen (2022b) to initialize the bootstrap at the same initial point on every iteration, using the point estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ returned from Algorithm 5. The primary goal of the bootstrap is to measure the stochastic variation of an estimator around the parameter of interest. If we perform the bootstrap with random initialization, we will capture additional uncertainty that arises from estimating different local modes of the log-likelihood function. Initializing at the same point also avoids the label switching identifiability problem, which occurs when the probability distribution remains identical after some parameters are permuted. Additionally, this saves on computational time because we

are also not performing many random initializations. Our bootstrap procedure is summarized in the Algorithm 6.

Algorithm 6 Bootstrap procedure for obtaining confidence intervals

Require: $\{(\mathbf{X}_i, \mathbf{Y}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n$, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$, B (a large number, say 1,000)

- 1: **for** $b \in 1, \dots, B$ **do**
 - 2: Sample n draws uniformly with replacement from $\{1, 2, \dots, n\}$. Put these into index set I_b .
 - 3: Set the b th bootstrapped data set $D_b^* = \{\mathbf{X}_i, \mathbf{Y}_{i,\mathbf{R}_i}, \mathbf{R}_i\}_{i \in I_b}$.
 - 4: Obtain $(\boldsymbol{\beta}^{*(b)}, \boldsymbol{\theta}^{*(b)})$ using Algorithm 5 on data set D_b^* initialized with $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$.
 - 5: **end for**
 - 6: **return** $\{\boldsymbol{\beta}^{*(b)}\}_{b=1}^B, \{\boldsymbol{\theta}^{*(b)}\}_{b=1}^B$
-

We now describe how to construct confidence intervals for a given parameter, using $\theta_{1,1}$ as an illustrating example. We first obtain a point estimate $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ using Algorithm 5. Then, we run Algorithm 6 using $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ as the initialization point for a large number of iterations B . We estimate the variance of $\theta_{1,1}$ using the bootstrap samples via the sample variance of $\{\theta_{1,1}^{*(b)}\}_{b=1}^B$

$$\widehat{V} = \frac{1}{B-1} \sum_{b=1}^B \left(\theta_{1,1}^{*(b)} - \frac{1}{B} \sum_{b'=1}^B \theta_{1,1}^{*(b')} \right)^2.$$

We form a 95% confidence interval using a normal approximation

$$(\hat{\theta}_{1,1} - 1.96 \cdot \sqrt{\widehat{V}}, \hat{\theta}_{1,1} + 1.96 \cdot \sqrt{\widehat{V}}).$$

Other approaches such as the percentile method described in Chen (2022b) will also be equivalent asymptotically since the bootstrap distribution will be asymptotically normal. The key advantage of this method is that it avoids the difficulty of obtaining a closed form expression for the Fisher information matrix, which can be difficult for complex mixture models. The process of the bootstrap is similar to that of attaining a point estimate because

we run Algorithm 5 repeatedly many times. However, unlike obtaining a point estimate, we use the same initialization point on many different data sets (bootstrap samples) rather than many different initialization points on the same data set. White and Murphy (2014) have also discussed how bootstrap methods can struggle when parameters are close to the boundary, so we recommend that post-hoc visualizations are performed to ensure that label-switching has not occurred. In our data analysis and simulations, the mixture components are well-separated, and in practice, we order the groups based on the parameter estimates to avoid the label switching problem.

5 Cluster Analysis

A feature of the mixture model that we propose is that we are able to cluster individuals into different groups according to the predictive probability. This allows us to discover latent groups inside our data. We assume that the mixture of binomial product experts model has already been fit successfully using Algorithm 5, and we have an estimator $(\hat{\beta}, \hat{\theta})$.

Each mixture component can be thought of as a cluster, so the method we propose can be used to perform model-based clustering (Bouveyron et al., 2019). To simplify the problem, we first consider the case where we do not have any missing outcome variables. That is, the data we have is only latent incomplete. For each observation i , define the probabilities as follows

$$\hat{\pi}_k(\mathbf{X}_i, \mathbf{Y}_i) := P(Z = k | \mathbf{X}_i, \mathbf{Y}_i; \hat{\beta}, \hat{\theta}) = \frac{w_k(\mathbf{X}_i; \hat{\beta}_k) p_k(\mathbf{Y}_i; \hat{\theta}_k)}{\sum_{k'=1}^K w_{k'}(\mathbf{X}_i; \hat{\beta}_{k'}) p_{k'}(\mathbf{Y}_i; \hat{\theta}_{k'})}. \quad (3.10)$$

The quantity $\hat{\pi}_k(\mathbf{X}_i, \mathbf{Y}_i)$ is the predictive probability that individual i is from component k . We compute the probability of belonging to each class given the data $\mathbf{X}$ and $\mathbf{Y}$, leading to a probability vector $\hat{\pi}(\mathbf{X}_i, \mathbf{Y}_i) = (\hat{\pi}_1(\mathbf{X}_i, \mathbf{Y}_i), \dots, \hat{\pi}_K(\mathbf{X}_i, \mathbf{Y}_i)) \in \mathbb{R}^K$. It is straightforward to create a cluster assignment of each individual by assigning each observation to the component

with the maximum probability. Namely, we assign individual i to the cluster with

$$\hat{C}_i = \operatorname{argmax}_k \hat{\pi}_k(\mathbf{X}_i, \mathbf{Y}_i).$$

In the context of the neuropsychological test scores, we can interpret the latent class as a measure of latent cognitive ability over multiple cognitive domains. An individual's cognitive ability is a complex summary of many different attributes that we hope to measure using the neuropsychological test scores. The latent class can be predicted using the test scores and the baseline covariates.

When the data is missing, we perform clustering using the observed data. For any $r \in \mathcal{R}$, note that there is a closed form expression for $P(Z = k | \mathbf{X}_i, \mathbf{Y}_r)$. Define the following quantity

$$\hat{\pi}_{k,r}(\mathbf{X}_i, \mathbf{Y}_{i,r}) := P(Z = k | \mathbf{X}_i, \mathbf{Y}_{i,r}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \frac{w_k(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_k) p_{k,r}(\mathbf{Y}_{i,r}; \hat{\boldsymbol{\theta}}_k)}{\sum_{k'=1}^K w_{k'}(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_{k'}) p_{k',r}(\mathbf{Y}_{i,r}; \hat{\boldsymbol{\theta}}_{k'})}.$$

For observation i and every $k \in [K]$, we compute $\hat{\pi}_{k,\mathbf{R}_i}(\mathbf{X}_i, \mathbf{Y}_{i,\mathbf{R}_i})$. This quantity can be interpreted as the estimated probability of arising from component k given the data $\mathbf{X}_i$ and $\mathbf{Y}_{i,\mathbf{R}_i}$. When there are missing outcome variables, we compute these probabilities explicitly and use them to perform clustering. We summarize our overall proposed methodology in Figure 3.1. From the observed data, we run Algorithm 5 (which comprises Algorithms 3 and 4) to obtain a point estimate. We use the point estimate and the observed data in the Algorithm 6 to construct confidence intervals using the bootstrap. Cluster assignments can also be obtained using the point estimates.

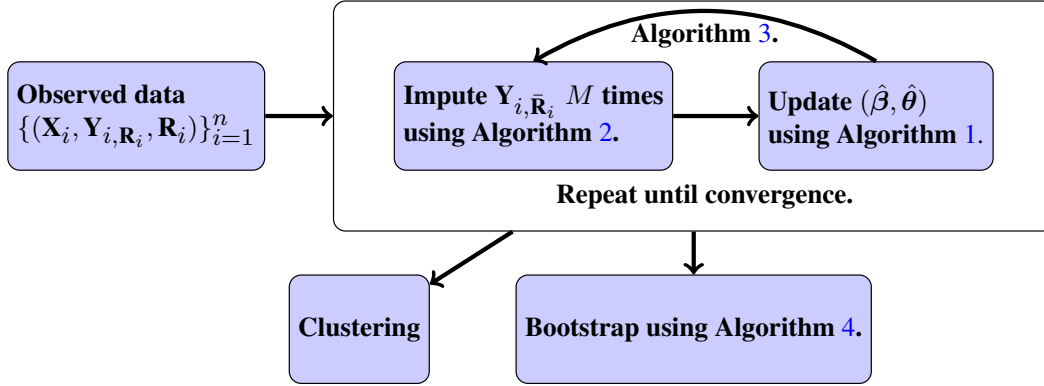


Figure 3.1: This flowchart describes overall inference procedure.

6 Simulations

We now examine the performance of our method in finite samples using simulated data. We have $p = 2$ covariates, $d = 4$ outcome variables, and $K = 3$ mixtures. Note that $\eta = 2$, $\eta = 2.5$, and $\eta = \infty$ correspond to approximately 75%, 85%, and 100% complete cases. We generate the data through a mixture of binomial product experts model $p(y|x)$, and the parameter η affects the strength of the missingness via the selection model $P(R = r|x, y)$. Further details on the data generating process and the role of η are described in Appendix B.5.1. Each of the simulation studies is evaluated using the mean-squared error $\text{MSE}_{\theta} = \mathbb{E}[\|\hat{\theta} - \theta\|_2]$ and $\text{MSE}_{\beta} = \mathbb{E}[\|\hat{\beta} - \beta\|_2]$. For the simulations in Table 8, we estimate the mean-squared error by simulating $U = 1000$ data sets in each scenario and average the L_2 norm of the error over all the data sets. The estimated MSEs are calculated as follows

$$\widehat{\text{MSE}}_{\beta} = \frac{1}{U} \sum_{i=1}^U \sum_{k \neq 1} \|\hat{\beta}_k^{(i)} - \beta_k\|_2, \quad \widehat{\text{MSE}}_{\theta} = \frac{1}{U} \sum_{i=1}^U \sum_k \|\hat{\theta}_k^{(i)} - \theta_k\|_2, \quad (3.11)$$

where the superscript (i) denotes the estimated parameter of the i th generated data set. We also construct 95% confidence intervals for each parameter and report the coverage over all

Table 3.1: These results are the estimated MSEs and estimated coverage after imputing $M = 20$ times from our model and running Algorithms 5 and 6 for $U = 1000$ replicates. The first row contains results for the estimation of θ . The second row contains results for the estimation of β .

MSE$_{\theta}(\times 100)$				Estimated Coverage of θs			
	Sample Size				Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$		$n = 500$	$n = 1000$	$n = 2000$
$\eta = \infty$	0.203	0.099	0.05	$\eta = \infty$	0.946	0.948	0.950
$\eta = 2.5$	0.192	0.093	0.047	$\eta = 2.5$	0.943	0.943	0.945
$\eta = 2$	0.203	0.104	0.054	$\eta = 2$	0.939	0.939	0.932

MSE$_{\beta}(\times 100)$				Estimated Coverage of βs			
	Sample Size				Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$		$n = 500$	$n = 1000$	$n = 2000$
$\eta = \infty$	12.7	6.05	3.08	$\eta = \infty$	0.949	0.951	0.947
$\eta = 2.5$	41.5	20.6	10.5	$\eta = 2.5$	0.952	0.951	0.944
$\eta = 2$	40.5	20.0	10.1	$\eta = 2$	0.956	0.951	0.950

Table 3.2: These results are the estimated MSEs after imputing $M = 20$ times using *mice* for $U = 200$ replicates. The left table contains results for the estimation of θ . The right table contains results for the estimation of β .

MSE$_{\theta}(\times 100)$				MSE$_{\beta}(\times 100)$			
	Sample Size				Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$		$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.183	0.096	0.045	$\eta = 2.5$	39.4	19.0	10.2
$\eta = 2$	0.194	0.105	0.045	$\eta = 2$	46.0	19.6	11.6

the $U = 1000$ data sets. We check convergence of Algorithms 3 and 5 by calculating the L_2 norm of the difference between the current and old parameter estimates and comparing against a set tolerance of $\epsilon = 10^{-4}$. In the first column of Table 8, for a given η , we observe that the estimated MSE decreases at a roughly linear rate as we increase the sample size, which is expected in the parametric setting. This demonstrates that the estimation procedures for both the no missing data (Algorithm 3) and missing data cases (Algorithm 5) are consistent. The MSE for β is higher when there is missingness than when there is no missingness, which suggests that the estimating β becomes more difficult.

In the second column of Table 8, we generally observe nominal coverage as well. There is slight undercoverage of θ when $\eta = 2$. We attribute this to the fact that we do not have a closed-form expression for the MLE and there are many approximations in the algorithms. For example, Algorithm 5 is a stochastic EM algorithm because it relies on a prespecified number of imputations M . Increasing the number of imputations will decrease the Monte Carlo error at the cost of computational time. It is possible that the gradient of the log-likelihood surface can be fairly small in the neighborhood of the true MLE, leading to early stopping of the algorithms and returning a near-solution but not the true MLE. Moreover, when there is missing data, the effective sample size is lower than the number of observations n , which can be another reason for the slight undercoverage in finite samples. For most cases, the coverage is close to nominal, which suggests that the bootstrap procedure and the choice of initializing it at the original estimate returned by Algorithm 5 works in practice.

For an additional comparison, we conduct a second simulation using the same data generating process. We first perform imputation using the *mice* R package (van Buuren and Groothuis-Oudshoorn, 2011a), generating $M = 20$ imputed datasets. We then fit our model using Algorithm 3 and compute the MSEs. The results are reported in Table 3.2.

All simulations in this section were run on a cluster with 48 cores and a parallelized

Table 3.3: For the simulations in Tables 8 and 3.2, we report the average **computation time in minutes** and the standard deviation in parentheses across all the randomly generated data sets. These were performed on a cluster using 48 cores. The first table contains the times for obtaining both a point estimate and a confidence interval using Algorithms 5 and 6. The second table contains the times for only obtaining a point estimate with *mice* and Algorithm 3.

Algorithms 5 and 6 (Point Estimate and Bootstrap CI)				<i>mice</i> and Algorithm 3 (Point Estimate)			
	Sample Size				Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$		$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	2.0 (2.1)	2.8 (0.5)	5.5 (1.1)	$\eta = 2.5$	8.1 (0.91)	14.2 (0.83)	28.2 (3.8)
$\eta = 2$	2.4 (1.9)	3.9 (3.5)	6.2 (1.9)	$\eta = 2$	8.1 (0.76)	14.5 (0.89)	27.4 (2.1)

implementation. Our first observation is that this two-step approach—imputing with *mice* followed by Algorithm 3—is substantially slower than Algorithms 5 and 6. As a result, we limit the *mice* experiment to 200 total repetitions for estimating the MSE and do not report coverage results since bootstrap-based interval estimation would be prohibitively time-consuming. For reference, obtaining *both* a point estimate and a bootstrapped confidence interval with our method requires approximately 3 to 6 times less computation time than obtaining *only* a point estimate after imputing with the *mice* method. This can be seen in Table 3.3.

Interestingly, we observe that the MSEs from the *mice* imputation and Algorithm 3 pipeline are comparable to those from our method. However, we believe this is largely due to the well-separated mixture components in the data generating process. In general, we do not recommend using *mice* in this context, as its imputation model may not be compatible with the mixture model being fit, and may not lead to asymptotically valid estimates in more challenging settings. Model uncongeniality is a known problem in the imputation literature (Meng, 1994). In contrast, our method imputes from the same model that is estimated and fit, ensuring consistency.

Further simulations on the complete case data and the influence of the number of imputations are also included in Appendix B.5.2. When fitting the model on complete case data, we observe that the MSE does not decrease at a linear rate due to bias, and that the confidence intervals are not asymptotically valid. Moreover, the results appear fairly stable as we vary the number of imputations M across 10, 20, and 50, suggesting that 10 imputations are already sufficient. To ensure better accuracy of our estimates while maintaining computational efficiency, we use $M = 20$ imputations in the real data analysis.

7 *Application to the NACC data*

As mentioned earlier, the data set that motivates our model is from the National Alzheimer’s Coordinating Center (NACC) Uniform Data Set, collected from the years 2005 to 2019. This is a longitudinal data set that comprises individuals of varying cognitive status: cognitively normal to mild cognitive impairment (MCI) to dementia. Neuropsychological test scores are typically assessed annually with some baseline covariates collected upon entry to the study. However, there is also some missingness present in the outcome variables. The missingness can be due to various reasons. For example, some tests may no longer be administered after a given time because they have been replaced with another one. On the other hand, some individuals may have test scores missing because of a data recording error or they are too sick to take more tests.

7.1 *Description of outcome variables and covariates*

Our primary goal is to measure the cognitive ability of the Alzheimer’s disease patients while measuring its association with baseline covariates. Following similar analysis in [Brenowitz et al. \(2017\)](#), we focus on four cognitive domains: long-term memory (episodic memory), attention, language, and executive function. Specifically, we used the Logical Memory Story A immediate and delayed recall to assess memory, Digit Span Forward and Backward tests to evaluate attention, Animal listing test to measure language ability, and Trail Making Tests Parts A and B to measure executive function. Additionally, we used the Mini-Mental State Examination (MMSE) to assess overall cognitive impairment, resulting in a total of eight outcome variables. The Trail A and B tests are assessed based on the time it takes to complete a given task, so a higher score is considered worse. On the other hand, for the other tests, higher scores are indicative of stronger cognitive ability.

We include four baseline covariates in this study: age, education, sex, and race. Age was

kept on a yearly scale, and education was dichotomized based on whether an individual had obtained a college degree. Of particular interest was the association between education and cognitive decline. This analysis was motivated by the cognitive reserve hypothesis, which suggests that some mechanisms can provide individuals more resilience to cognitive decline. We would like to explore this hypothesis to see if individuals with higher levels of education are more resistant to cognitive impairment (Meng and D’Arcy, 2012; Thow et al., 2018).

For each individual, we have a CDR (Clinical Dementia Rating) score, which is assigned by a medical professional. The CDR score is based on a clinician’s judgment on an individual’s cognitive ability and is the standard procedure to determine whether someone has dementia or not. A CDR score 0 refers to cognitively normal, a score 0.5 refers to a mild cognitive impairment (MCI) state, and a score of at least 1 refers to dementia. Note that the CDR score is all based on a clinician’s judgment, which is very different from the neuropsychological tests (exam-based). This score takes values in the set $\{0, 0.5, 1, 2, 3\}$. Here, 0 indicates normal cognitive ability, 0.5 is mild cognitive impairment (MCI), 1 is mild dementia, 2 is moderate dementia, and 3 is severe dementia. Thus, we expect that a healthy group would contain individuals with lower CDR scores than an unhealthier group. We will use the CDR score to interpret some of our results, but we emphasize that we *do not* use this variable in the model.

Since there are 8 outcome variables, there are up to $2^8 = 256$ missing data patterns. In this data set, we observe 93 missing data patterns with 7 patterns comprising over 95% of the data. In particular, there are $n_{cc} = 25041$ fully observed cases and $n = 41181$ total observations, so that approximately only 60% of the individuals have completely observed outcomes.

In Figure 3.2, we plot the CDR distribution for the complete cases, the individuals missing at least one outcome variable, and the entire data set, respectively. If the data was missing

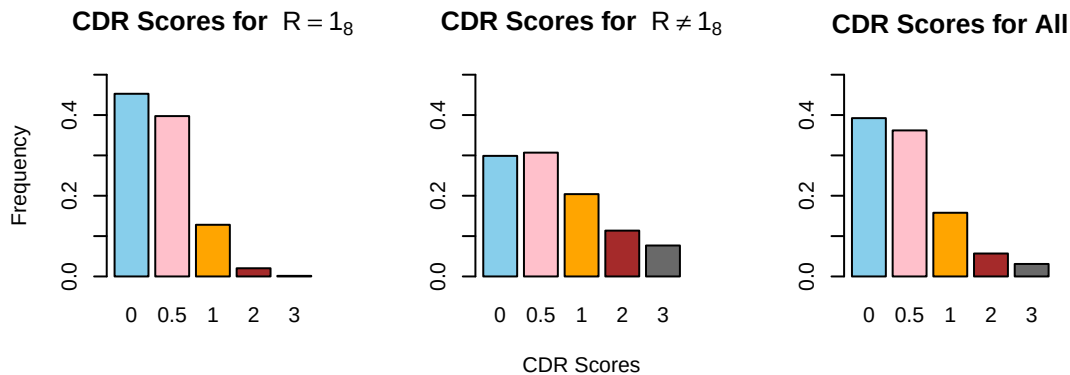


Figure 3.2: The CDR score distributions for the complete cases, the individuals missing at least one outcome variable, and the entire data set are provided in the left, middle, and right panels, respectively.

completely at random, we would expect the CDR score distribution to be mostly constant across all missing patterns of the data. However, we observe that for the complete cases, the dementia group (CDR score of 1, 2, or 3) makes up less than 20%, but the dementia group is close to 40% of the data that has at least one missing outcome variable. This suggests that the dementia group is severely underrepresented when we only perform analysis on the complete cases. Thus, from the plots, we can visually postulate the MCAR assumption to be unreasonable.

We select the model using prior knowledge. Since the scientific community previously determined that five levels for the CDR score was appropriate, we also choose the same number. A similar number of groups will help with interpretability, and five is small enough that the number of parameters is manageable.

7.2 Point estimates and confidence intervals

The data analysis was conducted on a 2022 Macbook Air with 8 cores, and it took about a day to fit the model and complete the bootstrap with 1000 bootstrap samples. In Table 3.4,

we report the point estimates of the test score means and the corresponding standard errors in parentheses. There are a few takeaways. As reported, Classes 1 and 5 can be interpreted as the most and least healthy individuals, respectively. Note that in 6 out of 8 tests (omitting TRAIL A and TRAIL B), a higher score corresponds to better performance. On the other hand, TRAIL A and TRAIL B are scored using time-to-completion in seconds, so a higher score corresponds to a worse performance. For every test aside from TRAIL B, we observe a clear monotonic behavior in the test score means from Class 1 to Class 5, which suggests that the model is reasonable and fairly interpretable. For TRAIL B, Classes 4 and 5 have very similar mean test scores that are close to the maximum possible test score of 300. Since a test score of 300 indicates that the individual timed out and did not complete the test, this suggests that individuals from Classes 4 and 5 have comparable performance on the TRAIL B test.

Additionally, the results also suggests the different tests may have value at distinguishing individuals from different latent cognitive classes. For example, TRAIL A and TRAIL B are routinely used to measure executive function. Classes 4 and 5 represent the most unhealthy individuals in the population, but individuals in both perform similarly on TRAIL B. On the other hand, the mean test scores in Classes 4 and 5 for TRAIL A are very different, which suggests that TRAIL A can be a good discriminator between very unhealthy and moderately unhealthy individuals. As TRAIL B is a more complex task than TRAIL A, this seems to agree with our intuition.

In Table 3.5, we report the point estimates of the coefficients in the logistic regressions. One initial observation is that the coefficients of Age and Education are all positive and negative at the $\alpha = 0.05$ significance level, respectively, across all classes. This implies that increasing age is associated with lower cognitive ability and higher education may provide some protection against dementia. The latter is known as the cognitive reserve hypothesis in

Table 3.4: These tables contain point estimates of the test score means for each latent class. In the title of each column, we report the maximum possible score for a given test in parentheses. In each cell, the standard errors are reported in parentheses.

Outcome Variables					
		MMSE (30)	LOGIMEM (25)	MEMUNITS (25)	DIGIF (12)
Class	1	28.9 (0.01)	13.4 (0.04)	12.0 (0.05)	8.8 (0.02)
	2	26.8 (0.03)	9.1 (0.05)	7.1 (0.06)	7.7 (0.02)
	3	25.3 (0.06)	7.6 (0.08)	5.5 (0.08)	7.1 (0.03)
	4	22.5 (0.06)	5.3 (0.05)	3.5 (0.05)	6.6 (0.03)
	5	14.4 (0.12)	3.1 (0.05)	1.8 (0.04)	5.3 (0.04)
		DIGIB (12)	ANIMALS (77)	TRAILA (150)	TRAILB (300)
Class	1	7.2 (0.05)	20.9 (0.05)	28.4 (0.09)	65.9 (0.2)
	2	5.7 (0.02)	15.9 (0.05)	44.2 (0.17)	117.9 (0.43)
	3	4.9 (0.03)	13.5 (0.07)	57.2 (0.39)	209.7 (0.84)
	4	4.2 (0.02)	11.0 (0.06)	66.7 (0.36)	298.5 (0.08)
	5	2.7 (0.03)	6.9 (0.07)	145.3 (0.20)	295.2 (0.09)

Table 3.5: This table contains point estimates for the coefficients for each of the covariates and each latent class. In each cell, the standard errors are reported in parentheses.

		Covariates			
		Age	Education	Race	Sex
Class	2	0.06 (0.001)	-0.62 (0.028)	-0.37 (0.026)	0.72 (0.037)
	3	0.07 (0.002)	-0.9 (0.039)	0.45 (0.034)	0.93 (0.043)
	4	0.06 (0.002)	-1.06 (0.035)	0.36 (0.032)	0.79 (0.04)
	5	0.05 (0.002)	-1.14 (0.0041)	0.41 (0.0035)	0.93 (0.0041)

the context of education (Meng and D'Arcy, 2012; Thow et al., 2018). We can compare the magnitudes of the coefficient of Education to the coefficient of Age. For example, in Class 2, we observe $0.62/0.06 \approx 10$, and in Class 5, we observe $1.14/0.05 \approx 23$. This suggests that having a college degree or higher education may be equivalent to being 10-20 years younger in terms of cognitive ability.

7.3 *Latent classes and clustering*

We report two clustering results using models fit on the complete cases (Figure 3.3) and the entire data (Figure 3.4). For each set of clustering results, we visualize the clusters by plotting the CDR score composition of each of the five clusters via barplots. Furthermore, we report the mean of the CDR scores for each group, which gives a rough estimate on the group-specific clinical cognitive ability. From the complete case results (CC) in Figure 3.3, we generally see that the first clusters are primarily composed of healthy individuals as individuals with CDR scores of 0 take up larger proportions. As we progress to the middle latent classes, healthier individuals take up fewer proportions and individuals with mild cognitive impairment (CDR score of 0.5) take up larger proportions. Then, in Group 5, the majority of the individuals have some form of dementia.

For the clustering that used the entire data with the missing at random assumption (Figure 3.4), we observe a similar trend with the healthier individuals, the MCI individuals, and the dementia individuals being the dominant proportion in the earlier, middle, and later latent classes. However, the results from clustering the entire data set are more stark. In contrast with the complete clustering, Groups 4 and 5 contain a significantly higher number of individuals. Moreover, our model is able to detect a group of individuals with high dementia, as evidenced by Group 5's mean CDR score of 1.55. Unlike the complete case clustering, Group 5 has almost 75% of individuals with some form of dementia. Many of these individuals would be omitted from the analysis if the missing data was not accounted for. Thus, accounting for the missingness leads to a clustering result with a stronger correlation with clinical assessments (CDR score).

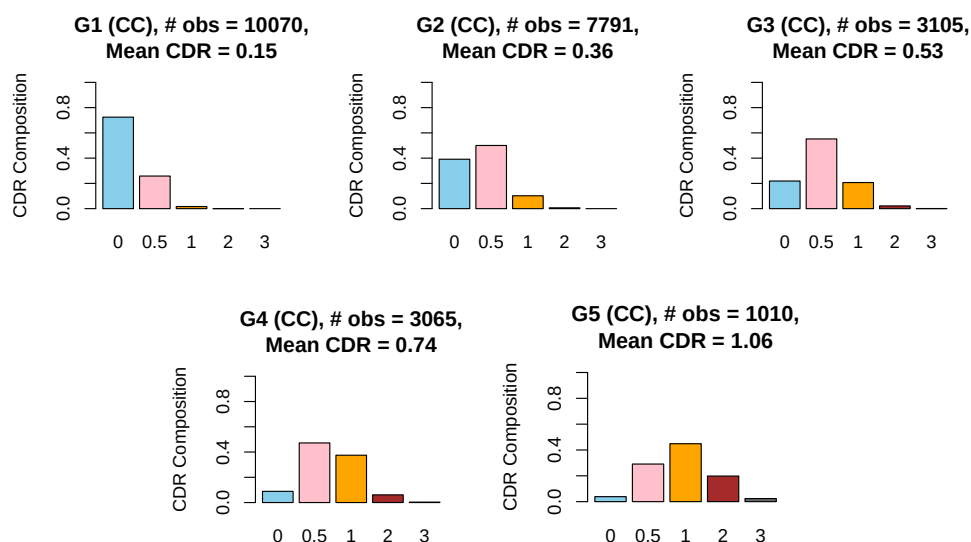


Figure 3.3: Clustering on complete data only. These barplots summarize the composition of each of the five latent groups. We order the groups from most healthy to least healthy. This is reflected in the mean CDR score of each group.

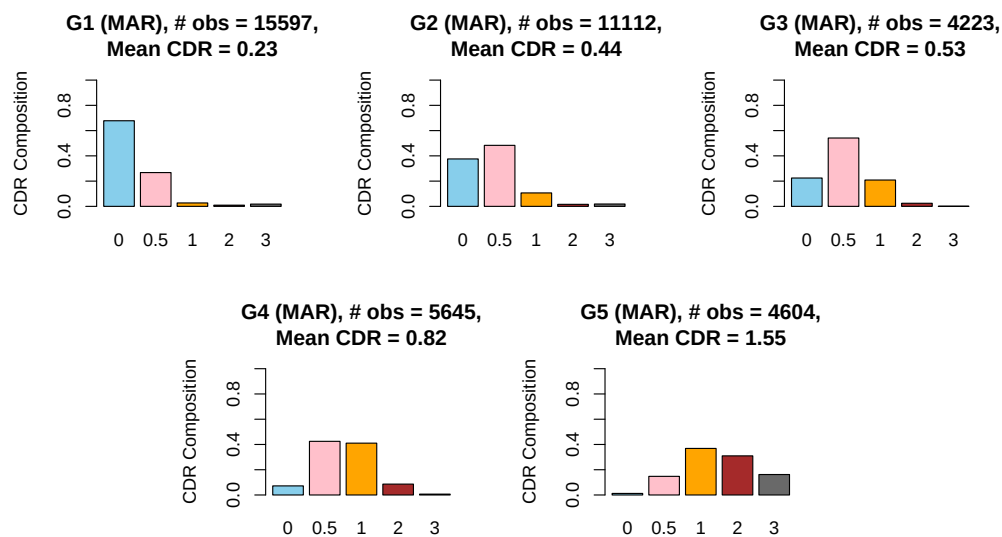


Figure 3.4: Clustering on the entire data with MAR assumption. These barplots summarize the composition of each of the five latent groups. We order the groups from most healthy to least healthy. This is reflected in the mean CDR score of each group.

8 Discussion

In this chapter, we have proposed a mixture of binomial experts model for modeling neuropsychological test score data. This model builds on classical ideas from the latent variable and mixture of experts literature. Through this model, we are able to construct a latent representation of an individual’s cognitive ability using their test scores and relate it to baseline covariates. Because the Alzheimer’s disease data set is both enriched with multivariate information and plagued with missing data, we address both of these issues. We outline how to perform estimation and inference under the MAR assumption. Further comments on simulations are included in Appendix 8.

There are several avenues for extending this chapter. Our presented model has two key components: the weights $w_k(x; \beta_k)$ and the component probabilities $p_k(y; \theta_k)$. Due to the local independence assumption, the component probabilities $p_k(y; \theta_k)$ reduce to a product of marginal distributions, which is a binomial product in this chapter. One straightforward way to generalize this work is modifying these two working models to other parametric families. In our setting, the outcome data is discrete and bounded because we have neuropsychological test scores. However, often, outcome data can come in a variety of different forms such continuous, discrete, and mixed, as well as bounded and unbounded. For instance, in continuous and count data, respectively, the Gaussian and Poisson distributions may be applicable. Access to individual question level data may allow us to leverage ideas from item response theory and the Rasch model (Rasch, 1960). Additionally, since there is longitudinal information in the data, there may be a way to incorporate time in the model, so that model-based clustering can be performed on entire trajectories rather than individuals at given a time point. There has been previous work on mixture of experts models applied to time series and longitudinal data (Waterhouse et al., 1995; Huerta et al., 2003; Tang and Qu, 2016), and it would be interesting to extend these frameworks to a model for multivariate

longitudinal discrete data. Model selection is also another avenue for future exploration as we selected our model using prior knowledge. We included comments on AIC and BIC in Appendix B.6. Also, since we have 33 ADRC centers collecting data, we may introduce a random effect model in the weight to handle the effect from different centers.

The MAR assumption is fundamental to our analysis of the NACC data as well as the inference procedure (Algorithm 5), but it is not the only option. Missing not at random (MNAR) assumptions encompass a rich class of potentially plausible assumptions. It would be interesting to explore MNAR assumptions in the context of latent variable modeling using methods such as pattern graphs (Chen, 2022a; Cheng et al., 2022; Suen and Chen, 2021). Additionally, perturbing the MAR as a form a sensitivity analysis remains largely an open question. We note that departures from missing at random would likely result in a significant modification of the elegant estimation procedure outlined in this chapter because the ignorability condition would no longer hold. Missing not random assumptions often require stronger modeling assumptions because the relationship between $\mathbf{Y}$ and $\mathbf{R}$ needs to be modeled directly. We leave this for future work.

Appendix B: Chapter 3 Appendices

B.1 Detailed Derivation of the EM Algorithm Update Equations

There are two EM algorithms described in this paper. The inner EM algorithm (Algorithm 3) is used to estimate the mixture of binomial product experts model while the outer EM algorithm (Algorithm 5) is used to fit the aforementioned model under a missing at random assumption. When there is no missingness, Algorithm 5 reduces to Algorithm 3. We discuss the details of both algorithms and the multiple imputation algorithm in the following three subsections.

B.1.1 Latent Variable EM

The E-step is easily obtained by calculating the $Q_{\text{LC},n}$ function. Since the only variable that is not conditioned on is Z , this is simply equivalent to changing all of the indicator variables in $\ell_{\text{LC},n}$ to conditional probabilities on $\mathbf{X}$ and $\mathbf{Y}$.

Thus, the $Q_{\text{LC},n}$ function writes as

$$\begin{aligned} & Q_{\text{LC},n}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) \\ &:= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log w_k(\mathbf{X}_i; \beta_k) + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log p_k(\mathbf{Y}_i; \theta_k), \end{aligned} \quad (3.12)$$

where

$$\hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) := P(Z = k | \mathbf{X}_i, \mathbf{Y}_i; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) = \frac{w_k(\mathbf{X}_i; \hat{\beta}_k^{(t)}) p_k(\mathbf{Y}_i; \hat{\theta}_k^{(t)})}{\sum_{k'=1}^K w_{k'}(\mathbf{X}_i; \hat{\beta}_{k'}^{(t)}) p_{k'}(\mathbf{Y}_i; \hat{\theta}_{k'}^{(t)})}.$$

We now derive the update equations in the M-step. Note that we update $(\boldsymbol{\beta}, \boldsymbol{\theta})$ via

$$(\widehat{\boldsymbol{\beta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}) = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}}{\operatorname{argmax}} Q_{\text{LC},n}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}).$$

Taking a partial derivative of (3.12) with respect to $\theta_{k,j}$, we obtain

$$\begin{aligned} & \frac{\partial}{\partial \theta_{k,j}} \frac{1}{n} \sum_{i=1}^n \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log \left(\binom{N_j}{Y_{i,j}} \theta_{k,j}^{Y_{i,j}} (1 - \theta_{k,j})^{N_j - Y_{i,j}} \right) \\ &= \frac{\partial}{\partial \theta_{k,j}} \frac{1}{n} \sum_{i=1}^n \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \cdot (Y_{i,j} \log \theta_{k,j} + (N_j - Y_{i,j}) \log(1 - \theta_{k,j})) \\ &= \frac{1}{n} \sum_{i=1}^n \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \cdot \left(\frac{Y_{i,j}}{\theta_{k,j}} - \frac{N_j - Y_{i,j}}{1 - \theta_{k,j}} \right). \end{aligned}$$

Setting this derivative equal to 0 yields the desired update equation for $\boldsymbol{\theta}$

$$\widehat{\theta}_{k,j}^{(t+1)} = \frac{\sum_{i=1}^n \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \cdot \frac{Y_{i,j}}{N_j}}{\sum_{i=1}^n \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)}.$$

We now discuss the update procedure for $\boldsymbol{\beta}$. We see that

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log p_k(\mathbf{Y}_i; \theta_k)$$

is the log-likelihood function with weights $\widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$ rather than $1(Z_i = k)$, so we can apply a standard package that fits a logistic regression model using gradient ascent. However, instead of treating the dependent variable as $(1(Z_i = 1), 1(Z_i = 2), \dots, 1(Z_i = K))$, we use the dependent variable $\mathbf{W}_i(\mathbf{X}_i, \mathbf{Y}_i) \equiv (\widehat{\pi}_1^{(t)}(\mathbf{X}_i, \mathbf{Y}_i), \widehat{\pi}_2^{(t)}(\mathbf{X}_i, \mathbf{Y}_i), \dots, \widehat{\pi}_K^{(t)}(\mathbf{X}_i, \mathbf{Y}_i))$.

This completes the derivation of Algorithm 3.

B.1.2 Multiple Imputation under Missing at Random

The missing at random property equivalently can be written as

$$\begin{aligned} P(\mathbf{R} = r|x, y) = P(\mathbf{R} = r|x, y_r) &\iff p(x, y, r) = p(x, y_r, r) \cdot p(y_{\bar{r}}|y_r, x) \\ &\iff p(y_{\bar{r}}|y_r, x, r) = p(y_{\bar{r}}|y_r, x) \end{aligned}$$

for all $r \in \mathcal{R}$.

Therefore, the imputation distribution $p(y_{\bar{r}}|y_r, x, r) \equiv p(y_{\bar{r}}|y_r, x)$ is constructed using the global model $p(y|x)$ that is fit on all of the data regardless of the missing data pattern.

Algorithm 4 requires the estimation of the conditional distribution $p(y_{\bar{r}}|y_r, x; \boldsymbol{\beta}, \boldsymbol{\theta})$ for every $r \in \mathcal{R}$. In the following derivation of the imputation distribution, we drop the parameters and reference to the iteration t for readability. For any $r \in \mathcal{R}$, we have

$$\begin{aligned} p(y_{\bar{r}}|y_r, x) &= \frac{p(y|x)}{p(y_r|x)} \\ &= \frac{p(y|x)}{\sum_{y_{\bar{r}}} p(y|x)} \\ &= \frac{\sum_{k=1}^K w_k(x) \prod_{j=1}^d p_{k,j}(y_j)}{\sum_{y_{\bar{r}}} \sum_{k=1}^K w_k(x) \prod_{j=1}^d p_{k,j}(y_j)} \\ &= \frac{\sum_{k=1}^K w_k(x) \prod_{j=1}^d p_{k,j}(y_j)}{\sum_{k=1}^K w_k(x) \sum_{y_{\bar{r}}} \prod_{j=1}^d p_{k,j}(y_j)} \\ &= \frac{\sum_{k=1}^K w_k(x) \prod_{j=1}^d p_{k,j}(y_j)}{\sum_{k=1}^K w_k(x) \prod_r p_{k,j}(y_j)} \\ &= \sum_{k=1}^K \underbrace{\frac{w_k(x) \prod_{j \in r} p_{k,j}(y_j)}{\sum_{k=1}^K w_k(x) \prod_{j \in r} p_{k,j}(y_j)}}_{W_{k,r}(x, y_r)} \prod_{j \in \bar{r}} p_{k,j}(y_j) \\ &= \sum_{k=1}^K W_{k,r}(x, y_r) \prod_{j \in \bar{r}} p_{k,j}(y_j). \end{aligned}$$

Therefore, with a little algebra and the local independence assumption, we can show that the conditional distribution (and imputation distribution) remains a (reweighted) mixture of binomial products distribution, whose computation is fairly straightforward.

B.1.3 Missing at Random EM

Standard maximum likelihood inference can be viewed as constructing an M -estimator that estimates the population MLE

$$(\beta_0, \theta_0) = \underset{\beta, \theta}{\operatorname{argmax}} \mathbb{E}[\log p(\mathbf{Y}|\mathbf{X}; \beta, \theta)].$$

By the law of iterated expectation, observe that

$$\begin{aligned} \mathbb{E}[\log p(\mathbf{Y}|\mathbf{X}; \beta, \theta)] &= \sum_r \mathbb{E}[\mathbb{E} \log p(\mathbf{Y}|\mathbf{X}; \beta, \theta) | \mathbf{X}, \mathbf{Y}_r, \mathbf{R} = r] \cdot 1(\mathbf{R} = r) \\ &\stackrel{\text{MAR}}{=} \sum_r \mathbb{E}[\mathbb{E} \log p(\mathbf{Y}|\mathbf{X}; \beta, \theta) | \mathbf{X}, \mathbf{Y}_r] \cdot 1(\mathbf{R} = r). \end{aligned} \quad (3.13)$$

$$= \mathbb{E} \left[\sum_{r \in \mathcal{R}} \mathbb{E}[\log p(\mathbf{Y}|\mathbf{X}; \beta, \theta) | \mathbf{X}, \mathbf{Y}_r] \cdot 1(\mathbf{R} = r) \right]. \quad (3.14)$$

By converting equation (3.13) to the sample form, we obtain

$$\begin{aligned} Q_{\text{FC},n}(\beta, \theta | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}) &= \frac{1}{n} \sum_{i=1}^n \sum_{r \in \mathcal{R}} Q_{\text{FC},r}(\beta, \theta | \mathbf{X}_i, \mathbf{Y}_{i,r}; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}) \cdot 1(\mathbf{R}_i = r), \\ &= \frac{1}{n} \sum_{i=1}^n Q_{\text{FC},\mathbf{R}_i}(\beta, \theta | \mathbf{X}_i, \mathbf{Y}_{i,\mathbf{R}_i}; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}), \end{aligned} \quad (3.15)$$

where $Q_{\text{FC},r}(\beta, \theta | \mathbf{X}, \mathbf{Y}_r; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}) = \mathbb{E}[\log p(\mathbf{Y}|\mathbf{X}) | \mathbf{X}, \mathbf{Y}_r; \hat{\beta}^{(t)}, \hat{\theta}^{(t)}]$. By construction, we expect that $Q_{\text{FC},n}(\beta, \theta | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}; \hat{\beta}^{(t)}, \hat{\theta}^{(t)})$ has expectation equal to $\mathbb{E}[\log p(\mathbf{Y}|\mathbf{X}; \beta, \theta)]$. Unfortunately, we are not able to compute (3.15) because we do not have an explicit closed form

expression for $Q_{\text{FC},r}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}_r; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)})$. Thankfully, we can approximate it stochastically.

For large enough M , we expect

$$\frac{1}{M} \sum_{m=1}^M \log p(\tilde{\mathbf{Y}}^{(m;t)} | \mathbf{X}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) \approx Q_{\text{FC},r}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}_r; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}),$$

where $\tilde{\mathbf{Y}}^{(m;t)} = (\mathbf{Y}_r, \mathbf{Y}_{\bar{r}}^{(m;t)})$, and for each m , $\mathbf{Y}_{\bar{r}}^{(m;t)} \stackrel{\text{iid}}{\sim} p(y_{\bar{r}} | \mathbf{Y}_r, \mathbf{X}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)})$. Thus, to complete the E-step, we replace the sample version above with a stochastic approximation

$$Q_{\text{FC},n}^{(M)}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) = \frac{1}{M} \sum_{m=1}^M \frac{1}{n} \sum_{i=1}^n \log p(\tilde{\mathbf{Y}}_i^{(m;t)} | \mathbf{X}_i; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}),$$

where $\tilde{\mathbf{Y}}_i^{(m;t)} = (\mathbf{Y}_{i,\mathbf{R}_i}, \tilde{\mathbf{Y}}_{i,\mathbf{R}_i}^{(m;t)})$. To actually perform this stochastic approximation in practice, we simply multiply impute M times and stack the imputed data sets together.

The M-step is equivalent to maximizing the latent incomplete likelihood on the stacked imputed data set $\tilde{D}_t = \{(\mathbf{X}_i, \tilde{\mathbf{Y}}_i^{(m;t)})\}_{i=1,\dots,n; m=1,\dots,M}$. Namely, we have

$$(\hat{\boldsymbol{\beta}}^{(t+1)}, \hat{\boldsymbol{\theta}}^{(t+1)}) = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}}{\operatorname{argmax}} \frac{1}{n} \sum_{(\mathbf{X}_i, \tilde{\mathbf{Y}}_i^{(m;t)}) \in \tilde{D}_t} \ell_{\text{LI}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_i, \tilde{\mathbf{Y}}_i^{(m;t)}).$$

We showed in Section 2 that this can be maximized via an EM algorithm (Algorithm 3). Thus, the M-step of the outer EM algorithm is performed by Algorithm 3. This completes the derivation of Algorithm 5.

B.2 Identification Theory

We provide sufficient conditions for generic identifiability in our mixture of binomial product experts model and restate the proposition here.

[Sufficient conditions for generic identifiability] Suppose the following conditions hold.

A1 Each mixture is distinct such that $\theta_k \neq \theta_{k'}$ when $k \neq k'$.

A2 The number of outcome variables d and the number of mixtures K satisfies the bound

$$d \geq 2\lceil \log_{1+\min_j N_j} K \rceil + 1.$$

A3 The logistic regression model $w_k(x; \beta_k)$ for all $k \in [K]$ is identifiable.

Then, the mixture of binomial product experts is generically identifiable up to permutation of the parameters.

Note that the bound in Assumption A2 is likely to be able to be relaxed even further because it was built off of a general latent class model, where each mixture is assumed to be a product of multinomials rather than binomials. A mixture of binomial products is a submodel within the mixture of multinomial products model. A similar bound is discussed in Allman et al. (2009) when all of the outcome variables have the same dimension.

The sufficient conditions for the identifiability of mixtures of binomials in the one dimensional case has been examined in Teicher (1961). Identification of the logistic regression model can be shown using standard assumptions such as the existence of an invertible design matrix and lack of complete separation. More recently, Ouyang and Xu (2022) have established sufficient and necessary conditions for the identifiability of latent class models with covariates.

B.3 Comments on Code Implementation

B.3.1 Numerical Stability

We use the logsumexp trick, which is commonly used for numerical stability (McElreath, 2020). We are often interested in calculating probabilities of the form

$$(p_1, \dots, p_K),$$

where $\sum_{k=1}^K p_k = 1$. In our work, probabilities are often formed by the product of multiple terms in the range $[0, 1]$. Naive implementations can result in underflow, so we work with log-probabilities. Additionally, consider the scenario where

$$p_k = \frac{\exp(t_k)}{\sum_{k'=1}^K \exp(t_{k'})}.$$

Then, we have

$$p_k = \exp \left(t_k - \left[c + \log \sum_{k'=1}^K \exp(t_{k'} - c) \right] \right)$$

for any constant c . The form inside the brackets is known as the logsumexp trick. If we pick $c = \max_{k'} t_{k'}$, then we can avoid most numerical underflows since each $t_{k'}$ is shifted.

B.3.2 Random Initializations

We randomly initialize β and θ using draws from uniform distributions. The intercept of each β_k is drawn from $U[-1, 1]$ while each slope term is drawn from $U[-0.5, 0.5]$. We draw each $\theta_{k,j}$ from $U[0.1, 0.9]$, so that it is initialized away from the boundary. All draws are performed independently.

B.4 Remarks on Alternative EM Algorithms

Remark 3. For completeness, we provide some remarks on an alternative EM algorithm for fitting the model. Our proposed method comprises an EM algorithm nested within an outer EM algorithm. Alternatively, one can impute the missing outcome and latent class variables simultaneously using the distributions $P(Z = k|x, y_r; \hat{\beta}^{(t)}, \hat{\theta}^{(t)})$ and $p(y_{\bar{r}}|Z = k, y_r, x; \hat{\beta}^{(t)}, \hat{\theta}^{(t)})$ and update $(\hat{\beta}^{(t+1)}, \hat{\theta}^{(t+1)})$ with the explicit MLE estimates. When there is no missingness, this procedure mimics the Stochastic EM procedure described by Celeux and Diebolt (1985) and Celeux and Diebolt (1987). A similar method has been explored in the

missing at random setting under a Gaussian mixture model with no covariates by [Serafini et al. \(2020\)](#). Alternating between these two steps until convergence provides a consistent estimate. However, we do not recommend this method for our model in practice because time to convergence is often longer than Algorithm 5 for the same number of imputations M . We anticipate this is due to the increased stochastic variation that is observed by imputing Z on each iteration.

Remark 4. There has been previous work on model-based clustering with missing data. [Serafini et al. \(2020\)](#) discussed how to estimate Gaussian mixture models in the presence of missing at random data with the EM algorithm and Monte Carlo methods in the E-step. We provide comments on how our approach compares to a similar approach in Remark 3. Our approach differs because we utilize two EM algorithms, one nested within the other. Unlike some of the previous work in the model-based clustering with missing data literature, we take a more broad approach by incorporating covariates in our model as well as describing an inference procedure that quantifies uncertainty. [Sportisse et al. \(2023\)](#) described model-based clustering with missing not at random assumptions using a likelihood-based approach and an EM algorithm. Although there has been rich literature on missing not at random assumptions, there are still open problems with the missing at random assumption (such as performing sensitivity analysis) and sometimes missing not at random assumptions are not easily interpretable.

B.5 Simulations

B.5.1 Simulation Details

We describe the data generating process. We have $p = 2$ covariates and $d = 4$ outcome variables. The model is described as follows:

$$p(x) = N \left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} \right), \quad p(y|x) = \sum_{k=1}^3 w_k(x; \beta_k) \prod_{j=1}^4 \binom{N_j}{y_j} \theta_{k,j}^{y_j} (1 - \theta_{k,j})^{N_j - y_j},$$

where $\beta_1 = (0, 0, 0)$, $\beta_2 = (-1.5, 0.3, 0.4)$, and $\beta_3 = (-2, 0.5, 0.25)$. The parameters for the binomials are as follows

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.8 & 0.8 & 0.8 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}.$$

Thus, for a moderately sized model, we already have 18 parameters. To generate data for the missing data simulations, we use the same generating process as above, but we specify a selection model $P(R = r|x, y)$ to make the missing data. For every missing data pattern $r \neq 1111$, we generate the probabilities $P(R = r|x, y) = P(R = r|x, y_r)$ under the following scheme to ensure MAR holds. We have

$$\begin{aligned} \log \frac{P(R = 0001|x, y)}{P(R \neq 0001|x, y)} &= (-2 - \eta) - 0.25x_1 + 0.3x_2 + 0.15y_4, \\ \log \frac{P(R = 0110|x, y)}{P(R \neq 0110|x, y)} &= (-1 - \eta) + 0.3x_1 - 0.7x_2 - 0.1y_2 + 0.15y_3, \\ \log \frac{P(R = 1010|x, y)}{P(R \neq 1010|x, y)} &= (-2 - \eta) + 0.7x_1 - 0.4x_2 + 0.24y_1 - 0.15y_3, \\ \log \frac{P(R = 1110|x, y)}{P(R \neq 1110|x, y)} &= (-1 - \eta) + 0.2x_1 - 0.15x_2 + 0.15y_1 - 0.14y_2 + 0.05y_3, \\ P(R = 1111) &= 1 - \sum_{r \neq 1111} P(R = r|x, y). \end{aligned}$$

Here we treat $\eta \in [0, \infty)$ as a parameter that controls the amount of missingness. As $\eta \rightarrow \infty$, $P(R = r|x, y) \rightarrow 0$ for every $r \neq 1111$, thereby, leading towards $P(R = 1111|x, y) \rightarrow 1$ and consequently, a data set with no missing outcome variables.

B.5.2 Additional Simulations

We provide additional simulations to examine the effect of the number of imputations on the results of the Monte Carlo EM algorithm. We repeat the experiment for $\eta = 2, 2.5$ with $M = 10$ and $M = 50$ imputations. We do not repeat it for $\eta = \infty$ because the latter has no missing data and thus, requires no imputation. In this particular simulation, we actually observe similar performance as we vary from $M = 10, 20, 50$ imputations, suggesting that already at $M = 10$ imputations, the Monte Carlo error is sufficiently small. Additionally, since we are able to know the true generating mixture component for each data point, we compare the assigned clustering to the true generating mixture component. We report the average adjusted Rand index for $M = 10, 20, 50$ in Tables 3.6, 3.7, and 3.8. We observe fairly high and stable adjusted Rand indices across all sample sizes, amount of missingness η , and number of imputations M . This agrees with our expectations of our estimation procedure since the mixtures are well-separated in the generating process.

Furthermore, we run a set of simulations to determine the performance of a complete case analysis. The results are summarized in Table 3.9. We observe that the MSE does not decrease at a linear rate due to the bias associated with not accounting for the missing data and simply performing a complete case analysis. Thus, the estimates are not consistent. The adjusted Rand index is still mostly high with comparable performance to the estimation procedure that takes into account missingness. This can likely be attributed to the fact that the simulation problem has well-separated clusters. In our work, we are also interested in statistical inference and confidence intervals for estimating parameters. As sample size

Table 3.6: These are the results for $M = 10$ imputations. The first row contains results for the estimation of θ . The second row contains results for the estimation of β . The third row contains the average adjusted Rand index for each combination of η and sample size n .

MSE$_{\theta}(\times 100)$				Estimated Coverage of θs			
	Sample Size				Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$		$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.190	0.094	0.049	$\eta = 2.5$	0.940	0.940	0.937
$\eta = 2$	0.202	0.101	0.054	$\eta = 2$	0.939	0.938	0.931

MSE$_{\beta}(\times 100)$				Estimated Coverage of βs			
	Sample Size				Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$		$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	45.5	21.3	10.3	$\eta = 2.5$	0.953	0.952	0.953
$\eta = 2$	43.8	21.4	10.5	$\eta = 2$	0.956	0.950	0.949

Average Adjusted Rand Index			
	Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.935	0.936	0.937
$\eta = 2$	0.935	0.936	0.936

increases, the Estimated Coverage for the parameters generally decreases, away from the desired 95% coverage, implying that the procedure does not produce asymptotically valid confidence intervals. Thus, this simulation highlights the necessity of taking into account the missingness when constructing an estimation procedure and how complete case analysis can yield statistically invalid results.

Table 3.7: This table reports the average adjusted Rand index for each combination of η and sample size n for $M = 20$ imputations over all 1000 generated data sets.

Average Adjusted Rand Index			
	Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.936	0.936	0.937
$\eta = 2$	0.934	0.936	0.936

Table 3.8: These are the results for $M = 50$ imputations. The first row contains results for the estimation of θ . The second row contains results for the estimation of β . The third row contains the average adjusted Rand index for each combination of η and sample size n .

MSE$_{\theta}(\times 100)$			
	Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.189	0.094	0.049
$\eta = 2$	0.204	0.104	0.054

Estimated Coverage of θs			
	Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.942	0.941	0.940
$\eta = 2$	0.938	0.931	0.932

MSE$_{\beta}(\times 100)$			
	Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	44.6	21.8	10.7
$\eta = 2$	47.1	22.0	10.4

Estimated Coverage of βs			
	Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.956	0.947	0.951
$\eta = 2$	0.947	0.947	0.951

Average Adjusted Rand Index			
	Sample Size		
	$n = 500$	$n = 1000$	$n = 2000$
$\eta = 2.5$	0.935	0.936	0.937
$\eta = 2$	0.936	0.936	0.936

Table 3.9: These are the results for complete case analysis. The first row contains results for the estimation of θ . The second row contains results for the estimation of β .

MSE $_{\theta}(\times 100)$						Estimated Coverage of θ s					
Sample Size						Sample Size					
	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$		$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\eta = 2.5$	0.196	0.101	0.050	0.028	0.015	$\eta = 2.5$	0.943	0.936	0.936	0.914	0.887
$\eta = 2$	0.220	0.115	0.061	0.033	0.022	$\eta = 2$	0.938	0.939	0.913	0.887	0.843

MSE $_{\beta}(\times 100)$						Estimated Coverage of β s					
Sample Size						Sample Size					
	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$		$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\eta = 2.5$	49.6	25.1	13.4	7.66	4.80	$\eta = 2.5$	0.940	0.945	0.935	0.904	0.843
$\eta = 2$	59.0	28.3	17.1	10.7	7.71	$\eta = 2$	0.947	0.939	0.908	0.871	0.789

Average Adjusted Rand Index					
Sample Size					
	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\eta = 2.5$	0.936	0.936	0.937	0.937	0.937
$\eta = 2$	0.936	0.936	0.936	0.937	0.937

B.6 Comments on Model Selection

B.6.1 Model selection

In a standard data analysis, one key question is choosing the number of mixture components. In the case of no prior knowledge, we recommend using information criteria for model selection. The Akaike information criterion (AIC) [Akaike \(1973\)](#) and the Bayesian information criterion (BIC) [Schwarz \(1978\)](#) are popular methods for choosing the number of mixture components. Each of these information criteria works by adding a penalty term to the observed log-likelihood, defined above in [\(3.9\)](#).

Let $\nu(K)$ be the number of free parameters of the mixture of binomial product experts model with K mixtures. We showed earlier that $\nu(K) = K(p + d + 1) - p - 1$. As functions of K , the AIC and BIC write as

$$\text{AIC}(K) := 2\nu(K) - 2\ell_{\text{obs},n}, \quad \text{BIC}(K) := \nu(K) \log n - 2\ell_{\text{obs},n},$$

Table 3.10: These are the results of generating 100 data sets of size $n = 500$ with $\eta = 2, \infty$ and computing the AIC and BIC for each. We record for each of the 100 generated sets, which number of mixture of components the AIC and the BIC chose. The first table is for $\eta = \infty, n = 500$, and the second table is for $\eta = 2, n = 500$.

$\eta = \infty, n = 500$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
AIC	0	22	51	14	13
BIC	0	100	0	0	0

$\eta = 2, n = 500$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
AIC	0	47	31	13	9
BIC	0	100	0	0	0

respectively. One can choose the model that minimizes the AIC or BIC.

In another simulation, we repeatedly generate a data set of size $n = 500$ with $\eta = \infty$ and $\eta = 2$ (details in Appendix 8). Then, we fit models of varying sizes with $K = 2, 3, 4, 5, 6$ and create the AIC and BIC curves. For each data set, we find the numbers of mixtures that corresponds to the minimum AIC and BIC. Figure 3.5 provides an example of AIC and BIC curves for varying values of K for two specific data sets.

We generate 100 random data sets of size $n = 500$ with $\eta = 2, \infty$, and determine the number of mixture components selected by the AIC and BIC for each data set. The results are summarized in Table 3.10. We see that in general, the AIC is not as reliable as the BIC. The BIC has a higher penalty term than the AIC, and it chooses the recommended model for all data sets in the simulations. Therefore, this simulation suggests using BIC over AIC as an information criterion. We also encourage the use of prior scientific knowledge when possible for selecting a model.

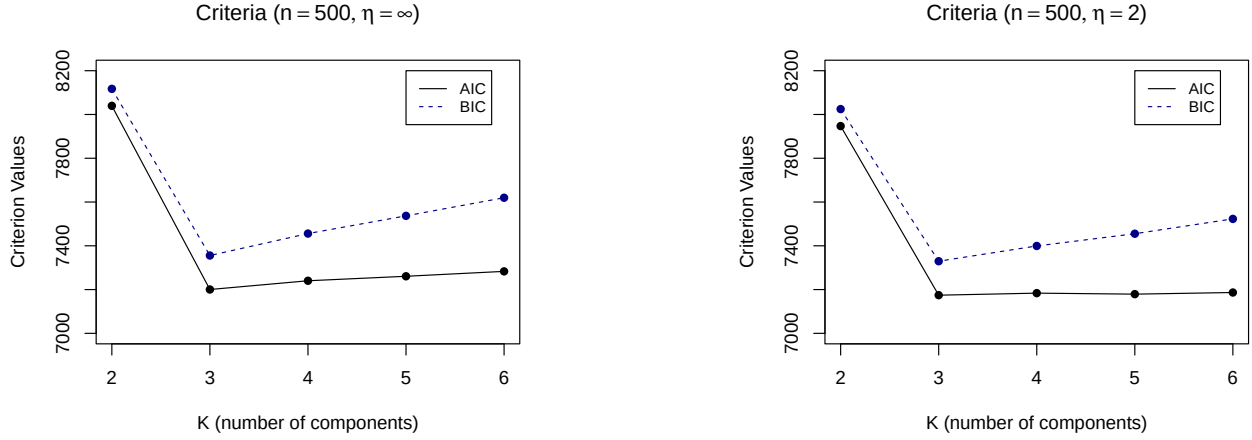


Figure 3.5: This figure depicts the AIC and BIC curves as we vary K for a given simulated data set of size $n = 500$ and $\eta = \infty, 2$.

B.7 Additional Comments on Clustering

Interestingly, the clustering procedure in the presence of missing data that is outlined in Section 5 can be related to a multiple imputation based approach as follows. Consider imputing M data sets and computing the following mean

$$\hat{\pi}_{k, \mathbf{R}_i}^{(M)}(\mathbf{X}_i, \mathbf{Y}_{i, \mathbf{R}_i}) := \frac{1}{M} \sum_{m=1}^M \hat{\pi}_k(\mathbf{X}_i, \tilde{\mathbf{Y}}_i^{(m)})$$

for every observation i . This procedure can be thought of constructing M plausible complete data sets and averaging the latent class probability vectors for each observation i over all of the M imputed data sets. Then, one can perform clustering using the probability $\hat{\pi}_{k, \mathbf{R}_i}^{(M)}(\mathbf{X}_i, \mathbf{Y}_{i, \mathbf{R}_i})$. As $M \rightarrow \infty$, we expect $\hat{\pi}_{k, \mathbf{R}_i}^{(M)}(\mathbf{X}_i, \mathbf{Y}_{i, \mathbf{R}_i}) \xrightarrow{p} \hat{\pi}_{k, \mathbf{R}_i}(\mathbf{X}_i, \mathbf{Y}_{i, \mathbf{R}_i})$ because for any $r \in \mathcal{R}$,

$$\mathbb{E}[P(Z = k | \mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}_r] = P(Z = k | \mathbf{X}, \mathbf{Y}_r).$$

Thus, for sufficiently large M , the two procedures are equivalent because higher M results

in more imputations, which reduces the Monte Carlo error that arises from the variability of the imputation. However, since there is an explicit form for the probability $\hat{\pi}_{k,r}(\mathbf{X}, \mathbf{Y})$, there is no need to do multiple imputation.

Remark 5. We use the notation $\hat{\pi}$ to emphasize that these probabilities are based on an *estimated model from a sample*. We estimate these probabilities by plugging our estimators into our statistical model and solving for the desired probabilities with Bayes' rule. In theory, there exist oracle probabilities $\pi_k(\mathbf{X}, \mathbf{Y}) := P(Z = k|\mathbf{X}, \mathbf{Y})$ and $\pi_{k,r}(\mathbf{X}, \mathbf{Y}_r) := P(Z = k|\mathbf{X}, \mathbf{Y}_r)$ based on the population. One can analyze clustering based on these population quantities, but in practice, we do not have access to these functions because they are parameterized by the true population parameters. Instead, we use $\hat{\pi}_k(\mathbf{X}, \mathbf{Y})$ and $\hat{\pi}_{k,r}(\mathbf{X}, \mathbf{Y}_r)$ as approximations to the population quantities. More theoretical analysis can be performed to compare the clustering of our sample to a population clustering, but this is out of the scope of this paper.

B.8 Proofs

B.8.1 Nonconcavity of the Latent Incomplete Log-Likelihood (Lemma 3.1)

We first demonstrate that the latent incomplete log-likelihood function is nonconcave.

Lemma 3.1 (Nonconcavity of the LI log-likelihood function). *The latent incomplete log-likelihood function*

$$\begin{aligned} \ell_{LI,n}(\beta, \theta; \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) &= \log \left(\prod_{i=1}^n \left(\sum_{k=1}^K w_k(\mathbf{X}_i; \beta_k) p_k(\mathbf{Y}_i; \theta_k) \right) \right) \\ &= \sum_{i=1}^n \log \left(\sum_{k=1}^K w_k(\mathbf{X}_i; \beta_k) p_k(\mathbf{Y}_i; \theta_k) \right). \end{aligned}$$

is not concave.

Proof of Lemma 3.1. We consider a simple counterexample showing nonconcavity. Take $n = 1$, $K = 2$, $p = 1$, and $d = 3$. We further assume that $N_1 = N_2 = N_3 = 5$. This is the log-likelihood where there is exactly one data point with two clusters, a single covariate, and three bounded discrete outcome variables. Suppose that $\mathbf{X} = 1$ and $\mathbf{Y} = (3, 1, 4)$. We have

$$\ell_{\text{LI},n}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}) = \log \left(\sum_{k=1}^2 w_k(X; \beta_k) \prod_{j=1}^3 \binom{N_j}{y_j} \theta_{k,j}^{N_j} (1 - \theta_{k,j})^{N_j - y_j} \right).$$

Consider two points $(\boldsymbol{\beta}^1, \boldsymbol{\theta}^1)$ and $(\boldsymbol{\beta}^2, \boldsymbol{\theta}^2)$, where the superscript indexes the data point. Let $\boldsymbol{\beta}^1 = (-1, -1)$, $\boldsymbol{\theta}^1 = (\theta_1^1, \theta_2^1)$, $\theta_1^1 = (0.7, 0.6, 0.3)$, $\theta_2^1 = (0.3, 0.5, 0.1)$, $\boldsymbol{\beta}^2 = (1, 6)$, $\boldsymbol{\theta}^2 = (\theta_1^2, \theta_2^2)$, $\theta_1^2 = (0.6, 0.4, 0.8)$, and $\theta_2^2 = (0.4, 0.3, 0.3)$. We now consider the line that connects these two points and show that it sits above the function.

Observe that

$$\ell_{\text{LI},n}(t\boldsymbol{\beta}^1 + (1-t)\boldsymbol{\beta}^2, t\boldsymbol{\theta}^1 + (1-t)\boldsymbol{\theta}^2) \approx -6.81 < -6.73 \approx t\ell_{\text{LI},n}(\boldsymbol{\beta}^1, \boldsymbol{\theta}^1) + (1-t)\ell_{\text{LI},n}(\boldsymbol{\beta}^2, \boldsymbol{\theta}^2)$$

for $t = 1/2$. Thus, nonconcavity is achieved. $\square$

B.8.2 Identification Proposition (Proposition 3.1)

Before we present the proof of this proposition, We first introduce a result from Allman et al. (2009) on generic identification. Then, we state two helpful lemmas and providing their proofs.

Theorem 3.1 (Theorem 4 of Allman et al. (2009)). *Let $S_1, \dots, S_d$ be d categorical variables and S_j has B_j categories. Consider a K -mixture model:*

$$P(S = s; \lambda) = \sum_{k=1}^K \omega_k \prod_{j=1}^d q(s_{k,j}; \lambda_{k,j}),$$

where $q(s_{k,j}; \lambda_{k,j})$ is a multinomial distribution on the B_j categories. Suppose there exists a partition C_1, C_2, C_3 of the set $\{1, 2, \dots, d\}$ such that $b_j = \prod_{j \in C_j} B_j$ and

$$\min\{K, b_1\} + \min\{K, b_2\} + \min\{K, b_3\} \geq 2K + 2.$$

Then, the model $P(S = s; \lambda)$ is generically identifiable.

The first lemma describes the generic identifiability of the mixture of binomial products model without covariates.

Lemma 3.2. *Consider the model*

$$\sum_{k=1}^K w_k \left(\prod_{j=1}^d \binom{N_j}{y_j} (\theta_{k,j})^{y_j} (1 - \theta_{k,j})^{N_j - y_j} \right).$$

If the dimension of the outcome variable Y and the number of components satisfy $d \geq 2\lceil \log_{1+\min_j N_j} K \rceil + 1$, then the model is generically identified up to permutation of the parameters.

Proof of Lemma 3.2. This result is a consequence of Theorem 4 from [Allman et al. \(2009\)](#) (which is itself also a result following Kruskal's theorem ([Kruskal, 1976, 1977](#))), and the proof follows a similar construction provided in Corollary 5 from the same paper.

We argue that the latent class model with K components and each Y_j having dimension N_j is generically identifiable under the aforementioned bound.

To simplify notation, let $M_j = 1 + N_j$ and $M_* = \min_j M_j$.

Let K be fixed. First, consider the case that $d = 2\lceil \log_{M_*} K \rceil + 1$. Observe that

$$M_*^{\lceil \log_{M_*} K \rceil - 1} < K \leq M_*^{\lceil \log_{M_*} K \rceil}.$$

Partition the set $\{M_j\}_{j \in [d]}$ into a singleton $P_3 = \{M_*\}$ and two sets P_1 and P_2 of equal size (each with cardinality $\lceil \log_{M_*} K \rceil$). Then, we have

$$\begin{aligned}\kappa_1 &:= \prod_{j \in P_1} M_j \geq M_*^{\lceil \log_{M_*} K \rceil}, \\ \kappa_2 &:= \prod_{j \in P_2} M_j \geq M_*^{\lceil \log_{M_*} K \rceil}, \\ \kappa_3 &:= \prod_{j \in P_3} M_j = M_*.\end{aligned}$$

Then, it follows that $\min(K, \kappa_1) = \min(K, \kappa_2) = K$ and $\min(K, \kappa_3) \geq 2$. Thus, we have

$$\min(K, \kappa_1) + \min(K, \kappa_2) + \min(K, \kappa_3) \geq 2K + 2.$$

When $d > 2\lceil \log_{1+\min_j N_j} K \rceil + 1$, one can partition $\{M_j\}_{j \in [d]}$ in a similar way such that κ_1 and κ_2 strictly increase. Therefore, the previous inequality will still hold.

Now, finally, note that a binomial model is a special case of a multinomial model and each Y_j having dimension N_j . So binomial product models are special cases of the latent class model of multinomial in Theorem 3.1. Generic identifiability of the larger model implies generic identifiability of the submodel. $\square$

The following lemma is used to show how the generic identifiability of the “no covariates” model can imply the generic identifiability of the model with covariates under certain assumptions.

Lemma 3.3. *Suppose the logistic regression model is identifiable, and each mixture is distinct. Then, the generic identifiability of the no covariate binomial product model implies the generic identifiability of the mixture of binomial product experts.*

Proof of Lemma 3.3. As each mixture is distinct, this implies that the model

$$\sum_{k=1}^K w_k(x; \beta_k) p_k(y; \theta_k)$$

is constructed such that $\theta_a \neq \theta_b$ for $a \neq b$. We organize the parameters into one $\alpha = (\beta, \theta) = (\beta_1, \dots, \beta_K, \theta_1, \dots, \theta_K)$ and $\alpha_k = (\beta_k, \theta_k)$. For $\sigma \in S_K$, where S_K is the permutation group comprising all $K!$ mappings from $[K]$ to $[K]$, we use the notation β_σ and θ_σ to denote the reordered tuples of β and θ , respectively, according to the permutation σ .

Now, consider a given covariate $x \in \mathcal{X}$, and suppose that

$$p(y|x; \alpha) = \sum_{k=1}^K w_k(x; \beta_k) p_k(y; \theta_k) = \sum_{k=1}^K w_k(x; \beta'_k) p_k(y; \theta'_k) = p(y|x; \alpha').$$

By generic identification of the no covariate binomial product model up to permutation, we have

$$\theta = \theta'_{\sigma_x}$$

for some permutation $\sigma_x \in S_K$. This permutation is indexed by x because this permutation may depend on the covariate x .

We now argue that this permutation is the same for any x , thereby implying that the entire mixture of binomial products model is unique up to permutation. We proceed using a proof by contradiction. Let x_1 and x_2 be the covariates of two distinct observations with corresponding permutations σ_{x_1} and σ_{x_2} such that $\sigma_{x_1} \neq \sigma_{x_2}$. Without loss of generality, assume that σ_{x_1} is the identity permutation. So, for $x = x_1$, we have

$$\theta_k = \theta'_k \quad \forall k \in [K] \tag{3.16}$$

by the identity permutation. For $x = x_2$, we have

$$\theta_k = \theta'_{\sigma_{x_2}(k)} \quad \forall k \in [K].$$

Since σ_{x_2} is not the identity permutation, there exists $g \in [K]$ such that $\sigma_{x_2}(g) \neq g$. Combining this fact with equation (3.16) via transitivity, we have

$$\theta'_g = \theta_g = \theta'_{\sigma_{x_2}(g)},$$

but this violates the distinct mixture assumption of $\theta_a \neq \theta_b$ for $a \neq b$. Thus, we have a contradiction, and the permutation σ_x must be invariant to the covariate x . Finally, since the logistic regression model is identifiable, the mixture of binomial product experts model is generically identified up to permutation. $\square$

We are now ready to present the proof of Proposition 3.1 (this proves the generic identifiability of our mixture of binomial product experts model), which is a synthesis of results from the previous two lemmas.

Proof of Proposition 3.1. Since the inequality $d \geq 2\lceil \log_{1+\min_j N_j} K \rceil + 1$ is satisfied by Assumption A2, we can invoke Lemma 3.2. This implies that the model without covariates is generically identified.

Then, under Assumptions A1 and A3, we can invoke Lemma 3.3. Thus, the model with covariates is also generically identified, thereby completing the proof. $\square$

B.8.3 Validity of the Bootstrap

In this subsection, we discuss the validity of the bootstrap for this MLE estimator. Before we discuss the theoretical results for the bootstrap, we state the Berry-Esseen bound, which

will be useful in the argument.

Lemma 3.4 (Berry-Esseen bound). *Suppose $Z_1, Z_2, \dots, Z_n$ are i.i.d. random variables such that $\mathbb{E}[Z_i] = \mu$, $\mathbb{E}[Z_i^2] = \sigma^2 > 0$, and $\mathbb{E}[|Z_i|^3] < \infty$. Then, there exists a constant $C > 0$ such that*

$$\sup_z |P(\sqrt{n}(\bar{Z}_n - \mu) \leq z) - \Phi(z)| \leq \frac{C\mathbb{E}[|Z_i|^3]}{\sigma^3\sqrt{n}}.$$

To demonstrate the validity of the bootstrap, we want to argue that the bootstrap distribution is asymptotically equivalent to the sampling distribution of the estimator. The main idea of the proof is to exploit the fact that the estimator $(\hat{\beta}, \hat{\theta})$ has an asymptotically linear form (that is, the MLE behaves much like a sample mean asymptotically) and converges jointly to a multivariate normal centered at (β, θ) at $\sqrt{n}$ -rate.

Define the statistic $\hat{T}_n := T_n(\mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) = \sqrt{n} \cdot (\frac{1}{n} \sum_{i=1}^n \ell(\beta, \theta; \mathbf{X}_i, \mathbf{Y}_i) - \mathbb{E}[\ell(\beta, \theta; \mathbf{X}, \mathbf{Y})])$ as the centered and rescaled log-likelihood function. Let P_* denote the conditional law taking the data $\mathbf{X}_{1:n}, \mathbf{Y}_{1:n}$ fixed and P be the law under the true generating distribution.

More precisely, we wish to show the following uniform bound

$$\sup_t |P_*(T_n^* \leq t | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) - P(\hat{T}_n \leq t)| \xrightarrow{P} 0.$$

We perform the following decomposition via the triangle inequality

$$\begin{aligned} \sup_t |P_*(T_n^* \leq t | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) - P(\hat{T}_n \leq t)| \leq \\ \underbrace{\sup_t |P_*(T_n^* \leq t | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) - \Phi(t; \hat{\sigma})|}_{(A)} + \underbrace{\sup_t |\Phi(t; \hat{\sigma}) - \Phi(t; \sigma)|}_{(B)} + \underbrace{\sup_t |P(\hat{T}_n \leq t) - \Phi(t; \sigma)|}_{(C)}. \end{aligned}$$

Observe by smoothness of the normal CDF and the fact that $\hat{\sigma} \xrightarrow{P} \sigma$, Term (B) is $O_p(n^{-1/2})$. Next, Term (C) is constructed as the difference between the sampling distribution

of $\hat{T}_n$ and the normal CDF. The sampling distribution is simply constructed via the MLE, so there is no asymptotic bias. Therefore, by the Berry-Esseen bound, Term (C) is $O_p(n^{-1/2})$.

The more challenging term to control is Term (A), which we will accomplish via another application of the Berry-Esseen bound. Under the law P_* , the data $\mathbf{X}_{1:n}, \mathbf{Y}_{1:n}$ are fixed, so for some $K \in \mathbb{R}$, we have

$$\begin{aligned} \sup_t |P_*(T_n^* \leq t | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) - \Phi(t; \hat{\sigma})| &\leq \frac{K \frac{1}{n} \sum_{i=1}^n |\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_i, \mathbf{Y}_i)|^3}{\hat{\sigma}^3 \sqrt{n}} \\ &\xrightarrow{p} \frac{K \mathbb{E}[|\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_i, \mathbf{Y}_i)|^3]}{\sigma^3 \sqrt{n}} \\ &= O_p(n^{-1/2}). \end{aligned}$$

where the last convergence follows from the convergence of $\frac{1}{n} \sum_{i=1}^n |\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_i, \mathbf{Y}_i)|^3 \xrightarrow{\text{a.s.}} \mathbb{E}[|\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_i, \mathbf{Y}_i)|^3] < \infty$ and $\hat{\sigma} \xrightarrow{p} \sigma$.

Thus, we have

$$\sup_t |P_*(T_n^* \leq t | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) - P(\hat{T}_n \leq t)| = O_p(n^{-1/2}).$$

Since we have shown the result for the centered and rescaled log-likelihood function, the consistency of the bootstrap follows from standard M-estimation theory (as the MLE is an M-estimator), as desired.

B.9 Further Comments on the Real Data Analysis

B.9.1 Reproducibility of the Real Data Analysis

We report the distribution of the maximum log-likelihood estimates in Figure 3.6. We observe that most of the random initializations converge to a local maximum of the log-likelihood function, as evidenced by the mode of the distribution. This distribution indicates

how important it is to have a sufficiently large number of random initializations in order to properly explore the parameter space. If the real data analysis was to be repeated, we expect to obtain a similar histogram of log-likelihood values.

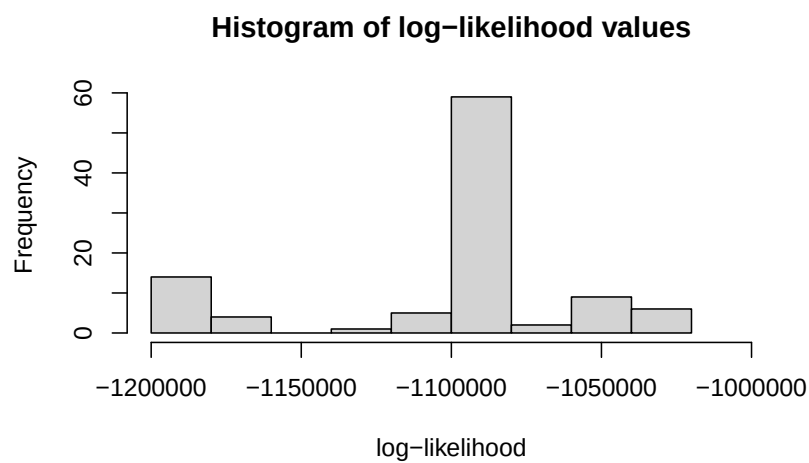


Figure 3.6: This figure is a histogram of the final log-likelihood values, reported over 100 random initializations, obtained using Algorithm 5 on the real data.

Chapter 4

MODELING MULTIVARIATE MISSINGNESS WITH TREE GRAPHS AND CONJUGATE ODDS

1 *Introduction*

Missing data is a pervasive issue across a wide range of disciplines, including healthcare, social sciences, economics, and machine learning. The absence of data points can arise due to various reasons, such as survey nonresponse, equipment failure, or privacy concerns, and the manner in which data is missing greatly influences the validity of statistical analyses. When data sets are large, there is a higher chance for missingness.

Rubin's seminal framework classifies missing data mechanisms into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Standard statistical techniques such as likelihood-based inference are often effective when the data are MCAR or MAR. However, MNAR data present a fundamentally harder problem: the missingness mechanism is nonignorable, meaning that the probability of missingness depends on the unobserved values themselves. This coupling between the missingness process and the data-generating process renders conventional techniques inadequate or biased without strong, and often untestable, assumptions. Ignoring the missing data can be problematic because it can bias the results if the missingness is informative, and it leads to higher uncertainty due to the reduction in sample size. In this chapter, we focus on a general class of missing not at random assumptions.

MNAR is notoriously difficult to model, as it requires explicit assumptions about the underlying missingness process. Unlike MCAR and MAR, which allow for consistent param-

eter estimation under appropriate modeling strategies, MNAR requires the incorporation of an additional model to describe the missing data mechanism. However, this mechanism is often unknown or difficult to justify based on observed data alone, making model specification highly sensitive to assumptions. The problem is further exacerbated in the case of nonmonotone missingness and multivariate data, where missing values do not follow a simple cascading pattern but instead occur sporadically across multiple variables.

Common data analysis methods utilize complete data sets, so imputation can be vital for downstream analysis. However, direct modeling of imputation distributions can be challenging due to the high dimensionality and interdependence among variables. Our objective is to develop models that are not only statistically sound but also interpretable, computationally efficient, and conducive to scalable imputation strategies.

Popular iterative imputation techniques such as multiple imputation by chained equations (*mice*) and MissForest have been widely adopted due to their flexibility and ease of use (van Buuren and Groothuis-Oudshoorn, 2011b; Stekhoven and Bühlmann, 2011). Nevertheless, these methods often implicitly assume MAR or rely on potentially incompatible conditional distributions. For instance, *mice* is constructed by specifying many conditional distributions, but these distributions may conflict with one another, potentially resulting in an incoherent or non-existent joint model. Additionally, MissForest is a deterministic single imputation based on random forests, which can lead to difficulty in uncertainty quantification and consistent estimates. Moreover, when the underlying missingness is MNAR or otherwise misspecified, these techniques can produce biased or incoherent results.

A recently proposed class of nonparametric identifying restrictions is that of pattern graphs (Chen, 2022a). While these graphs form a broad and flexible class, selecting an appropriate structure for data analysis remains a challenge. In this chapter, we introduce a smaller subclass, termed *tree graphs*, which offer reduced modeling complexity, connections

to existing assumptions, and a bridge towards a closed-form modeling strategy. We discuss related algebraic, modeling, and asymptotic theories.

This chapter centers around two key components: the tree graph and conjugate odds property. As outlined in Section 2, the tree graph provides a structured approach to constructing interpretable missing data assumptions that ensure nonparametric identification. In Section 3, the conjugate odds approach enables flexible and intuitive modeling of conditional probability distributions. Together, these elements form the foundation of our framework, which is designed to model missing data distributions.

More broadly, our main contributions are as follows:

1. **Properties, enumeration, and construction of tree graphs.** We introduce the concept of the tree graph (Definition 4.2) and its algebraic properties (Proposition 4.1, Proposition 4.2). Tree graphs are a directed graph between missing data patterns, and each tree graph corresponds to a missing data assumption that can nonparametrically identify the full data distribution. Additionally, we cast existing missing not at random (MNAR) assumptions into the tree graph framework.
2. **Modeling via conjugate odds.** We describe a property of parametric odds models to preserve the conditional data distribution within a given parametric family, thereby introducing the concept of conjugacy (Definition 4.3). One critical example of conjugacy is through the logistic odds model and the exponential family (Proposition 4.5) and mixture of exponential families (Corollary 4.2). We include further examples in this chapter and Appendix C.2.
3. **Modeling missing data distributions through conjugate odds and tree graphs.** We unify the ideas presented in Sections 2 and 3. We first demonstrate that by combining the idea of conjugate odds and the tree graph, one can construct closed-form solutions of the missing data distributions (Theorem 4.2).

4. **Strategies for tree graph selection.** We provide practical guidance on tree graph selection (Section 5). We also discuss generalizations of existing missing data assumptions that fall within the tree graph framework (Proposition 4.7).
5. **Inference.** We outline multiple inference procedures using explicit asymptotic statistics theory and various bootstrap methods (Algorithm 11).
6. **Sensitivity analysis.** We describe how to perform sensitivity analysis to a tree graph missing data assumption. We describe approaches that perturb a given tree graph assumption within and outside the set of tree graphs (Section 7).
7. **Simulations.** In Section 8, we provide some simulations to show finite sample performance of our methods. In our simulations, under correct model specification and the correct tree graph assumption, we observe consistency and nominal coverage.
8. **Real data.** In Section 9, we analyze two types of data: multivariate bounded discrete neuropsychological test data from the National Alzheimer’s Coordinating Center and multivariate continuous data in which we are able to highlight the flexibility of conjugate odds and the Gaussian kernel density estimator.

1.1 Related work

Missing data mechanisms that are missing not at random (MNAR) pose fundamental challenges because the distribution of interest is not identifiable without additional assumptions, and this is an active area of research. Early frameworks introduced two broad modeling strategies for MNAR data: selection models and pattern-mixture models. Selection models specify a joint model for the data and missingness indicators (often via a model for the missingness probability conditional on the unseen data), as exemplified by the influential dropout model from Diggle and Kenward (1994). These models directly incorporate the dependence

of missingness on underlying values, but require untestable parametric assumptions for identifiability. Little’s seminal 1993 work introduced pattern-mixture models (Little, 1993); these stratify the population by missing data patterns and model the distribution within each pattern. This approach offers transparency in how missingness affects the distribution, but it still needs identifying restrictions to be estimable. Subsequent studies such as those from Ibrahim et al. (2005) compared and refined these strategies for generalized linear models.

A variety of MNAR assumptions have been proposed to ensure identifiability of the full-data distribution. One line of work assumes no self-censoring, meaning each variable’s missingness does not depend on its own unobserved value (Shpitser, 2016; Sadinle and Reiter, 2017; Malinsky et al., 2020). Another class of identifying assumptions leverages auxiliary information. For example, Miao and Tchetgen Tchetgen (2016) demonstrated that an observed “shadow” variable (instrumental variable for missingness) can identify MNAR models without specifying the full missingness mechanism. Others have imposed constraints like complete-case missing value (CCMV) restrictions (Little, 1993; Tchetgen Tchetgen et al., 2018) or other structured assumptions on monotone missingness (Molenberghs et al., 1998).

Graphical models have emerged as a powerful way to represent and reason about missing data mechanisms. In missing data DAGs or m-DAGs, one augments the usual directed acyclic graph of variables with nodes for missingness indicators and edges capturing dependencies (Mohan et al., 2013; Tian, 2015; Mohan and Pearl, 2021). Recent work from Phung et al. (2025) uses a pattern directed acyclic graph along with an m-DAG to help with identification of the full-data distribution. These representations make explicit the conditional independence assumptions underlying missingness.

An alternative graphical perspective is provided by pattern graphs (Chen, 2022a). We emphasize that the *pattern graph* proposed by Chen (2022a) is separate from the *pattern directed acyclic graph* from Phung et al. (2025) because the former object corresponds to a

specific MNAR assumption while the latter object is used in combination with an m-DAG to help with identification of the full-data distribution. Also, a pattern graph differs from the aforementioned graph methods because a pattern graph is not a conventional graphical model (where each node corresponds to a variable). More precisely, each pattern graph represents a MNAR assumption where each node is a missingness pattern, and it yields inverse probability weighted and regression adjustment approaches. Further work by [Zamanian et al. \(2023\)](#) studied the sensitivity analysis within the pattern graph framework. The pattern graph set is very large, and in [Section 2](#), we introduce a subclass called tree graphs to simplify the modeling procedure.

Within the last few years, [Franks et al. \(2020\)](#) discussed Tukey’s factorization as an alternative way to model the full-data distribution. Tukey’s factorization is fairly general and provides another option aside from the selection or pattern-mixture model approaches to model the full-data distribution. Our work utilizes this clever factorization. Originally, [Franks et al. \(2020\)](#) used Tukey’s factorization in the setting with a single variable that could be missing, but we expand greatly on the idea by considering multivariate data with multiple nonmonotone missing patterns that are MNAR. Towards this end, we extend the idea of Tukey’s factorization to the multivariate setting (more details in [Section 4](#)).

1.2 Notation

We use a capital boldface variable to denote a vector-valued random variable. In this chapter, we consider a general problem setup, where $\mathbf{X} = (X_1, X_2, \dots, X_d) \in \mathbb{R}^d$ is a random vector of variables. Each of the d variables can possibly be missing for a total of up to 2^d missing patterns. Let $\mathbf{R} \in \mathcal{R} \subseteq \{0, 1\}^d$ be the random binary vector that describes the missing pattern associated with $\mathbf{X}$.

We write $R_j = 0$ if and only if variable X_j is missing. For a fixed pattern r , let $\mathbf{X}_r =$

$(X_j : r_j = 1)$ denote the observed random variables and $X_{\bar{r}} = (X_j : r_j = 0)$ denote the missing random variables. When we write “For j in r ,” this refers to the indices that contain 1. For example, suppose $\mathbf{X} = (X_1, X_2, X_3, X_4)$ and $r = 1001$. We have $\mathbf{X}_r = (X_1, X_4)$ and $\mathbf{X}_{\bar{r}} = (X_2, X_3)$. Then, the statement “For j in r ,” corresponds to “For j in 1, 4.”

We assume that the complete data is generated by sampling i.i.d. from the joint distribution $p(x, r)$, and the resulting associated pattern r generates the observed data. In this chapter, we will use the terminology *full-data distribution* and *pattern-specific joint distribution* to refer to $p(x, r)$ and $p(x|r)$, respectively.

2 Tree Graphs and Identification Theory

2.1 Pattern graphs

Chen (2022a) originally proposed pattern graphs as a way to model nonmonotone missingness and nonparametrically identify the full-data distribution $p(x, r)$. A pattern graph is a directed graph of missing data patterns that encodes a missing data assumption capable of nonparametrically identifying the full data distribution. In the paper, he proposed selection odds and pattern-mixture model formulations with respect to a given graph and showed that the two are equivalent. Estimation procedures using inverse probability weighting, regression adjustment, and semiparametric efficiency theory were explored. Building on this work, we focus on a strict subset of pattern graphs called *tree graphs*.

In this subsection, we first broadly summarize the previous work by introducing the notion of a pattern graph. We impose a partial order on the patterns in $\mathcal{R}$, where for any distinct $s, r \in \mathcal{R}$, we say $s > r$ if and only if the observed variables in pattern r are also observed in pattern s . From this partial order, we can construct a directed graph of all patterns, which forms the aforementioned *pattern graph*.

Definition 4.1. A regular pattern graph is a directed graph of all patterns in $\mathcal{R}$ such that

1. **Single source node.** Pattern $1_d := (\underbrace{1, 1, 1, 1, \dots, 1}_{d \text{ times}})$ is the only source.
2. **Regularity.** If there is an arrow present in the graph T from pattern s to pattern r , then $s > r$.

The second property refers to the regularity and ensures that the graph is directed in a way that preserves the partial ordering of the patterns. Pattern graphs represent information flow, which translates to a specific missing data assumption. Since we have a partial ordering of the missing patterns, we assume that a pattern borrows information from its parents to model its missing data. Denote $\text{PA}_T(r)$ as the set of parents for pattern r in graph T . Formally, the pattern-mixture model of $(\mathbf{X}, \mathbf{R})$ factorizes with respect to T if, we have

$$p(x_{\bar{r}}|x_r, R = r) \stackrel{T}{=} p(x_{\bar{r}}|x_r, R \in \text{PA}_T(r)) \quad \forall r \in \mathcal{R}. \quad (\text{P1})$$

Equation (P1) represents the pattern mixture model factorization property, which states that the extrapolation distribution under pattern r can be identified using information from its parents. Additionally, the selection odds model of $(\mathbf{X}, \mathbf{R})$ factorizes with respect to the graph if

$$\frac{P(R = r|x)}{P(R \in \text{PA}_T(r)|x)} \stackrel{T}{=} \frac{P(R = r|x_r)}{P(R \in \text{PA}_T(r)|x_r)} \quad \forall r \in \mathcal{R}. \quad (\text{P2})$$

Equation (P2) represents the selection odds factorization property, which states that the conditional odds of a pattern r against its parents only depends on the commonly observed variables. The pattern graph enjoys equivalence between the selection odds and pattern-mixture model factorization, which is formalized in Theorem 4.1. The pattern-mixture model formulation illuminates the kind of assumption imposed by pattern graph T . In particular, T associates each pattern r with a set $\text{PA}_T(r)$ comprising closely related patterns whose observed variables also include those of r .

Theorem 4.1 (Selection odds and pattern-mixture model factorizations for pattern graphs, Theorem 4 from [Chen \(2022a\)](#)). *Consider a pattern graph T , and suppose that $p(x_r, r) > 0$ for all $x_r \in \mathcal{X}_r$ and $r \in \mathcal{R}$. The selection odds and pattern-mixture model factorizations with respect to T are equivalent; that is, (P1) and (P2) are equivalent.*

Additionally, a pattern graph can be viewed as a local missing at random assumption. For a given pattern graph, we have the following assumption

$$1(R = r) \perp X_{\bar{r}} \mid X_r, R \in \{r, \text{PA}(r)\}.$$

This means if we restrict our universe to the set comprising a given pattern and its parents, then we arrive at a missing at random like assumption. The indicator of being in pattern r is independent of the missing variables under pattern r given the observed variables under pattern r and being restricted to the universe of r and $\text{PA}(r)$.

2.2 Definition and algebraic properties

[Chen \(2022a\)](#) demonstrated that pattern graphs encompass an exceptionally broad class of missing data assumptions. But there are some obvious shortcomings due to the flexibility of the graph. We now focus on a particular subclass of pattern graphs known as tree graphs. Tree graphs represent a rich subset of pattern graphs, exhibiting notable algebraic and statistical properties that facilitate simpler estimation procedures and graph construction. The term *tree graph* is used to reflect its graphical structure, which resembles a tree with a single root node. Moreover, several established missing data assumptions from the literature can be incorporated into the pattern graph framework and reformulated as tree graphs.

Definition 4.2 (Tree graph). A tree graph is a regular pattern graph in which every pattern $r \neq 1_d$ has exactly one directed path originating from 1_d .

Proposition 4.1. *Every tree graph corresponds to a unique missing not at random (MNAR) assumption that nonparametrically identifies the full data distribution. Additionally, the selection odds $P(R = r_\ell|X)/P(R = 1_d|X)$ admits the following identification formula*

$$\frac{P(R = r_\ell|X)}{P(R = 1_d|X)} \stackrel{\text{tree graph}}{=} \prod_{i=1}^{\ell} \frac{P(R = r_i|X_{r_i})}{P(R = r_{i-1}|X_{r_i})},$$

where $1_d =: r_0 \rightarrow r_1 \rightarrow r_2 \rightarrow \cdots \rightarrow r_\ell$ is the unique path in the tree graph from the source 1_d to pattern r_ℓ .

In this chapter, we denote the set of tree graphs formed from d variables as $\mathcal{T}_d$. For brevity, we omit the subscript d and simply write $\mathcal{T}$ when the context makes it clear that we are considering d variables. From the definition, we can see that a tree graph is a directed graph of the patterns in which there is a unique single path from the source 1_d to a given pattern r . In classical graph theory, this structure is also known as an *arborescence*. Proposition 4.1 highlights a key identification result for tree graphs. That is, a tree graph is a MNAR assumption that automatically nonparametrically identifies the full data distribution. MNAR assumptions are notably difficult to formulate. To make this graphical formulation more concrete, we include two specific examples of common missing data assumptions that can be reframed as tree graphs. In Example 3, we discuss the complete-case missing value (CCMV) assumption (Little, 1993; Tchetgen Tchetgen et al., 2018).

Example 3 (Complete-case missing value (CCMV)). *Our first example is the complete-case missing value, which is equivalent to*

$$p(x_{\bar{r}}|x_r, R = r) \stackrel{CCMV}{=} p(x_{\bar{r}}|x_r, R = 1_d)$$

for all $r \in \mathcal{R}$ (Little, 1993; Tchetgen Tchetgen et al., 2018). This can be viewed as a

relaxation of a complete case analysis to an assumption that does not place constraints on the observed data. In particular, the complete case distribution is only used to define the extrapolation distributions.

In contrast, a complete case analysis makes the assumption that

$$p(x|R = r) \stackrel{CCA}{=} p(x|R = 1_d)$$

for any $r \in \mathcal{R}$. The right-hand side of the equation is the complete-case distribution while the left-hand side is the distribution of the data under a given pattern $R = r$. Since the LHS decomposes as $p(x|R = r) = p(x_{\bar{r}}|x_r, R = r)p(x_r, R = r)$, a complete-case analysis implicitly places an assumption on the observed data and thus, may not agree with the observed data. The CCMV assumption bypasses this by only placing assumptions on the distribution of the missing variables, conditional on the observed data, and can be viewed as a first step above a naive CCA. For a visual example, we visualize the CCMV assumption in Figure 4.1 for $d = 3$ variables.

From a graphical perspective, the CCMV assumption represents the most natural tree graph, as it forms the shallowest structure. More broadly, tree graphs can be viewed as generalizations of the CCMV assumption, allowing for more complex paths from 1_d to the remaining patterns. In Example 4, we discuss another tree graph assumption, nearest-case missing value (NCMV) in the context of monotone missingness.

Example 4 (Nearest-case missing value under monotone missingness). *In our second example, we consider a setting of monotone missingness in which the missing patterns form an ordered set. For simplicity, we assume that the missingness arises from dropout such that if variable X_j is missing, then variable $X_{j'}$ is also missing for any $j' > j$. For notational convenience, we denote each pattern by a positive integer that denotes the index of the first*

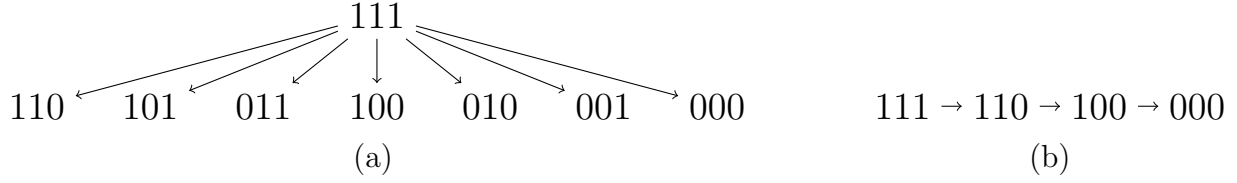


Figure 4.1: (a) The CCMV tree graph for $d = 3$ variables. (b) The NCMV tree graph for $d = 3$ variables.

0 in the missing pattern such that $\mathcal{D} = \{0, 1, 2, 3, 4, 5, 6, \dots, d\}$. Then, the set $\mathcal{D}$ has a one-to-one correspondence with $\mathcal{R} = \{1_d, 1_{d-1}0, 1_{d-2}00, \dots, 0_d\}$, where the subscript denotes the number of 1s and 0s. Letting D denote the random variable associated with $\mathcal{D}$, the NCMV assumption is equivalent to

$$p(x_{>d}|x_{\leq t}, D = d) = p(x_{>d}|x_{\leq d}, D = d + 1),$$

$$\frac{P(D = d|x)}{P(D = d + 1|x)} = \frac{P(D = d|x_t)}{P(D = d + 1|x_t)}$$

for all $d \in \mathcal{D}$. We visualize the NCMV assumption in Figure 4.1 for $d = 3$ variables.

From Examples 3 and 4, we see that previously proposed assumptions from the literature can be cast in the tree graph framework. Through Proposition 4.2, we now introduce equivalent definitions for tree graphs.

Proposition 4.2 (Equivalence definitions for tree graphs). *Let T be a pattern graph. The following statements are equivalent:*

1. **Unique directed path from 1_d .** The pattern graph T is a tree graph with d variables.
2. **Single parent.** Every pattern $r \neq 1_d$ in T has exactly one parent.
3. **Minimal.** There are $2^d - 1$ edges in T . That is, T achieves the lower bound on the number of edges that a pattern graph must have.

Several key practical insights arise from these properties. First, since these formulations are equivalent, the proposition provides multiple possible equivalent definitions of a tree graph. Moreover, as each pattern has exactly one parent, this gives us a straightforward method to both enumerate the class of tree graphs and construct a specific tree graph. The construction process is further discussed in Section 5. Next, minimality is closely linked to model complexity. Missing not at random assumptions can be notably exponentially complex. As the existence of an edge requires fitting an additional selection odds model, minimality ensures that the model complexity for the global model $p(x, r)$ is minimized within the space of pattern graphs and selection models.

2.3 Enumeration

Chen (2022a) demonstrated that the size of the pattern graph set is astronomical as a function of the number of variables d . In contrast, tree graphs form a much smaller subset of pattern graphs as each pattern $r \neq 1_d$ has only one parent. However, despite this reduction, the number of tree graphs still grows significantly with the dimension d . This is formalized with a lower bound, which is presented in the following proposition.

Proposition 4.3 (Enumeration of tree graphs). *The number of tree graphs is super-exponential in the number of variables d . In particular, $\log |\mathcal{T}_d| \gtrsim d \cdot 2^d$.*

The size of the tree graph class grows rapidly. The exact form of our lower bound is provided in the proof, but we note that when $d = 5$, we have $|\mathcal{T}_d| \geq 2^{18}$, and when $d = 6$, we have $|\mathcal{T}_d| \geq 2^{66}$. Observe that the size of the class is largely due to the fact that the number of missing patterns is 2^d , exponential in the number of variables. However, in practice, many of these patterns may not be observed in a given real data set. For example, when the missingness is monotone, the number of missing data patterns is d . Through some careful algebra, one can show that $\log |\mathcal{T}_d|$ under monotone missingness grows at least as fast as d^2

(up to constant factors), which is still substantial. Thus, the tree graph set remains a rich class of missing not at random assumptions while having significant simplifications in the resulting model complexity.

3 Conjugate Odds Families and Domain Adaptation

In the previous section, we established that tree graphs are a graphical representation for a MNAR assumption that can identify the full data distribution. As tree graphs are a subset of pattern graphs, one may apply any of the modeling strategies discussed in [Chen \(2022a\)](#) such as the inverse probability weighting, regression adjustment, or multiply robust approaches. However, we can see that there is another using the idea of Tukey’s factorization ([Tukey, 1986](#)).

Learning another distribution $p(x|A = a')$ given knowledge of the distribution $p(x|A = a)$ is a common problem in statistics and machine learning. This problem is commonly framed within the context of domain adaptation and transfer learning, two highly related fields that focus on transferring knowledge gained from a source domain to improve performance in a target domain with different distributional properties.

In domain adaptation, the goal is to adapt a model that has been trained on data from one domain (the source domain) to perform well on data from a related but different domain (the target domain). This can happen, for example, when a model trained on a certain type of data (such as synthetic data or data collected in one environment) is tested in a different, real-world environment. Domain adaptation methods aim to mitigate the effects of this distribution shift, often by learning shared representations or aligning distributions to reduce discrepancies between source and target data. Transfer learning also involves leveraging knowledge from one task or distribution to improve learning on a related task or distribution, even when the tasks or domains differ. The central idea is that useful patterns

learned from the source domain can be transferred to the target domain, where only limited or no labeled data may be available. In transfer learning, common assumptions include covariate shift, label shift, and concept shift (Sugiyama and Kawanabe, 2012).

We consider a broad view from a generative modeling perspective, in which it can be useful to borrow a preexisting model and shift it to adapt to another domain. Domain adaptation is also related to density ratio estimation, in which the ratio of a target density to a source density is desired (Sugiyama et al., 2012). In the machine learning community, domain adaptation has been shown to be particularly powerful for image and language tasks and can be used to reduce time for model fitting (Kouw and Loog, 2021). In this chapter, we apply the idea of domain adaptation to missing data by considering the complete cases $R = 1_d$ as the source domain and another pattern $R = r$ as the target domain.

Often, the sample size in the target population can be limited, so directly estimating the target distribution may be infeasible or unrealistic. In the missing data context, this is very plausible as it is common for multiple missing patterns to be observed with some being rare and having limited observations.

3.1 Exponential tilting

A natural starting point of this discussion is the idea of exponential tilting (also known as exponential change of measure). Exponential tilting is a fundamental transformation in probability that modifies a base distribution by reweighting it with an exponential factor (Esscher, 1932). We present the idea in the one dimensional case as a motivation. Given a probability density distribution function $p_0(x)$ on $\mathcal{X} \subseteq \mathbb{R}$, the exponentially tilted version of the distribution with respect to a parameter λ takes the form

$$p_1(x) \propto p_0(x)e^{\lambda T(x)} \quad \Longleftrightarrow \quad p_1(x) = \frac{p_0(x)e^{\lambda T(x)}}{\mathbb{E}_{p_0}[e^{\lambda T(X)}]},$$

where $T(x)$ is a chosen function of x , and the denominator $\mathbb{E}_{p_0}[e^{\lambda T(X)}]$ denotes the normalization constant. The interpretation is that parts of the support in which $e^{\lambda T(x)}$ is small will be downweighted, and parts of the support in which $e^{\lambda T(x)}$ is large will be upweighted. Typically, $T(x)$ can be chosen flexibly to arbitrarily reweight the original distribution, but specific choices can lead to simpler parametric models. In Figure 3.1, we observe the effect of tilting a parabolic distribution via an exponential factor; the general unimodal shape is preserved, but the tilting factor introduces a left skewness.

Often, $T(x)$ is chosen to be a sufficient statistic in exponential family models, and the denominator ensures normalization. This transformation effectively shifts the natural parameter of the underlying exponential family distribution, resulting in a new distribution within the same family. Exponential tilting has applications in importance sampling, large deviation theory, and Bayesian inference, where it facilitates likelihood ratio computations and posterior updates.

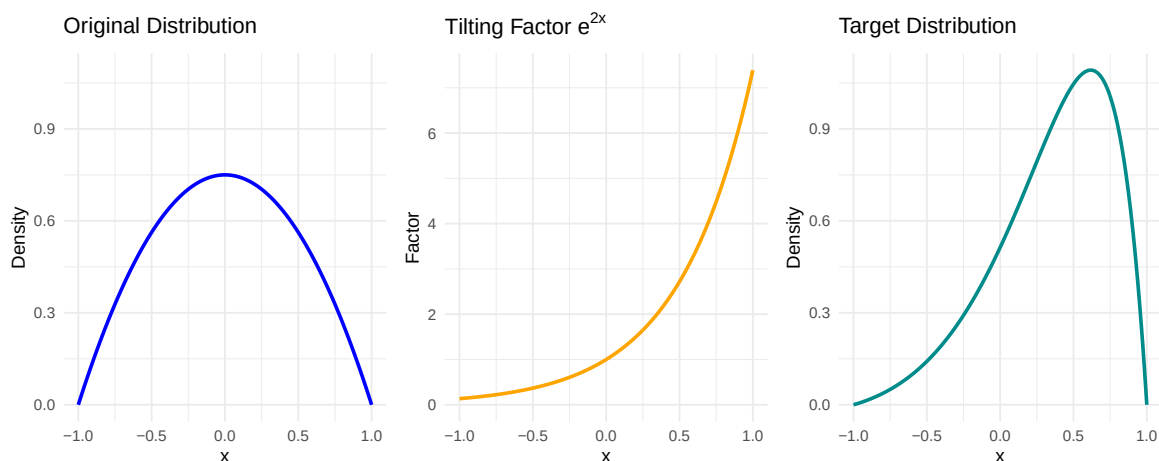


Figure 4.2: The left and right panels depict the original and target distribution (after tilting) on the interval $[0, 1]$. The middle panel is the exponential tilting factor e^{2x} , which corresponds to the weight placed at each part of the original distribution.

A key property of exponential tilting is that it preserves the exponential family structure. If the base distribution belongs to an exponential family with natural parameter η_0 , then the tilted distribution corresponds to a simple shift in the natural parameter, $\eta_\lambda = \eta_0 + \lambda$. This property enables efficient statistical computations, as it allows reweighting while maintaining sufficient statistics and conjugate relationships. In importance sampling, for instance, exponential tilting is used to construct alternative distributions that concentrate probability mass in regions of interest, reducing variance in Monte Carlo estimates. Similarly, in large deviation theory, it provides a means of characterizing rare event probabilities by biasing the original distribution toward extreme outcomes.

Exponential tilting also plays a central role in logistic regression and generalized linear models (GLMs). In binary classification, logistic regression models the log-odds of an outcome as a linear function of input covariates. This formulation implies that the class-conditional densities belong to an exponential family and differ only by an exponential tilt factor. More generally, GLMs express the conditional mean of a response variable in terms of an exponential family distribution, where the tilting parameter is a function of the explanatory variables. This connection underscores how exponential tilting facilitates flexible modeling while maintaining analytical tractability.

3.2 Generalizations to a conjugate odds property

In this section, we discuss a general modeling strategy in which a parametric model is posited for the source domain and under nice conditions, a simple parametric model can also be obtained for the target domain.

A starting point is to first consider the factorization

$$p(x|A = a') \propto p(x|A = a) \cdot \frac{P(A = a'|x)}{P(A = a|x)}. \quad (4.1)$$

From this factorization, we see that the source distribution can be perturbed towards the target distribution by multiplying by an odds factors. In some situations, the odds and the target distributions have nice forms. This leads to the following concept of a conjugate odds.

Definition 4.3 (Conjugate odds). Let $A \in \mathcal{A}$ be a categorical random variable that is auxiliary to the primary data X . Suppose that $p(x|A = a)$ and $p(x|A = a')$ belong to the same probability model $\mathcal{P}$. Then, we say the model formed by

$$\mathcal{O}(\mathcal{P}) := \{O_{a',a}(x) := P(A = a'|x)/P(A = a|x) : \forall a, a' \in \mathcal{A}\}$$

is a *conjugate odds* for $\mathcal{P}$.

In the definition for conjugate odds, we use the term conjugacy to relate it to the Bayesian literature and the idea of conjugate priors. In Bayesian analysis, conjugate priors offer an algebraic convenience in that it provides a closed-form expression for the posterior given a specific likelihood function, thereby bypassing the need for numerical integration or computational methods. In this chapter, we say that a given odds family is *conjugate* to a given family if it satisfies Definition 4.3.

Proposition 4.4 (Conjugate odds holds under mixtures). *Suppose that $\mathcal{O}(\mathcal{P})$ is a conjugate odds for probability model $\mathcal{P}$. Then, $\mathcal{O}(\mathcal{P})$ is a conjugate odds for the probability K -mixture model, where each component is an element of $\mathcal{P}$,*

$$\mathcal{M}_K(\mathcal{P}) := \left\{ p = \sum_{j=1}^K w_j p_j \mid p_j \in \mathcal{P}, \sum_{j=1}^K w_j = 1, w_j > 0 \ \forall j \right\}.$$

We note that in general, a rejection sampling scheme is also possible. One can posit a distribution for $p(x|A = a)$, and fit any odds model for $O_{a',a}(x)$. This can be any binary classifier, which extends this methodology to a suite of machine learning tools. Then, as long

as the odds factor $O_{a',a}(x)$ is bounded, then we can do a rejection sampling scheme by using $p(x|A = a)$ as a proposal distribution to shift towards our desired $p(x|A = a')$. A bounded odds factor is reasonable if the variables $\mathbf{X}$ belong to a bounded set. Although we no longer have a closed-form expression for the target distribution, we are able to perform sampling. This idea is further explored in Section 4.

3.3 Logistic odds

Now, we provide our first example of a conjugate odds family by demonstrating that logistic regression is a conjugate odds for the exponential family.

Proposition 4.5 (Exponential family, vector-valued random variable). *Suppose that $p(x|A = a)$ belongs to the exponential family parameterized by $\eta \in \mathcal{H}$,*

$$p(x|A = a; \eta) = h(x)g(\eta) \exp(\eta^\top T(x))$$

Then, the associated odds model

$$O_{a'}(x; \gamma) := \log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma^\top T(x), \quad \gamma_0 := \log \frac{P(A = a')}{P(A = a)} + \log \frac{g(\eta')}{g(\eta)}$$

holds if and only if

$$p(x|A = a'; \eta') = h(x)g(\eta') \exp((\eta')^\top T(x)),$$

where $\eta' := \eta + \gamma \in \mathcal{H}$.

A natural corollary of Proposition 4.5 is the following result, which establishes a link between a logistic regression model and an exponential tilting factor.

Corollary 4.1 (Exponential tilting and logistic regression). *Imposing a logistic regression model on the odds $P(A = a'|x)/P(A = a|x)$ is equivalent to tilting a distribution $p(x|A = a)$*

by an exponential factor.

Proposition 4.5 has numerous applications, as the exponential family encompasses a broad class of parametric distributions for both discrete and continuous random variables, including the normal, exponential, binomial, Poisson, and negative binomial distributions. Since exponential tilting via logistic regression corresponds to a translation in the natural parameter space, the range of possible values for the natural parameter is of fundamental importance. This proposition further implies that when logistic regression is performed using the sufficient statistics of an exponential family, the fitted coefficients of these statistics directly determine the parameterization of the new distribution $p(x|A = a')$.

One key element of the proposition is the final condition $\eta' := \eta + \gamma \in \mathcal{H}$. Although any pair of identically parameterized exponential family distributions permits a logistic regression representation of the odds, not all logistic regression and exponential family distribution pairs yield an exponential family representation for the target distribution. Similarly, not all exponential tiltings lead to an exponential family and may not even result in a valid distribution. This discrepancy arises when the translation shifts the natural parameter beyond its valid domain. Generally, this issue is mitigated when the natural parameter belongs to an unbounded space. For instance, in the case of the binomial distribution, the natural parameter $\eta := \log p/(1 - p)$ belongs to $\mathbb{R}$, and any translation stays within the set.

Example 5 (One dimensional Gaussian). *Suppose that $X|A = a \sim N(\mu, \sigma^2)$. The sufficient statistic is $T(x) = (x, x^2)$ with natural parameter $\eta = (\mu/\sigma^2, -\frac{1}{2\sigma^2})$ and $g(\eta_1, \eta_2) = \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \cdot \sqrt{-2\eta_2}$. Suppose that*

$$\log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma_1 x + \gamma_2 x^2,$$

where $\gamma = (\gamma_1, \gamma_2)$. Then, $X|A = a' \sim N(\mu', (\sigma^2)')$ such that

$$\mu' := \frac{\mu/\sigma^2 + \gamma_1}{1/\sigma^2 - 2\gamma_2}, \quad (\sigma^2)' := \frac{1}{1/\sigma^2 - 2\gamma_2}.$$

Remark 6. If the $\gamma_2 \geq 1/(2\sigma^2)$, then this suggests that there is model misfit because $(\sigma^2)' \leq 0$, violating the nonnegativity of the variance assumption.

Example 6 (Binomial). Suppose that $X|A = a \sim \text{Binomial}(N, p)$ with N known. The sufficient statistic is $T(x) = x$ with natural parameter $\eta = \log p/(1 - p)$. Suppose that

$$\log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma_1 x$$

where $\gamma = \gamma_1$. Then, $X|A = a' \sim \text{Binomial}(N, p')$ with

$$p' := \frac{\exp(\text{logit}(p) + \gamma_1)}{1 + \exp(\text{logit}(p) + \gamma_1)}.$$

The notion of conjugacy extends to a mixture model, where each component belongs to the same parametric family. This is formalized in Corollary 4.2. There are a few illuminating examples that fall under these specific conditions such as the Gaussian mixture model and binomial product mixture model (Suen and Chen, 2023).

Corollary 4.2 (Mixture of exponential family). Suppose that

$$p(x|A = a) = \sum_{k=1}^K w_k \cdot h(x)g(\eta_k) \exp(\eta_k^\top T(x)), \quad \log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma^\top T(x).$$

Then, we have

$$p(x|A = a') = \sum_{k=1}^K \tilde{w}_k \cdot h(x)g(\eta_k + \gamma) \exp((\eta_k + \gamma)^\top T(x)),$$

where

$$\tilde{w}_k := \frac{w_k \cdot g(\eta_k)}{g(\eta_k + \gamma)} \bigg/ \sum_{k'=1}^K \frac{w_{k'} \cdot g(\eta_{k'})}{g(\eta_{k'} + \gamma)}.$$

Example 7 (Gaussian mixture model with isotropic variance). *Since there are not many convenient options for off-the-shelf modeling of multivariate continuous data, practitioners often use a Gaussian mixture model for its flexibility and relatively easy associated estimation procedure. Suppose that*

$$p(x|A = a) = \sum_{k=1}^K w_k \cdot \prod_{j=1}^d \frac{1}{\sigma_{k,j}^2 \sqrt{2\pi}} \exp \left(-\frac{1}{2\sigma_{k,j}^2} (x_j - \mu_{k,j})^2 \right),$$

$$\log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma_1^\top x + \gamma_2^\top x^2.$$

Then,

$$p(x|A = a') = \sum_{k=1}^K w_k \cdot \prod_{j=1}^d \frac{1}{\sigma'_{k,j}{}^2 \sqrt{2\pi}} \exp \left(-\frac{1}{2\sigma'_{k,j}{}^2} (x_j - \mu'_{k,j})^2 \right),$$

where

$$\mu'_{k,j} := \frac{\mu_{k,j}/\sigma_{k,j}^2 + \gamma_{1,j}}{1/\sigma_{k,j}^2 - 2\gamma_{2,j}}, \quad (\sigma_{k,j}^2)' := \frac{1}{1/\sigma_{k,j}^2 - 2\gamma_{2,j}}, \quad \text{and} \quad w'_k = w_k \cdot \frac{g(\eta_k)}{g(\eta_k + \gamma)} \bigg/ \sum_{k=1}^K \frac{w_k \cdot g(\eta_k)}{g(\eta_k + \gamma)}$$

for $g(\eta_1, \eta_2) = \prod_{j=1}^d g(\eta_{j,1}, \eta_{j,2}) = \prod_{j=1}^d \exp \left(\frac{\eta_{j,1}^2}{4\eta_{j,2}^2} \right) \cdot \sqrt{-2\eta_{j,2}}$ and $\text{logit}(z) := \log(z/(1-z))$.

Example 8 (Binomial product mixture model). *Previously, [Suen and Chen \(2023\)](#) introduced the binomial product mixture model to model multivariate discrete data. Suppose that*

$$p(x|A = a; w, p) = \sum_{k=1}^K w_k \cdot \prod_{j=1}^d \binom{N_j}{x_j} p_{k,j}^{x_j} (1 - p_{k,j})^{N_j - x_j}, \quad \log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma^\top x.$$

Then,

$$p(x|A = a'; w', p') = \sum_{k=1}^K w'_k \cdot \prod_{j=1}^d \binom{N_j}{x_j} p'_{k,j}{}^{x_j} (1 - p'_{k,j})^{N_j - x_j},$$

where

$$p'_{k,j} = \frac{\exp(\text{logit}(p_{k,j}) + \gamma_j)}{1 + \exp(\text{logit}(p_{k,j}) + \gamma_j)} \quad \text{and} \quad w'_k = w_k \cdot \frac{g(\eta_k)}{g(\eta_k + \gamma)} \bigg/ \sum_{k=1}^K \frac{w_k \cdot g(\eta_k)}{g(\eta_k + \gamma)}$$

$$\text{for } g(\zeta) = \prod_{j=1}^d \frac{1}{(1 + \exp(-\zeta_j))^{N_j}}.$$

Example 9 (Gaussian kernel density estimator). Suppose that $p(x|A = a)$ is fit nonparametrically using a kernel density estimator with a product Gaussian kernel as follows

$$\hat{p}(x|A = a) = \frac{1}{nh^d} \sum_{i=1}^n \prod_{j=1}^d K\left(\frac{x_j - X_{i,j}}{h}\right),$$

where $K(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)$.

Thus, the KDE is a Gaussian mixture model with n components (equally weighted), each being a multivariate Gaussian centered at each data point with covariance matrix $\text{diag}(h^2, h^2, \dots, h^2)$. Then, from Example 7, it follows that $\hat{p}(x|A = a')$ is a weighted Gaussian kernel density estimator.

3.4 Power law odds

We provide an additional family of examples through a power law family. When the odds can be modeled using a power law family, we can expand the family of distributions that we have conjugate odds for. Modeling the odds using a power function is a nontraditional method, but it is similar to logistic regression in that it can be interpreted as a linear classifier with a more gradual boundary.

Proposition 4.6 (Power function family, Pareto distribution). *Suppose that $p(x|A = a)$ is a Pareto distribution*

$$p(x|A = a; \alpha, \beta) = \begin{cases} \frac{\alpha\beta^\alpha}{x^{\alpha+1}} & x \geq \beta \\ 0 & o.w. \end{cases}.$$

Then, the associated odds model

$$O_{a'}(x; \gamma) := \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 x^\gamma, \quad \gamma_0 := \frac{P(A = a')}{P(A = a)} \cdot \frac{\alpha'}{\alpha} \cdot \beta^{\alpha' - \alpha}$$

holds if and only if

$$p(x|A = a'; \alpha', \beta) = \begin{cases} \frac{\alpha'\beta^{\alpha'}}{x^{\alpha'+1}} & x \geq \beta \\ 0 & o.w. \end{cases},$$

where $\alpha' = \alpha - \gamma \in \mathbb{R}^+$.

Remark 7. We can also consider an odds model of the form

$$\frac{P(A = 0|x)}{P(A = 1|x)} = \beta_0 + \beta_1 x^{\beta_2}.$$

This shares similar properties with logistic regression in that the odds are always nonnegative for $x > 0$.

If the odds model is generalized to a sum of K terms, then the resulting distribution $p(x|A = a')$ will be a mixture of Pareto distributions with the same shape parameter β . Since the original distribution $p(x|A = a)$ has support in the positive reals, fitting the odds model with a polynomial can be done, provided the polynomial is strictly nonnegative. We provide further examples in [Appendix C.2](#).

4 Tree Graphs and Conjugate Odds

Our goal is now to describe an easy way to construct estimates of the imputation distribution $p(x_r|x_r, R = r)$ and the conditional distribution $p(x|R = r)$. We demonstrate that both of these tasks can be achieved in one shot by unifying the two frameworks (tree graph and conjugate odds) through the idea of Tukey’s factorization.

Definition 4.4 (Tukey’s factorization, (Tukey, 1986)). Consider a univariate $Y \in \mathbb{R}$ that is observed if $R = 1$ and not observed if $R = 0$. We have the following factorization

$$p(y|R = r) = p(y|R = 1) \cdot \frac{P(R = r|y)}{P(R = 1|y)} \cdot \frac{P(R = 1)}{P(R = r)}.$$

Introduced by Tukey in a discussion (Tukey, 1986), the advantage of the above factorization is that identifies the missing data distribution as a product of two terms (one of which is $p(y|R = 1)$ and can be estimated easily) and an odds term, which can be easier to think about and can naturally arise in many applications. The key observation is that the above equation is reminiscent of an aforementioned factorization for tilting a distribution (Equation (4.1)). The term $p(y|R = 1)$ is directly identifiable from the observed data. The odds term $P(R = r|y)/P(R = 1|y)$ depends on unobserved data, but can be identified using the tree graph framework. From here, we can expect to borrow the tools from conjugate odds framework to tilt the complete case distribution $p(x|1_d)$.

Franks et al. (2020) previously built on the idea of Tukey’s factorization as an alternative method from pattern-mixture models and selections models for modeling the full-data distribution. In their work, they discuss this modeling strategy with a a single variable and two possible missing patterns. We extend this work to handle the multivariate case. Tukey’s original factorization can be naturally generalized to a multivariate setting, as seen in the following definition that we propose.

Definition 4.5 (Multivariate Tukey’s factorization). Consider a multivariate $X \in \mathcal{X} \subseteq \mathbb{R}^d$ with an associated missing pattern R . We have the following factorization

$$p(x|R=r) = p(x|R=1_d) \cdot \frac{P(R=r|x)}{P(R=1|x)} \cdot \frac{P(R=1_d)}{P(R=r)}.$$

As in the univariate case, the above factorization demonstrates that $p(x|R=r)$ is proportional to a product of two terms: the complete case distribution $p(x|R=1_d)$ and an odds term $P(R=r|x)/P(R=1_d|x)$. That is, this is another factorization for tilting a distribution as in Equation (4.1). Importantly, this selection odds term is not directly identifiable without further assumptions. However, Proposition 4.1 shows that under a tree graph assumption, these selection odds admits an elegant identification formula and can be estimated using the observed data.

Assumption 1 (Absolute continuity with respect to the complete case distribution). *The distribution $p(x|R=r)$ is absolutely continuous with respect to the complete case distribution $p(x|R=1_d)$ for any $r \neq 1_d$.*

When utilizing Tukey’s factorization, one implicitly is making an assumption. Assumption 1 arises from the nonnegativity of the selection odds nonnegative: if $p(x|R=1_d) = 0$, then $p(x|r) = 0$ must hold. If the complete case distribution satisfies a positivity condition where $p(x|R=1_d) > 0$ for all $x \in \mathcal{X}$, then this assumption will be trivially satisfied. For instance, the mixture models presented in Section 3 satisfy this positivity condition since each mixture component has positive probability on all of $\mathcal{X}$.

We now harmonize the two frameworks with the following theorem.

Theorem 4.2 (Modeling pattern-specific joint distributions using tree graphs and conjugate odds). *Suppose the following conditions hold:*

1. *The missingness mechanism is specified using a tree graph assumption.*

2. The odds model for the selection odds is conjugate to the $p(x|R = 1_d)$ model.

Then, the pattern-specific joint distributions $p(x|R = r)$ for all r belong to the same family as $p(x|R = 1_d)$.

Theorem 4.2 is very powerful because it combines a tree graph and the conjugate odds property in the missing data context and demonstrates how that can lead to elegant modeling of the pattern-specific joint distributions $p(x|R = r)$.

Example 10 (Tree graph with logistic regression and Gaussian model). *Suppose that a tree graph assumption holds, and the selection odds can all be modeled using a logistic regression. Then, if $p(x|R = 1_d)$ belongs to an exponential family, then $p(x|R = r)$ for any $r \in \mathcal{R}$ is also exponential family.*

In particular, suppose that

- $p(x|R = 1_d)$ is a multivariate Gaussian
- all the selection odds can be modeled using logistic regression ($\log[P(R = r|x)/P(R \in PA_T(r)|x)] = \gamma_{0,r} + \gamma_r^\top T(x_r)$ with $T(\cdot)$ being a linear function)

then all the pattern-specific joint distributions $p(x|R = r)$ for all r are multivariate Gaussian. This idea generalizes to exponential family models and mixtures of exponential family models due to Propositions 4.4 and 4.5.

A further consequence of Definition 4.5 is that imposing a tree graph assumption and models for the odds leads to an explicit closed-form expression form $p(x|R = r)$. Notably, this distribution factorizes as $p(x_{\bar{r}}|x_r, r)p(x_r|r)$, so the aforementioned procedure models both the observed data distribution and the missing data distribution in one shot. This is an advantage over other methods such as *mice*, which are able to generate Monte Carlo estimates from the imputation distribution but do not specify a form for the density of the observed or missing data distributions. Because we obtain a specific form for the pattern-

specific joint distribution $p(x|R = r)$ due to conjugacy, it is easier to interpret and also perform imputation without having to refit anything.

4.1 Imputation via a conjugate odds approach

In some settings, estimating a joint model $p(x, r)$ is not the end goal. For example, some might want to complete the data using an imputation. As previously mentioned, we are able to obtain a closed-form expression for the imputation distribution due to conjugacy. We outline this in Algorithm 7.

Algorithm 7 Conjugate odds imputation under logistic regression and exponential family

Require: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n$, a tree graph T

- 1: Fit the complete case model $p(x|1_d; \theta)$ with natural parameter η .
 - 2: **for** $r \neq 1_d$ **do**
 - 3: Fit the logistic regression model $O_r(x_r; \beta_r) := P(R = r|x_r)/P(R \in \text{PA}_T(r)|x_r)$ under the tree graph T
 - 4: **end for**
 - 5: **for** $r \neq 1_d$ **do**
 - 6: Compute the selection odds with respect to the source 1_d via $O_{r:1_d}(x; \lambda_r) := \prod_{r'} O_{r'}(x_{r'}; \beta_{r'})$.
 - 7: Form the parameter of the distribution $p(x|R = r; \theta_r)$ via $\theta_r := \eta^{-1}(\eta + \lambda_r)$.
 - 8: **end for**
 - 9: **for** $m = 1, 2, \dots, M$ **do**
 - 10: **for** $i = 1, 2, \dots, n$ **do**
 - 11: **if** $\mathbf{R}_i \neq 1_d$ **then**
 - 12: Impute $\widetilde{\mathbf{X}}_{i,\mathbf{R}_i}^{(m)}$ using the distribution $p(x_{\bar{r}}|x_r, R = r)$.
 - 13: Set the m -th imputed data to be $\widetilde{\mathbf{X}}_i^{(m)} := (\mathbf{X}_{i,\mathbf{R}_i}, \widetilde{\mathbf{X}}_{i,\mathbf{R}_i}^{(m)})$.
 - 14: **end if**
 - 15: **end for**
 - 16: **end for**
 - 17: **return** $\left\{ \widetilde{\mathbf{X}}_i^{(m)} \right\}_{\substack{i=1,\dots,n \\ m=1,\dots,M}}$
-

4.2 *Rejection sampling for imputation*

When odds are not modeling using a conjugate odds, then there may be challenges in finding a closed form expression for the imputation distribution. However, provided that the odds terms are bounded away from infinity, it is possible to perform rejection sampling. The key requirement is that there exists a constant U_r such that the target density $f(x)$ satisfies $f(x) \leq U_r \cdot g(x)$ for all x , where $g(x)$ is the proposal density. Here we would simply take the proposal distribution to be the complete case distribution $p(x_{\bar{r}}|x_r, R = 1_d)$ and the target distribution to be $p(x_{\bar{r}}|x_r, R = r)$, our true imputation distribution. Since the odds terms are bounded, this ensures that such an U_r exists, making the rejection sampling procedure feasible. We outline this method in Algorithm 8. Although this approach may introduce additional computational overhead, it offers a flexible alternative when traditional sampling methods are not applicable due to the lack of a closed-form expression.

Algorithm 8 Imputation via rejection sampling under logistic regression and exponential family

Require: $\{(\mathbf{X}_i, \mathbf{R}_i)\}_{i=1}^n$, U_r (an upper bound on the odds for pattern r), a tree graph T

- 1: Fit the complete case model $p(x|1_d; \theta)$ with natural parameter η .
- 2: **for** $r \neq 1_d$ **do**
- 3: Fit the logistic regression model $O_r(x_r; \beta_r) := P(R = r|x_r)/P(R \in \text{PA}_T(r)|x_r)$ under the tree graph T
- 4: **end for**
- 5: **for** $r \neq 1_d$ **do**
- 6: Compute the selection odds with respect to the source 1_d via $O_{r:1_d}(x; \lambda_r) := \prod_{r'} O_{r'}(x_{r'}; \beta_{r'})$.
- 7: Form the parameter of the distribution $p(x|R = r; \theta_r)$ via $\theta_r := \eta^{-1}(\eta + \lambda_r)$.
- 8: **end for**
- 9: **for** $m = 1, 2, \dots, M$ **do**
- 10: **for** $i = 1, 2, \dots, n$ **do**
- 11: **if** $\mathbf{R}_i \neq 1_d$ **then**
- 12: **while** $Y^{(m)}$ is not accepted **do**
- 13: Sample a proposal $Y^{(m)} \sim p(x_{\bar{r}}|x_r, R = 1_d)$.
- 14: Accept $Y^{(m)}$ with probability $\frac{p(Y^{(m)}|x_r, R=r)}{U_r \cdot p(Y^{(m)}|x_r, R=1_d)}$.
- 15: **end while**
- 16: Set the m -th imputed data to be $\widetilde{\mathbf{X}}_i^{(m)} := (\mathbf{X}_i, \mathbf{R}_i, Y^{(m)})$.
- 17: **end if**
- 18: **end for**
- 19: **end for**
- 20: **return** $\left\{ \widetilde{\mathbf{X}}_i^{(m)} \right\}_{\substack{i=1, \dots, n \\ m=1, \dots, M}}$

5 *Strategies for Tree Graph Selection*

We established in Section 2.3 that the set of tree graphs grows super-exponentially with the number of variables. Consequently, selecting a plausible tree graph as a missing data assumption for analysis can be a challenging task. We have previously established in Proposition 4.1 that each tree graph is a MNAR assumption that cannot be rejected by the observed data. In practice, it is beneficial to have guiding principles that help narrow down the selection process in a systematic and justifiable manner.

From the single-parent property of Proposition 4.2, we know that selecting a tree graph is equivalent to assigning each pattern to a unique parent. This allows every tree graph G to be represented by a function $\text{PA}_G : \mathcal{R} \mapsto \mathcal{R}$ that satisfies the condition $\text{PA}_G(r) \neq r$ for all $r \in \mathcal{R}$. This functional representation provides a compact and computationally efficient way to encode tree structures, making it particularly useful for computer implementation. To construct a tree graph in a principled manner, we propose the following three guidelines. Each principle serves to restrict the vast space of all possible tree graphs by introducing constraints that are based on domain knowledge, theoretical assumptions, or empirical evidence. We present these principles in the order we recommend they be applied.

Guiding Principles

1. **Prior knowledge.** Select a parent for each pattern that follows prior or scientific knowledge.
2. **Partial ordering.** Select a parent for each pattern based on an existing partial ordering principle (such as CCMV or NCMV). We discuss generalizations of the NCMV assumption to nonmonotone data in a later subsection.
3. **Data-driven method.** Use the data to identify most relevant parents. We provide two methods based on distributional distance and effective sample size.

The reasoning behind this order is to prioritize the most reliable sources of information. We begin with prior knowledge, as it provides a strong, established foundation for selecting patterns. Next, we consider partial orderings, which offer additional structure based on existing theoretical frameworks. Finally, we apply data-driven methods, using empirical evidence to inform our choices. This three-pronged approach ensures that a tree graph is constructed using reasonable approaches.

5.1 *Prior knowledge*

The first and most fundamental principle is to leverage prior knowledge when selecting a parent for each pattern. Scientific insights, domain expertise, or well-established theoretical foundations can provide strong guidance in determining plausible parent-child relationships. For instance, in a biological setting, hierarchical dependencies between genetic markers may be informed by known pathways or functional interactions. Similarly, in causal inference, domain knowledge may suggest causal relationships between variables. By incorporating prior knowledge into the selection process, we ensure that the tree graph aligns with meaningful, interpretable structures that reflect real-world mechanisms.

Example 11 (Longitudinal study with missingness due to dropout). *Consider a longitudinal study where the same test is measured with at regular time intervals. Then, suppose there is monotone missingness due to dropout. We might hypothesize that individuals with missing pattern $R = 1000$ and $R = 0000$ are closely related because we might reason that individuals that never showed up to the study are most similar to individuals who only showed up to the first time point. Then, one can connect the patterns $1000 \rightarrow 0000$. NCMV*

Example 12 (Hierarchical data collection processes). *Suppose we have four collected variables: X_1 , which corresponds to a routine check-up measure such as blood pressure, X_2 , representing a disease state like chronic kidney disease (CKD), X_3 , which measures swelling (a common symptom of CKD), and X_4 , a clinical test result assessing kidney function. In medical settings, it is common for X_3 and X_4 to be recorded only when X_2 exceeds a certain threshold, indicating a more severe condition. Consequently, missing data patterns such as $R = 1000$, 0100 , and 1100 may arise. Since individuals missing X_3 and X_4 are likely healthier, it is plausible to infer hierarchical relationships between these missing patterns, such as $1100 \rightarrow 1000$ and $1100 \rightarrow 0100$, where the presence of both symptom and test data informs cases where only one or neither is recorded.*

Example 13 (Group similarity). *Suppose we have three variables: X_1 , a self-reported stress level, X_2 , alcohol consumption (e.g., self-reported drinks per week), and X_3 exercise habits (e.g., frequency of physical activity per week). There may be a social stigma associated with alcohol, which is related to underreporting and even missingness. We posit that the groups $R = 100$ and 101 are similar in that they are more likely to suffer from such social stigma, so we may suggest a relationship $101 \rightarrow 100$.*

5.2 Partial ordering

Some existing assumptions, such as the nearest case missing value (NCMV) assumption, are based on a partial ordering of missing data patterns but can also carry additional scientific interpretations. [Chen and Sadinle \(2019\)](#) introduced the concept of a donor-based identification strategy, which falls within the pattern graph framework. When the set of parents is restricted to a singleton, the donor-based identification restriction aligns with a tree graph assumption. Additionally, [Little \(1993\)](#) previously introduced the complete case missing value (CCMV) assumption.

As noted earlier, classical missing data assumptions apply in cases of monotone missingness. The NCMV assumption can be interpreted as associating each missing data pattern with a parent pattern that has exactly one additional observed variable. However, a major limitation of the original NCMV assumption is that it is limited to monotone missing data.

Here, we present a generalization of the nearest case missing value (GNCMV) assumption. In the classical NCMV case, we work in the monotone missing data setting, and each pattern has a parent that is the unique pattern that contains exactly one more observed variable. In the nonmonotone missing data setting, the possible parent is no longer unique because there are multiple possible patterns that contain one more observed variable. For example, in the monotone missing data situation 1000 would have parent 1100, but in the nonmonotone missing data setting, it could have parent 1100, 1010, or 1001.

Definition 4.6 (Generalized nearest case missing value (GNCMV)). A tree graph is called a *generalized nearest case missing value* assumption if every pattern in the graph has a parent that contains exactly one more observed variable.

Suppose the data is longitudinal; in this case, missing data patterns can be systematically grouped based on a structured imputation approach. Specifically, each binary missing pattern is associated with another pattern that contains exactly one additional observed

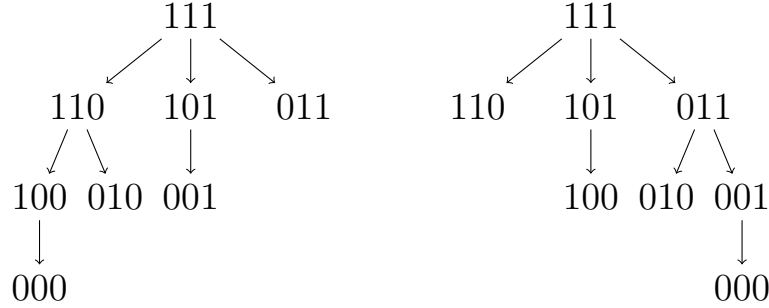


Figure 4.3: The left panel corresponds to the tree graph for $d = 3$ variables and the leftmost-first NCMV (LNCMV) assumption. The right panel corresponds to the tree graph for $d = 3$ variables and the rightmost-first NCMV (RNCMV) assumption.

variable (i.e., one more 1), where the leftmost 0 in the original pattern is replaced by a 1. For example, the pattern 0101 is associated with 1101, and 1010 is associated with 1110. This *leftmost first approach* extends the NCMV assumption to accommodate nonmonotone missingness. For instance, the missing patterns 1000 and 0100 may be closely related to 1100, as they share the same missing structure from Years 3 to 4. By prioritizing the leftmost missing entry when imputing data, this method provides a more structured and flexible framework for handling missing data across a broader range of scenarios. Additionally, this construction can be extended to a *rightmost first approach*. We visualize both of these ideas in Figure 4.3.

Proposition 4.7 (Generalized nearest-case missing value tree graphs). *Denote the subset of tree graphs that exhibit the GNCMV property as $\mathcal{T}_{\text{GNCMV}}$. We have $\log |\mathcal{T}_{\text{GNCMV}}| \gtrsim 2^d$. Moreover, for any $T \in \mathcal{T}_{\text{GNCMV}}$, T exhibits the following properties:*

1. *It achieves the maximum possible depth of d .*
2. *Every pattern r in T is positioned at the maximum possible distance from the source node 1_d , thereby corresponding to the most information flow.*

Observe that by pruning the nonmonotone patterns from each graph, we recover the

tree structure that would exist under the NCMV assumption with monotone missingness. Moreover, the k -th layer contains $\binom{n}{k}$ patterns. This assumption stands in direct contrast to the CCMV assumption, as each pattern is positioned at the maximum possible distance from the source, representing the opposite structural arrangement.

We might expect that assuming CCMV and performing a statistical analysis will yield estimators that remain relatively close to the complete case analysis's point estimates. On the other hand, if each observed distribution $p(x_r|r)$ differs significantly from the complete case distribution, the GNCMV assumptions may lead to point estimates that deviate more from the complete case analysis than those under CCMV.

5.3 Data-driven method

When attempting to infer the structure of a tree graph from data, a natural question is: how should we identify the most plausible parent nodes for a given node? Our data-driven method offers a principled way to rank candidate parents using observed distributions. The data-driven method is an approach one can use to rank potential parents from the data, thereby informing the tree graph structure from the existing data. By assumption, the tree graph asserts the following equality for every pattern r

$$p(x_{\bar{r}}|x_r, R = r) \stackrel{T}{=} p(x_{\bar{r}}|x_r, R = \text{PA}_T(r)).$$

Implicitly, we also assume that there exists a tree graph for which the above assumption holds. Thus, one natural idea is to only match extrapolation distributions if the observed data distributions under both $R = r$ and $R \in \text{PA}_T(r)$ are similar. More precisely, we would desire $d(p(x_r|R = r), p(x_r|R = \text{PA}_T(r)))$ to be small for some probability metric or divergence d . This motivates two possible modeling-based approaches here.

Parent-based modeling. In the first, suppose we obtain data $X_{1,r}, X_{2,r}, \dots, X_{n_r,r} \sim p(x_r|r)$ and attempt to determine which parent distribution $p(x_r|s)$ has the best fit, among all possible parents s . In practice, for each $s \in \text{PPA}(r)$, we fit a parametric model for $p(x_r|s)$ and estimate the expected log-likelihood calculated on the data. This procedure can be expressed in the population version as

$$\operatorname{argmax}_{s \in \text{PPA}(r)} \mathbb{E}_{X_r \sim p(x_r|R=r)} [\log p(X_r|R=s)].$$

The KL divergence provides an alternate perspective. Through a series of equalities, we have

$$\begin{aligned} \operatorname{argmin}_{s \in \text{PPA}(r)} D_{\text{KL}}(p(x_r|R=r) \parallel p(x_r|R=s)) &= \operatorname{argmin}_{s \in \text{PPA}(r)} \int_{-\infty}^{\infty} p(x_r|r) \log \frac{p(x_r|r)}{p(x_r|s)} dx_r \\ &= \operatorname{argmax}_{s \in \text{PPA}(r)} \int_{-\infty}^{\infty} p(x_r|r) \log p(x_r|s) dx_r \\ &= \operatorname{argmax}_{s \in \text{PPA}(r)} \mathbb{E}[\ell(X_r|s)]. \end{aligned}$$

This highlights the fact that the maximization procedure we propose is directly equivalent to picking the pattern that minimizes the sample version of the KL divergence between $p(x_r|R=r)$ and $p(x_r|R=s)$. Implementation of this procedure in practice can be most efficiently done by first estimating each model $p(x_r|r)$ for all $r \in \mathcal{R}$ and then storing each model. We present connections to the KL divergence, but we also note that one could certainly extend this to other distances. More generally, other f -divergences or metrics such as the Wasserstein distance could be explored, particularly when distributional smoothness or support mismatch is a concern. For example, the Hellinger distance can also be utilized and has the nice property that it is a bounded metric.

Child-based modeling. There is an alternative approach through a child-based model-

ing approach. In contrast to the above, suppose we obtain data $X_{1,s}, X_{2,s}, \dots, X_{n_s,s} \sim p(x_s|s)$ and attempt to determine which child distribution $p(x_r|r)$ has the best fit. We outline this in further detail in the Appendix. Provided the fitted models are stored in memory, both the parent-based and child-based approaches have similar computational complexity, but the parent-based method has an illuminating theoretical interpretation. In simulation, we demonstrate that both the parent-based and child-based modeling approaches are able to learn the correct tree graph given enough sample size in some settings. This is discussed in Section 8.

5.4 Random sampling and connections to model averaging

A tree graph can also be generated randomly from the set $\mathcal{T}$. First, we present Algorithm 9, where we show how to sample a tree graph uniformly from $\mathcal{T}$. We can randomly sample from the distribution of parents for each pattern $r \neq 1_d$. By considering the set of possible parents for each pattern r and choosing one uniformly at random, one can form a tree graph. Every tree graph in $\mathcal{T}$ will be equally likely to be selected.

If one performs this sampling and constructs the corresponding point estimator many times, the set of point estimators may be averaged to form a final estimate. We can view this as a form of model averaging.

Algorithm 9 Sampling a tree graph uniformly at random

Require: A set of missing patterns $\mathcal{R}$

- 1: **for** $r \in \mathcal{R}$ **do**
 - 2: **if** $r \neq 1_d$ **then**
 - 3: Define $\text{PPA}_r := \{s : s > r, s \in \mathcal{R}\}$ as the set of potential parents of pattern r .
 - 4: Uniformly sample $s_r \sim \text{PPA}_r$.
 - 5: Form the parent set of pattern r for graph T as follows $\text{PA}_T(r) = \{s_r\}$.
 - 6: **end if**
 - 7: **end for**
 - 8: **return** Tree graph T
-

We can extend this algorithm to randomly sample from an arbitrary distribution by combining Algorithm 9 with a rejection sampling scheme. We present Algorithm 10, which serves as a minor modification to Algorithm 9 by introducing an acceptance criterion but generalizes the sampling to arbitrary distributions over $\mathcal{T}$.

Algorithm 10 Sampling a tree graph from an arbitrary PMF $p(t)$

Require: A set of missing patterns $\mathcal{R}$ and a distribution $p(t)$ over $\mathcal{T}$

```

1: AcceptFlag = 0
2: while AcceptFlag = 0 do
3:   Sample  $T$  uniformly from the space of all tree graphs using Algorithm 9.
4:   Accept  $T$  with probability  $p(T)/\max_t p(t)$ .
5:   if Accepted then
6:     AcceptFlag = 1
7:   end if
8: end while
9: return Tree graph  $T$ 

```

Remark 8 (Bayesian and frequentist perspective). Algorithms 9 and 10 allow a data analyst to place a prior on the set of tree graphs and combine them into a single point estimate. However, we emphasize that while this somewhat mimics a Bayesian approach, this is not a fully Bayesian method because a prior is not place on the parameters but only on the missing data assumptions.

6 Inference

In this section, we describe a procedure for constructing confidence intervals. Recall that in the tree graph and conjugate odds framework, we fit two types of models: a complete case model $p(x|1_d)$ and conjugate odds models $O_r(x_r) := P(R = r|x_r)/P(R \in \text{PA}_T(r)|x_r)$ for every $r \neq 1_d$. This implicitly models the full-data distribution $p(x, r)$, which thereby implies specific forms for the distributions $p(x|r)$. While constructing confidence intervals for the parameters of $p(x|1_d)$ and $O_r(x)$ is fairly straightforward, it is more challenging to construct

confidence intervals for $p(x|R=r)$ for an arbitrary r . This is because such intervals require accounting for the full joint sampling distribution of the parameters, incorporating joint uncertainty across both model components. In the following subsection, we also describe an empirical bootstrap approaches to quantify uncertainty.

Definition 4.7 (Primary model). In the tree graph and conjugate odds setting, we have two types of models: a complete case model $p(x|1_d)$ and an odds model $O_r(x_r)$. We use the term primary model for pattern r to refer to the model that corresponds to the pattern r . If $r = 1_d$, then the primary model is the complete case model $p(x|1_d)$. Otherwise, it is the odds model $O_r(x_r)$.

Each of the primary models described above is fit using observed data from at most two patterns. This suggests that certain MLE parameters for the complete case model $p(x|1_d)$ and the odds models $O_r(x_r)$ may be independent. We formalize that in Proposition 4.8.

Proposition 4.8 (Independence of certain MLE parameters). *Let β_r and β_s be the parameters associated with the primary models for distinct patterns r and s . Suppose neither of these conditions hold:*

1. *One pattern is the parent of the other.*
2. *The two patterns are siblings.*

Then, the MLE estimators $\hat{\beta}_r$ and $\hat{\beta}_s$ are independent.

Based on the results in Proposition 4.8, we can specify exactly the form of the asymptotic covariance matrix through an undirected graph. We describe the idea in Corollary 4.3 and Example 14 describes how this can be applied.

Corollary 4.3 (Block structure of the asymptotic covariance matrix). *In a tree graph T , convert each edge to an undirected edge, and add an undirected edge between every pair of siblings. (This is similar to the idea of moralizing a directed graph except we connect the*

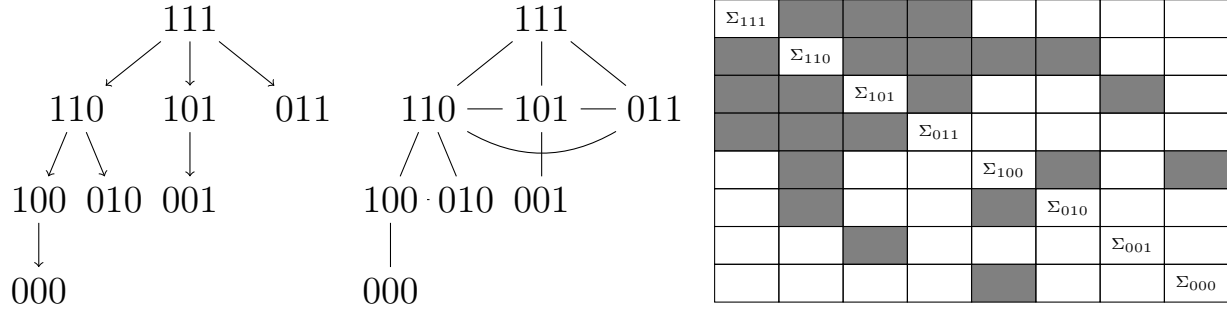


Figure 4.4: The left panel has the original tree graph, and the middle panel shows the resulting undirected graph after connecting the siblings. Let Σ_r denote the asymptotic covariance of the parameters associated with the primary model for pattern r . The block structure of the asymptotic covariance matrix is depicted in the right panel, where the white regions refer to blocks of 0s. All other regions are not guaranteed to be 0.

siblings rather than the parents.) Call the resulting undirected graph U . Then, the maximal cliques of U exactly determine the block structure of the asymptotic covariance matrix.

Example 14 (L-NCMV for 3 variables and its asymptotic covariance structure). *In this example, we consider the L-NCMV tree graph for 3 variables. The results are recorded in Figure 14. We obtain 4 maximal cliques: $\{111, 110, 101, 011\}$, $\{110, 100, 010\}$, $\{101, 001\}$, and $\{100, 000\}$.*

An interesting observation follows from Corollary 4.3. Since under the CCMV assumption, all models are fit using the complete case data, all MLEs will be correlated. In contrast, under a GNCMV assumption, all models are fit with minimal data shared. This leads to the idea of densest and sparsest asymptotic covariance matrices in Proposition 4.9.

Proposition 4.9. *CCMV leads to densest asymptotic covariance matrix. Any tree graph assumption belonging to $\mathcal{T}_{\text{GNCMV}}$ leads to the sparsest asymptotic covariance matrix.*

6.1 Empirical bootstrap

In the bootstrap, we can overcome performing any analytic computation. We have access to the joint bootstrap distribution, which mimics the joint sampling distribution. Our empirical bootstrap approach utilizes resampling from the empirical distribution (Efron, 1979). We describe the process of generating bootstrap samples and refitting the model to obtain bootstrap estimates in Algorithm 11. Because we are operating under a smooth parametric model, the bootstrap is asymptotically valid, and an argument similar to the one provided in Chapter 3 that uses the Berry-Esseen bound can be followed.

Algorithm 11 Empirical bootstrap procedure for obtaining confidence intervals

Require: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n, \hat{\boldsymbol{\theta}}, B$ (a large number, say 1,000)

- 1: **for** $b \in 1, \dots, B$ **do**
 - 2: Sample n draws uniformly with replacement from $\{1, 2, \dots, n\}$. Put these into index set I_b .
 - 3: Set the b th bootstrapped data set $D_b^* = \{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i \in I_b}$.
 - 4: Fit the complete case distribution $p(\mathbf{x}|\mathbf{R} = 1_d; \boldsymbol{\theta}^{*(b)})$ on $D_b^*[\mathbf{R} = 1_d]$.
 - 5: Obtain the selection odds $O_r(x; \hat{\boldsymbol{\beta}}^{*(b)}) := P(R = r|x)/P(R = \text{PA}(r)|x)$ for every pattern $r \in \mathcal{R}$ on the data $D_b^*[\mathbf{R} = r] \cup D_b^*[\mathbf{R} = \text{PA}(r)]$.
 - 6: **end for**
 - 7: **return** $\{\boldsymbol{\beta}^{*(b)}\}_{b=1}^B, \{\boldsymbol{\theta}^{*(b)}\}_{b=1}^B$
-

Since we are under a parametric model, every statistical functional is a function of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. In the situation that the statistical functional does not have a simple analytical form, we recommend computing a multiple imputation estimator, which serves as a Monte Carlo approximation. For every bootstrap estimate $(\boldsymbol{\beta}^{*(b)}, \boldsymbol{\theta}^{*(b)})$, we can construct the imputation distributions $p(x_{\bar{r}}|x_r, R = r)$ for every pattern r and multiply impute. After obtaining a completed data set, then we can compute the statistical functional by computing it on the multiply imputed data set. Then, afterwards, we may pool these estimates together to construct a confidence interval. We summarize this procedure in Algorithm 12. This

approach is generally computationally expensive because within each bootstrap iteration, we have to perform a multiple imputation step, but it overcomes the difficulty of finding a closed-form analytic expression for any general statistical functional we care about. We note that another procedure could take an inverse probability weighting approach. In particular, models for the odds will imply a closed form expression for the inverse propensity score ($1/P(R = 1_d|X)$), and this inverse propensity score can be used to reweight the complete cases. In general, however, we recommend a multiple imputation approach because this will be asymptotically more efficient than an IPW method.

Algorithm 12 Empirical bootstrap procedure with multiple imputation for obtaining confidence intervals

Require: $\{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n$, $\hat{\boldsymbol{\theta}}$, B (a large number, say 1,000), M , $S(\cdot)$ (statistical functional)

- 1: **for** $b = 1, 2, \dots, B$ **do**
- 2: Sample n draws uniformly with replacement from $\{1, 2, \dots, n\}$. Put these into index set I_b .
- 3: Set the b th bootstrapped data set $D_b^* = \{(\mathbf{X}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i \in I_b}$.
- 4: Form the bootstrap empirical distribution $\hat{P}^{*(b)}$.
- 5: Fit the complete case distribution $p(x|R = 1_d; \boldsymbol{\theta}^{*(b)})$ on $D_b^*[\mathbf{R} = 1_d]$.
- 6: Obtain the selection odds $O_r(x; \hat{\boldsymbol{\beta}}^{*(b)}) := P(R = r|x)/P(R = \text{PA}(r)|x)$ for every pattern $r \in \mathcal{R}$ on the data $D_b^*[\mathbf{R} = r] \cup D_b^*[\mathbf{R} = \text{PA}(r)]$.
- 7: **for** $r \neq 1_d$ **do**
- 8: Construct the imputation distribution $p(x_{\bar{r}}|x_r, R = r; \boldsymbol{\beta}^{*(b)}, \boldsymbol{\theta}^{*(b)})$ by renormalizing $O_r(x; \hat{\boldsymbol{\beta}}^{*(b)}) \cdot p(x|R = 1_d; \boldsymbol{\theta}^{*(b)})$.
- 9: **end for**
- 10: **for** $m = 1, 2, \dots, M$ **do**
- 11: **for** $i = 1, 2, \dots, n$ **do**
- 12: **if** $\mathbf{R}_i \neq 1_d$ **then**
- 13: Impute $\widetilde{\mathbf{X}}_{i,\mathbf{R}_i}^{(b,m)}$.
- 14: **end if**
- 15: **end for**
- 16: **end for**
- 17: Form the bootstrap empirical distribution $\hat{P}^{*(b),M}$.
- 18: Compute the statistical functional $S(\hat{P}^{*(b),M})$ on the completed imputed data.
- 19: **end for**
- 20: **return** $\{S(\hat{P}^{*(b),M})\}_{b=1}^B$

7 *Sensitivity Analysis*

In practical data analysis, it is essential to evaluate the influence of missing data assumptions on statistical estimators. Since such assumptions dictate the structure of the missing data mechanism, any misspecification can lead to biased or misleading inferences. In this chapter, we focus on the tree graph as the primary missing data assumption and consider a structured sensitivity analysis framework to assess its impact.

Broadly, sensitivity analysis approaches can be categorized into deviations within the tree graph set and deviations outside the tree graph set. The former considers alternative graph structures that remain within the tree graph set while the latter relaxes the tree structure entirely, allowing for more flexible relationships. Both types of deviations allow one to assess the robustness of the estimator to different levels of structural perturbation. We generally consider the former because that is most within the scope of this chapter. Regarding the latter, there may also be situations in which there exists no specific tree graph that can accurately describe the missingness mechanism. In these settings, incorrectly assuming a tree graph could lead to estimates that deviate far from the truth. Thus, it is critical to ensure that a tree graph assumption is plausible before blindly applying the method described in this chapter.

In all of these settings, the complete case model remains unchanged. However, models involving missing patterns (those dependent on assumptions about the missing data mechanism) are subject to perturbations. By systematically examining these perturbations, we aim to quantify the sensitivity of inference to the assumed missing data structure. This approach provides a principled way to assess the degree to which conclusions depend on a given tree graph assumption.

7.1 Deviation within the tree graph set

A natural approach to evaluating deviations in the tree graph framework is to consider a set of plausible tree graphs, denoted as $\widetilde{\mathcal{T}}$. These alternative graphs can be constructed by incorporating prior knowledge, existing partial orderings (such as the GNCMV framework) and data-driven methods. Exploring multiple tree structures allows us to assess the sensitivity of statistical inferences to different assumptions about the missing data mechanism.

Because the tree graph structure permits the use of conjugate odds imputation, as discussed earlier, we can perform statistical analyses for each tree in $\widetilde{\mathcal{T}}$, obtaining $|\widetilde{\mathcal{T}}|$ point estimates of the target parameter. Comparing these estimates provides insight into the impact of tree specification on inference, helping to determine whether certain structural choices lead to significant variation in results.

7.2 Perturb selection odds models via exponential tilting

Alternatively, one may consider deviating from the tree graph set, and there are multiple approaches one may take. We discuss a straightforward one here in terms of the exponential tilting of the selection odds models.

A given selection odds model $O_r(x_r) = \frac{P(R=r|x_r)}{P(R=s|x_r)}$ can be commonly estimated using logistic regression, especially under our conjugate odds framework. To incorporate sensitivity analysis and assess the robustness of inferences under potential deviations from the assumed selection model, we introduce a perturbation mechanism via exponential tilting (Kim and Yu, 2011; Shao and Wang, 2016; Zhao et al., 2017). With many odds model and many variables, there can be an exponential number of sensitivity parameters one can have. One approach is to consider variable-wise sensitivity parameters, where we have a sensitivity parameter for each of the d variables. The sensitivity parameter vector can be $\rho := (\rho_1, \rho_2, \dots, \rho_d)$.

Specifically, we modify the selection odds model by multiplying it with an exponential

adjustment term, leading to the perturbed selection odds model

$$O'_r(x) := O_r(x_r) \cdot \exp(\rho_{\bar{r}}^\top x_{\bar{r}}),$$

where $\rho_{\bar{r}}$ consists of the sensitivity parameters (one for each missing covariate under pattern r). The exponential tilting formulation allows for a flexible and interpretable perturbation of the selection model. By appropriately choosing $\rho_{\bar{r}}$, one can examine a range of plausible missing data mechanisms, thereby assessing the sensitivity of the resulting inference. Each element of $\rho_{\bar{r}}$ represents a potential deviation from the originally estimated selection model, effectively shifting the selection mechanism in a controlled manner.

Despite the introduction of the perturbation term, the new perturbed selection odds model remains within a parametric logistic regression framework. The exponential tilting approach does not alter the functional form of the selection odds model beyond a simple multiplicative adjustment. As a result, the model retains its parametric interpretability. When a given $\rho_j = 0$, this corresponds to no perturbation. Such a sensitivity parameter can be viewed as a coefficient in a linear model, and one can specify its range based on one's belief of the relative impact of the missing variables to that of the observed variables.

Remark 9. This approach is equivalent to the approach by [Franks et al. \(2020\)](#). In that approach, they consider a single variable Y that is subject to missingness and impose a parametric assumption on selection probability $P(R = 1|y) = \text{logit}(\alpha + \beta y)$, where β is given a prior distribution.

One can view their modeling approach as a special case of our framework with a Bayesian perspective. In our framework, their technique can be viewed as a tree graph approach with an exponential tilting sensitivity analysis. To see this, consider the simple tree $1 \rightarrow 0$ that

provides the identification assumption

$$P(R = 0|y)/P(R = 1|y) \stackrel{1 \rightarrow 0}{=} P(R = 0)/P(R = 1) =: O_0.$$

Then, β can be interpreted as a sensitivity parameter through which we can define a perturbed selection odds model that now depends on the variable y

$$O'_0(y) := O_0 \cdot \exp(\beta y).$$

8 Simulations

In our simulation study, we consider a setting with $d = 3$ bounded discrete variables. The data-generating process assumes that $p(x|R = 1_d)$ follows a mixture of binomial product distributions, while the selection mechanism is modeled such that the selection odds $P(R = r|x)/P(R = 1_d|x)$ follow a logistic regression model for all missing data patterns r . Under correct specification of the true tree graph, we assess consistency and coverage using an empirical bootstrap procedure.

For a given tree graph, the simulation procedure consists of the following steps:

- **Data Generation:** We specify $p(x \mid R = 1_d)$ as a mixture of binomial products, set the probabilities $P(R = r)$ for each missing data pattern, and specify the logistic regression coefficients γ_r and intercepts $\gamma_{0,r}$. This setup ensures that each conditional distribution $p(x \mid R = r)$ remains a mixture of binomial products.

Our data generating process is defined as follows:

R	111	110	101	100	010	001
$P(R)$	0.3	0.2	0.1	0.15	0.15	0.1

$$p(x \mid R = \mathbf{1}_d; \mathbf{w}_{cc}, \boldsymbol{\theta}_{cc}) \quad \text{with} \quad \mathbf{w}_{cc} = \begin{bmatrix} 0.3 & 0.5 & 0.2 \end{bmatrix}, \quad \boldsymbol{\theta}_{cc} = \begin{bmatrix} 0.70 & 0.75 & 0.70 \\ 0.50 & 0.50 & 0.40 \\ 0.20 & 0.30 & 0.10 \end{bmatrix}$$

The missingness mechanism is modeled using logistic regressions:

$$\begin{aligned} \frac{P(R = 110 \mid x)}{P(R = 111 \mid x)} &\propto \exp(0.1x_1 + 0.1x_2), & \frac{P(R = 101 \mid x)}{P(R = 111 \mid x)} &\propto \exp(0.3x_1 + 0.1x_3), \\ \frac{P(R = 100 \mid x)}{P(R = 101 \mid x)} &\propto \exp(-0.1x_1), & \frac{P(R = 010 \mid x)}{P(R = 110 \mid x)} &\propto \exp(0.1x_2), \\ \frac{P(R = 001 \mid x)}{P(R = 101 \mid x)} &\propto \exp(0.1x_3). \end{aligned}$$

- **Consistency Assessment:** For each given sample size, we generate $U = 200$ random data sets based on the data generating process. We estimate model parameters using an expectation-maximization (EM) algorithm for $p(x \mid R = \mathbf{1}_d)$ and standard logistic regression for the selection mechanism. Then, we report the estimated MSE over all $U = 200$ point estimates.
- **Coverage Evaluation:** We report the estimated coverage for our bootstrap approach over all $U = 200$ random data sets using Algorithm 11 and $B = 500$ bootstrap samples. Confidence intervals are constructed by using estimating the standard errors with the bootstrap estimates and then, adding and subtracting them from the point estimate.

In Table 8, we generally see that the estimated MSE decreases at a linear rate, indicating that we have consistent performance. Since our estimator is the MLE, it is also asymptotically efficient under this correct model specification. We also see that the coverage is roughly nominal and that the approach outlined in Algorithm 11 works well. For each data set, we also learn a tree graph using the parent-based and child-based modeling approaches, reporting the learned tree graphs in Figure 8 and their frequencies in Table 8. From here,

we can see that both data-driven methods generally select the correct tree graph with high frequency given a large enough sample size.

Mean Squared Error ($\times 100$)					
	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$	$n = 10000$
θ_{cc}	0.036	0.020	0.0092	0.0036	0.0019
w_{cc}	0.17	0.09	0.041	0.016	0.0073
β	0.29	0.14	0.07	0.026	0.013

Estimated Coverage					
	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$	$n = 10000$
θ_{cc}	0.94	0.94	0.94	0.95	0.94
w_{cc}	0.93	0.94	0.93	0.95	0.95
β	0.98	0.98	0.98	0.97	0.98

Table 4.1: These results are the estimated MSEs and estimated coverage after fitting a mixture model and running Algorithm 11 for $U = 200$ replicates.

Parent-Based Modeling				
	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Tree Graph 1	127	140	160	185
Tree Graph 2	73	60	40	15

Child-Based Modeling				
	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Tree Graph 1	183	194	198	200
Tree Graph 2	15	6	2	0
Tree Graph 3	2	0	0	0

Table 4.2: These results are of the data-driven methods over $U = 200$ random data sets.

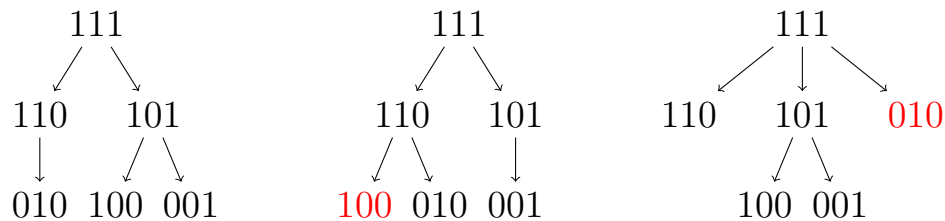


Figure 4.5: The left panel contains Tree Graph 1 (the tree graph used to generate the missing data). The middle and right panels contain Tree Graphs 2 and 3, respectively, and they correspond to graphs incorrectly learned by the parent-based and child-based modeling approaches. The nodes in the middle and right panels that are highlighted red indicate which ones were assigned to an incorrect parent.

9 Real Data Analysis

9.1 Mixture of binomial products

In our first example, we consider the analysis of neuropsychological test scores in the database of the National Alzheimer’s Coordinating Center (NACC)¹. The National Alzheimer’s Coordinating Center (NACC), funded by the NIH and NIA, oversees the largest longitudinal database on Alzheimer’s disease in the United States. It serves as a coordinating hub for 33 Alzheimer’s Disease Research Centers (ADRCs) across the country. This data set that comprises individuals of varying cognitive status: cognitively normal to mild cognitive impairment (MCI) to dementia. Each individual is assigned a CDR (clinical dementia rating) from clinician with 0 corresponding to cognitively normal, 0.5 corresponding to mild cognitive impairment, and 1, 2, and 3 corresponding to mild, moderate, and severe dementia, respectively.

Neuropsychological test scores are typically assessed annually with some baseline covariates collected upon entry to the study. However, there is also some missingness present in

¹<https://nacccdata.org/>

the outcome variables. The missingness can be due to various reasons. For example, some tests may no longer be administered after a given time because they have been replaced with another one. On the other hand, some individuals may have test scores missing because of a data recording error or they are too sick to take more tests.

9.2 *Description of outcome variables and covariates*

Our primary goal is to measure the cognitive ability of the Alzheimer’s disease patients. We focus on the following variable UDSBENTD, which is the total score for ten to fifteen minute delayed drawing of Benson figure. In the Benson figure test, a participant is presented with a diagram of a complex figure and is asked to copy it. After a period of about ten to fifteen minutes, they are asked to recopy it again from memory, and they are assigned a score from 0 to 17 based on how well it resembles the original figure. This test measures visuospatial, visual memory, and executive abilities. We look at individuals who entered the study from the years 2015 to 2019 and follow them for five years total, examining the repeated delayed Benson figure test score each year. We do not use the CDR score in the model, but we use it to help report and interpret the results.

We first plot the missing pattern distribution in Figure 4.6. We can initially observe that the complete cases are very small with $n_{cc} = 271$ individuals out of a possible $n = 13440$. Additionally, every possible pattern of the 16 possible is observed, ranging from $R = 10000$ to $R = 11111$. The distribution is primarily dominated by the monotone missing patterns 10000, 11000, 11100, 11110, and 11111, likely due to dropout. Of primary interest, we will examine the patterns 10000, 11000, 10100, and 10010 because they are some of the larger patterns.

We next plot four different tree graphs of interest: LNCMV, RNCMV, parent-based modeling, and child-based modeling. These four tree graphs are reported in Figure 4.7. In-

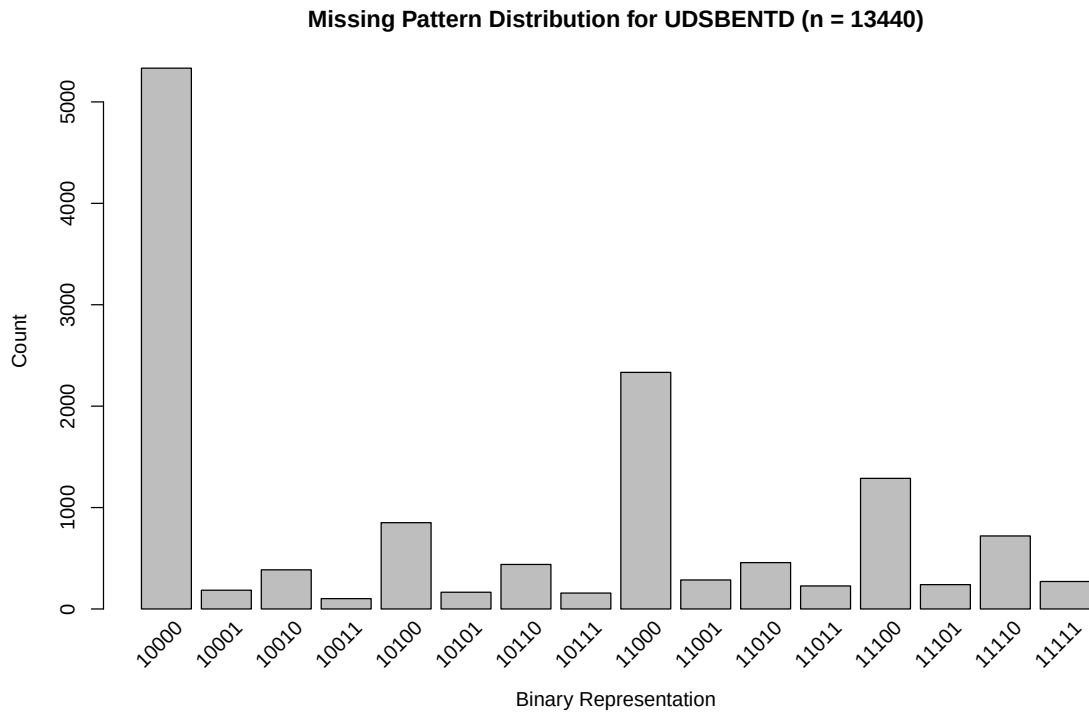


Figure 4.6: This is the missing pattern distribution for the UDSBENTD variable over 5 years.

terestingly, they all share similar maximum depth. The parent-based modeling is generally able to recover the LNCMV principle for many of the patterns, including most of the monotone missing patterns. For example, in the parent-based modeling tree graph, we observe the path $11100 \rightarrow 11000 \rightarrow 10000$, which is exactly the LNCMV principle applied to these patterns. On the other hand, the child-based modeling appears to incorporate a mix of both LNCMV and RNCMV principles when assigning a pattern to a given parent. Because the patterns 10000, 11000, 10100, and 10010 each share similar ancestors in both the LNCMV and parent-based modeling tree graph, we would expect that the two tree graphs lead to similar fitted distributions at the end.

In Figure 4.8, we plot the results of the fitted model $p(x|R = 1_d)$ using a mixture of

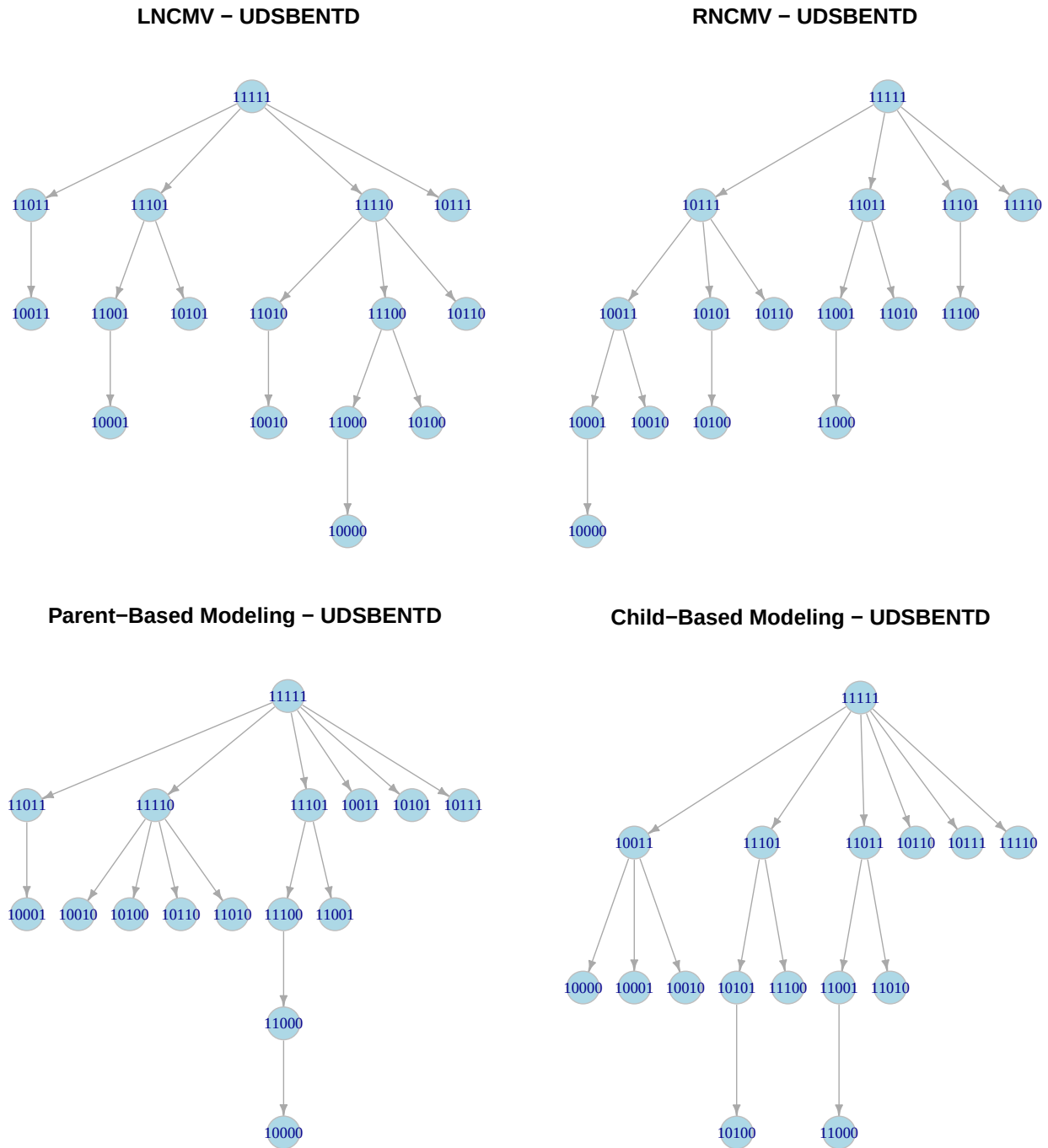


Figure 4.7: These are the tree graphs used for the UDSBENTD data analysis.

binomial products. We fit it using 5 clusters because of the recommendation from BIC. From the first five panels, we can see that it is roughly able to capture the shape of the marginal distributions. In the sixth panel, we include plots of the five clusters, each represented by a latent trajectory over the five time points. Each cluster is represented by a curve with the observed average CDR score reported at each dot. A given dot corresponds to the predicted mean UDSBENTD score from the model for a given cluster and year. Because the average CDR score is close to 0, the first two clusters represent cognitively normal people. There is also some evidence of a practice effect between years 1 and 3 because the scores increase over those years (Goldberg et al., 2015). The third clusters can be interpreted as mild cognitively impaired people because the average CDR score is close to 0.5. The fourth cluster appears to represent mild cognitively impaired people transitioning to dementia. The fifth cluster appears to be mostly mildly cognitive impaired or demented individuals.

In Figure 4.9, we plot our fitted model using the conjugate odds method against the observed marginal distributions as a diagnostic check. Because our method models both the imputation distribution and observed data distribution in one shot, it is important to perform this diagnostic check to have confidence in the imputation distribution results. The parent and LNCMV graphs have similar results while the CCMV and RNCMV graphs have similar results. So, we report results from RNCMV, parent-based modeling, and child-based modeling. All resulting models for the observed-data distribution appear to fit the data reasonably well in the majority of settings and the complete-case distribution (CCA) generally fails to capture the peak at 0 for most of the observed data distributions.

For the observed marginal distributions for patterns 11000 and 10010, the fit from the different tree graphs is comparable. From the marginal distributions $p(x_1|R = 10000)$ and $p(x_1|R = 10100)$, we see that the parent-based modeling tree graph yields fitted models that generally approximate the observed distributions better. Thus, for the following plots in

Figure 4.10, we report the fitted result for $p(x|R = r)$ using parent-based and contrast it with imputing with *mice* and then fitting a mixture of binomial products model. We see that the clusters across the different patterns 10000, 11000, 10100, and 10010 are generally

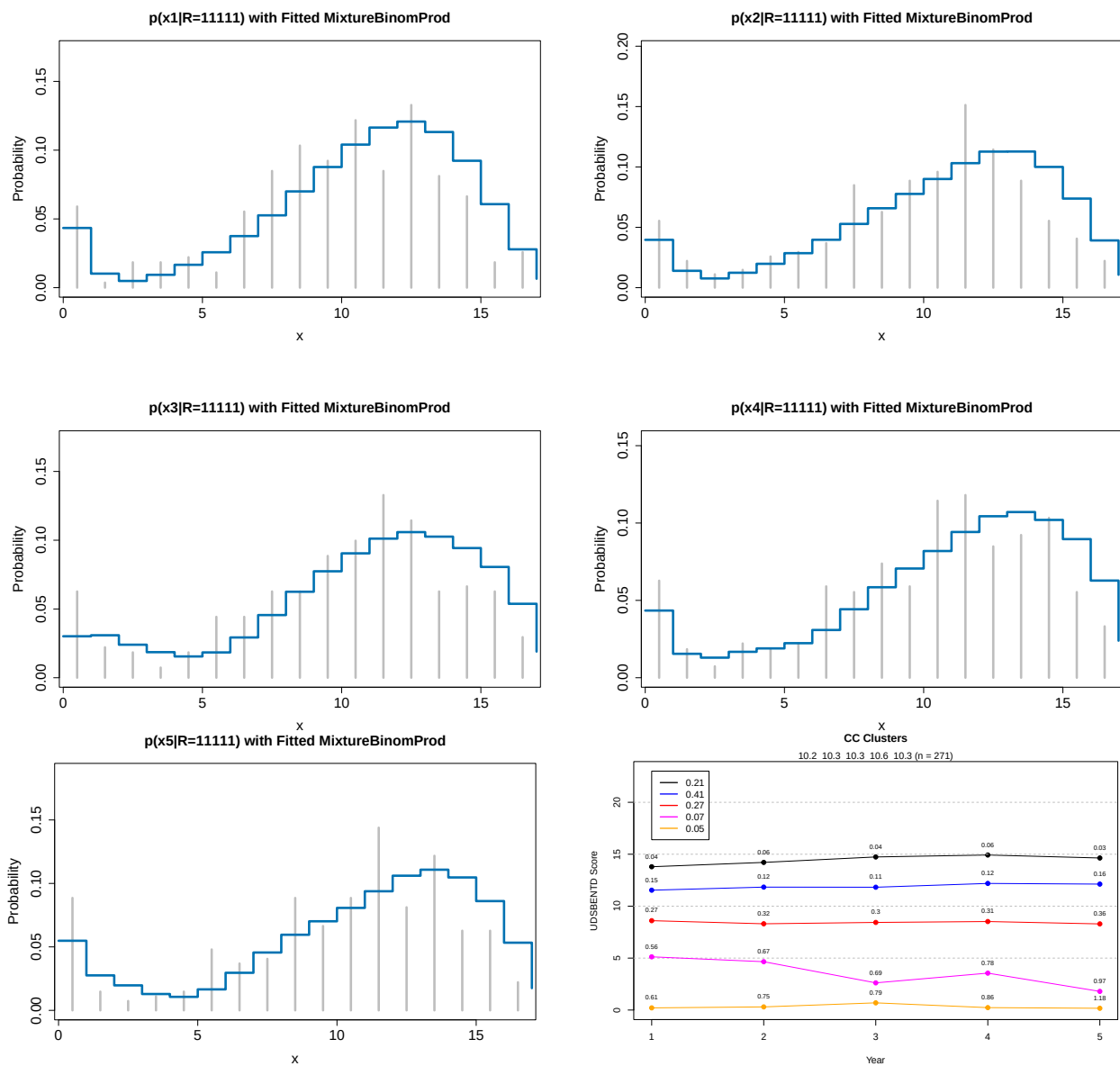


Figure 4.8: These are the results on fitting the mixture of binomial products model on the complete case distribution.

very similar for parent-based, but the weights change. For example, for $R = 10000$, there is more weight towards the unhealthier clusters. We also note that *mice* yields relatively similar clusters as well in terms of trends, but it may suffer from the model incompatibility problem (Meng, 1994), is longer to fit, and cannot handle MNAR data.

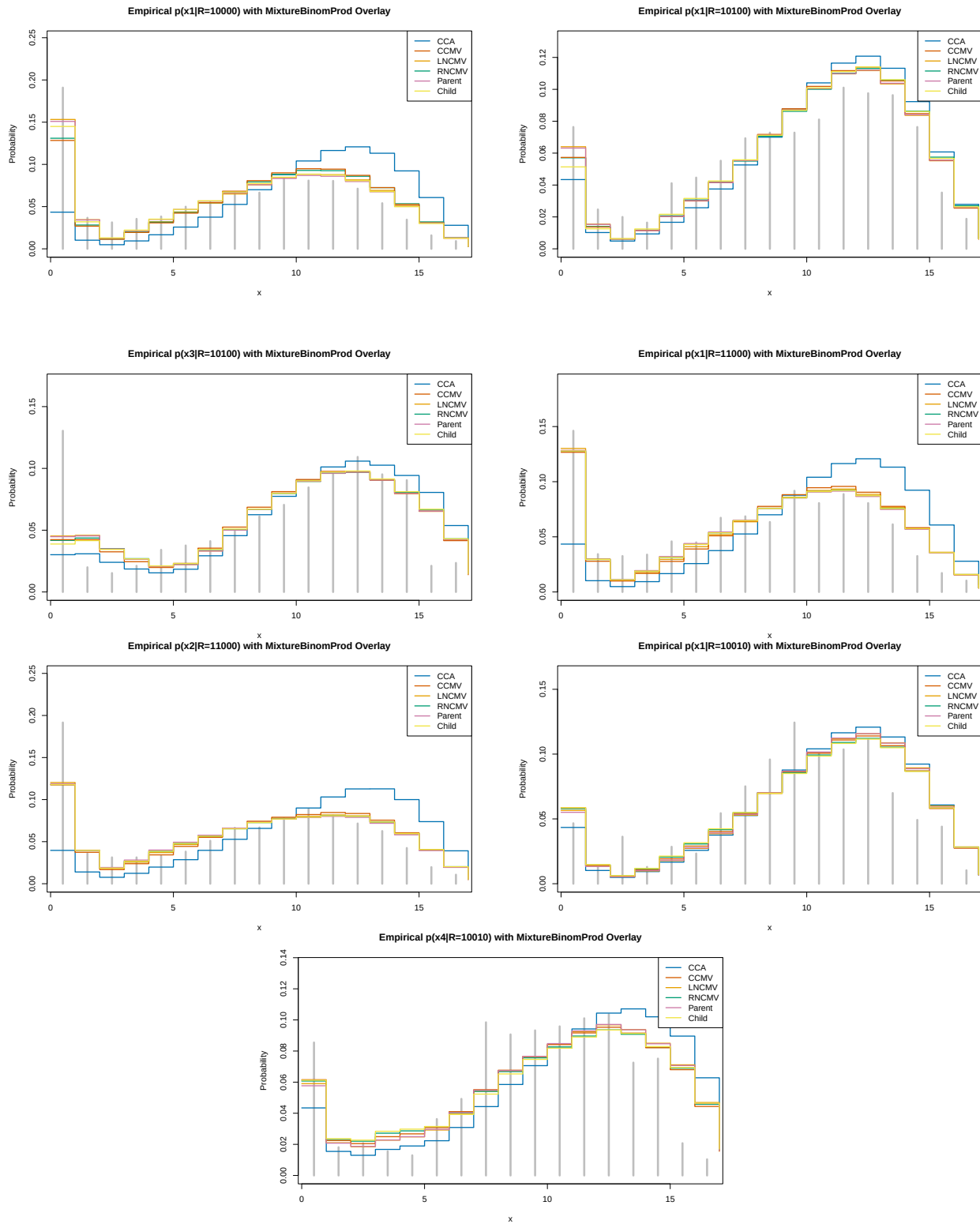


Figure 4.9: These are the results of the implied fitted distributions from different tree graphs on the observed marginal distributions for patterns 10000, 11000, 10100, and 10010. The parent and LNCMV graphs have similar results while CCMV and RNCMV graphs have similar results.

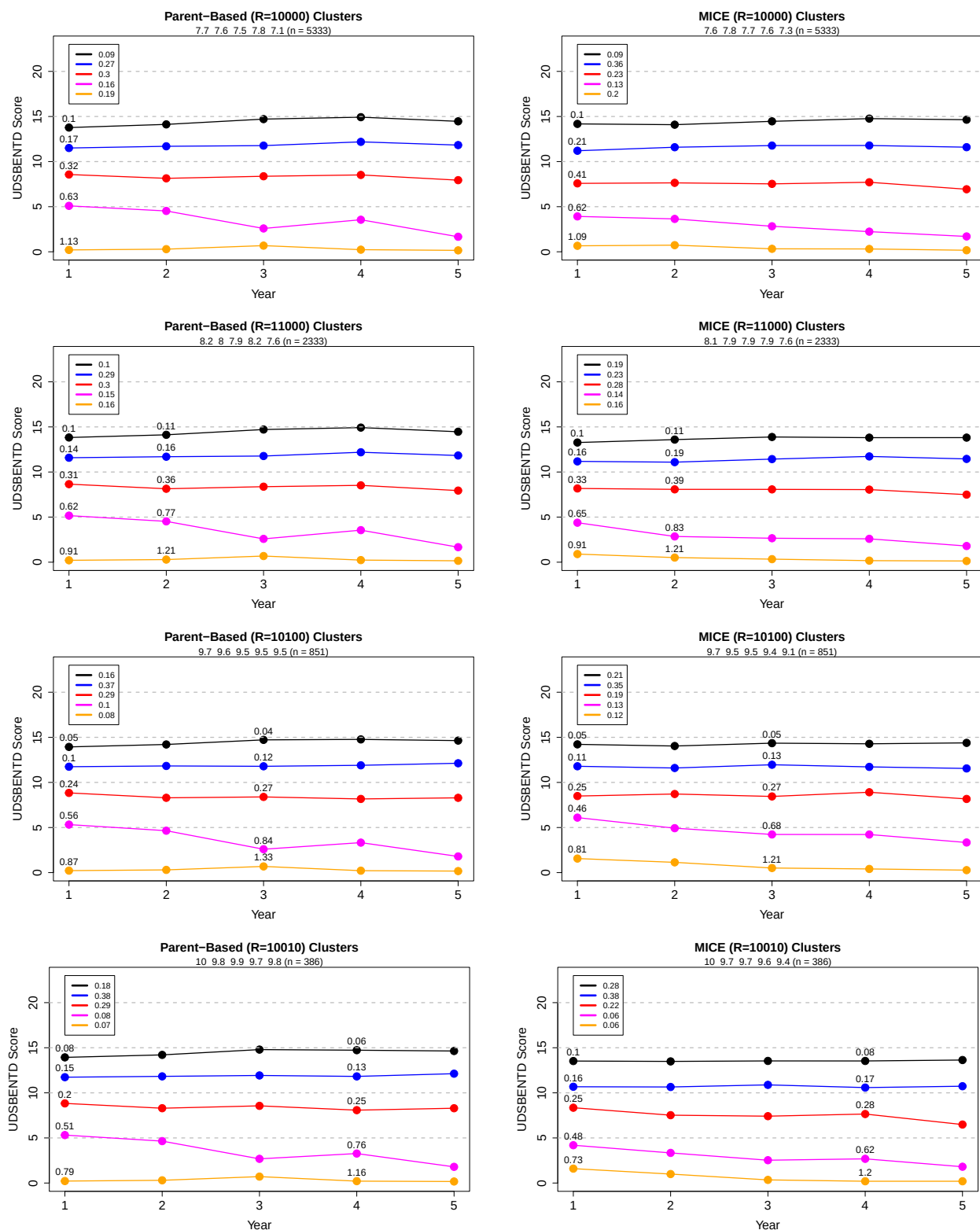


Figure 4.10: These are the results of the fitted trajectories. Each row corresponds to a pattern with 10000, 11000, 10100, and 10010, respectively. Column 1 is the parent-based modeling tree graph assumption, and column 2 is fitting the model after *mice* imputation.

9.3 Kernel density estimation

We now consider data in the continuous setting. We consider white vinho verde wine samples from the north of Portugal. This data consists of 4898 observations and was originally collected to model wine quality based on physicochemical tests. We select three continuous variables to study the modeling effect: pH, sulphate, and alcohol levels. Initially, we normalize the data such that it has a mean 0 and standard deviation 1.

We generate the missing data via a missing not at random mechanism 100 times through a tree graph and a prespecified selections odds model. On each iteration, we consider four density estimators for each conditional distribution $p(x|R = r)$. The first is a multivariate kernel density estimator using a Gaussian kernel on the complete-case data. Then, we construct our tree graph KDE, where we exploit the conjugate odds property with the Gaussian kernel. We also include an available case marginal Gaussian KDE estimator, where we fit the distribution based on all data that is observed for that dimension. For example, if we are considering dimension 3, then we pool the data from patterns 111, 101, and 001, and fit a one-dimensional KDE $\hat{p}(x_3|R \in \{001, 101, 111\})$. One clear disadvantage of the available case KDE is that we are unable to construct a joint KDE. Additionally, when there is missingness, we also perform *mice* imputation 20 times and construct the multivariate KDE on the *mice* imputed data.

In Figure 4.11, for each pattern–dimension pair, we plot the marginalized KDEs averaged over all 100 iterations. Of primary interest, we plot the tree graph KDE obtained after applying the conjugate odds property. For comparison, we also plot the complete case KDE, the available case KDE, and when the data is missing, the *mice* KDE. Since we have access to the true data and generate the missingness ourselves, we also can construct the oracle KDE, based on the true data. Thus, we include the oracle KDE, which is the KDE fitted using the true data. We expect the tree graph KDE to agree with the oracle KDE, and

largely, we observe that the tree graph KDE is able to generally identify the same shape as the oracle KDE. In contrast, the competing kernel density estimators generally do not capture the correct shape of the distribution, and it is clear they have different means and modes. We note that imputing with *mice* and fitting a KDE provides a similar result to the available case KDE, but it is not similar to the oracle KDE. This provides further evidence of the need to be careful when applying *mice*, especially when the data is missing not at random.

As we mentioned before, our method can model the missing data distribution and the observed data distribution in one shot, so we report both results. In the cases where we modeling the missing data distribution, we outline the plot in magenta. In the cases where we model the observed data distribution, we outline the plot in blue. Plots that are outlined in blue can be viewed more as diagnostic plots.

Each row of Figure 4.11 corresponds to the marginal distributions for patterns 110, 101, and 001, respectively. For pattern 110, the first and second plots correspond to marginal observed data distributions, and the third plot corresponds to a marginal missing data distribution. For pattern 101, the first and third plots correspond to marginal observed data distributions, and the second plot corresponds to a marginal missing data distribution. For pattern 001, the third plot corresponds to a marginal observed data distribution, and the first and second plots correspond to marginal missing data distributions, respectively.

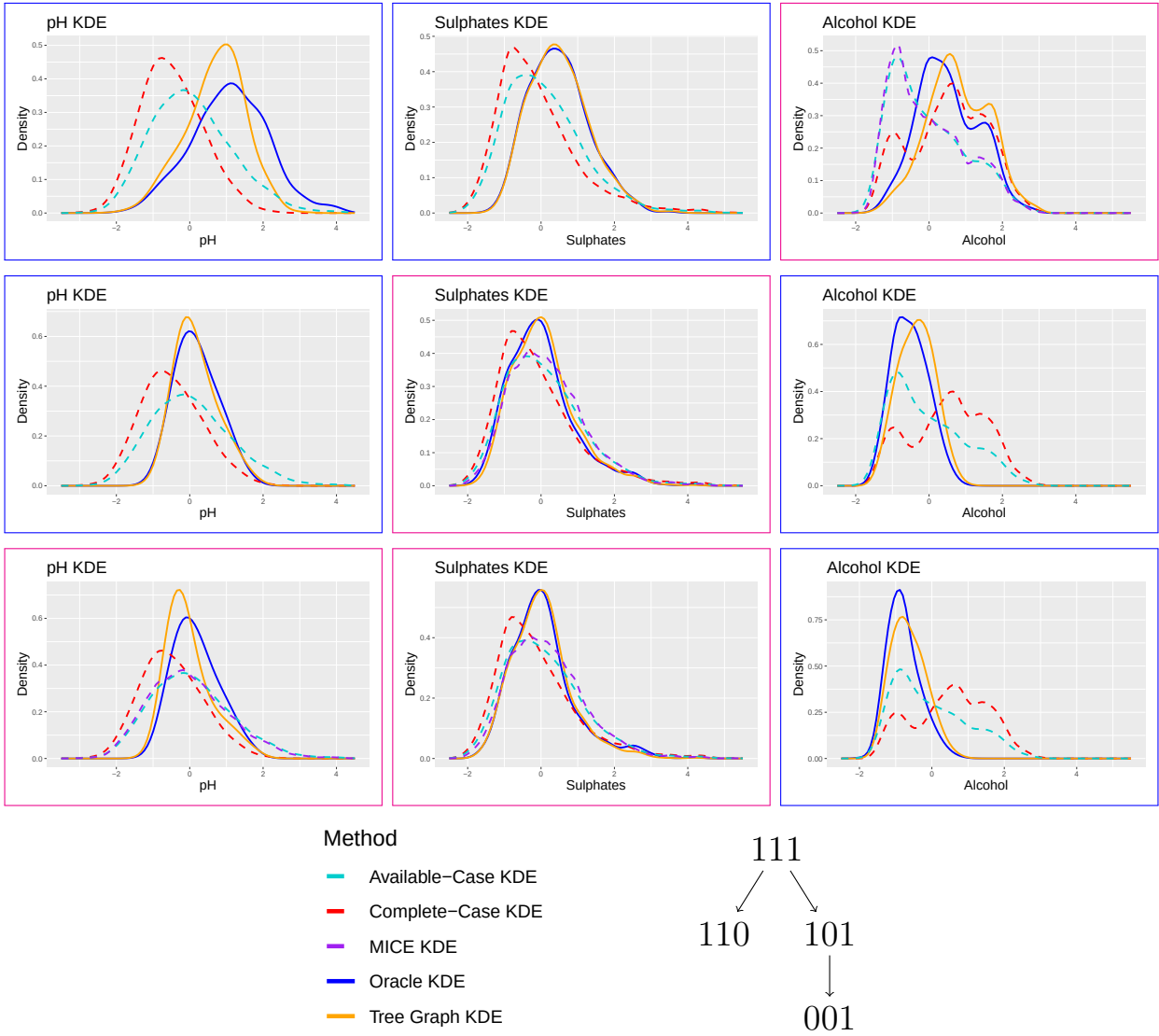


Figure 4.11: Each column corresponds to dimensions 1, 2, and 3, respectively. Rows 1, 2, and 3 depicts the fitted KDEs for the pattern 110, 101, and 001, respectively. The last row includes the plot of the tree graph used to generate the missignness.

10 Discussion

In this chapter, we introduced a new strategy for modeling multivariate missing not at random data. This strategy combines two frameworks: 1) the tree graph framework for identifying the selection odds and 2) the conjugate odds property to ensure simple modeling. We demonstrate that tree graphs form an incredibly rich subclass of the general pattern graph. Each tree graph represents a missing not at random assumption and provides an elegant form of the selection odds, thereby overcoming a shortcoming of a general pattern graph approach. Moreover, the conjugate odds property is introduced and used to model all distributions of the form $p(x|r)$. We provide examples of the conjugate odds property with applications to mixtures of exponential family models. Furthermore, we provide simulations to demonstrate finite sample performance of our method, and we analyze two data sets comprising multivariate discrete and multivariate continuous data.

There are several ways to extend the ideas in this chapter. As presented, our framework works using mixture of exponential family models with logistic odds and mixture of Pareto distributions with power law odds. Previous proposed models from the literature such as the mixture of binomial products (Suen and Chen, 2023) and the Rasch model (Rasch, 1960) could be utilized here. It would be interesting to explore other parametric families and determine what others might fall under this framework. When fitting the $p(x|1_d)$ model for the neuropsychological test score data, we used a mixture of binomial products model to account for the dependence between test scores from different time points. However, since the neuropsychological test score data is longitudinal by nature, there may be a more sophisticated way to incorporate time in the $p(x|1_d)$ model.

Furthermore, while we discussed multiple methods for choosing a tree graph and performing sensitivity analysis, this remains an active area of research. Since a tree graph is a nonparametric identifying restriction that cannot be rejected by the observed data, it is

critical to choose one that is plausible and aligns with our prior knowledge. We have outlined a few different principles, but there may be more extensions. For example, when performing a parent-based or child-based modeling approach, one could consider distributional distances such as the Wasserstein or Hellinger metrics. A natural way to conduct sensitivity analysis is through exponential tilting (Kim and Yu, 2011; Shao and Wang, 2016; Zhao et al., 2017), but there may be more other methods that exploit the geometry of the pattern graph space to interpolate between different tree graph assumptions. We leave this for future work.

Appendix C: Chapter 3 Appendices

C.1 Proofs

C.1.1 Pattern graphs

Proof of Theorem 4.1. We start from the pattern-mixture model factorization of the pattern graph and proceed with a series of equivalent statements to show the selection model factorization holds as well. For every $r \in \mathcal{R}$, we have by the pattern-mixture model factorization,

$$p(x_{\bar{r}}|x_r, R = r) = p(x_{\bar{r}}|x_r, R \in \text{PA}(r)).$$

Thus, it follows that for every $r \in \mathcal{R}$,

$$\begin{aligned} p(x_{\bar{r}}|x_r, R = r) = p(x_{\bar{r}}|x_r, R \in \text{PA}(r)) &\iff \frac{p(x, R = r)}{p(x_r, R = r)} = \frac{p(x, R \in \text{PA}(r))}{p(x_r, R \in \text{PA}(r))} \\ &\iff \frac{p(x, R = r)}{p(x, R \in \text{PA}(r))} = \frac{p(x_r, R = r)}{p(x_r, R \in \text{PA}(r))} \\ &\iff \frac{P(R = r|x)}{P(R \in \text{PA}(r)|x)} = \frac{P(R = r|x_r)}{P(R \in \text{PA}(r)|x_r)}. \end{aligned}$$

By positivity, division by $p(x_r, r)$ is valid. This is precisely the selection model factorization, so they are equivalent. $\square$

C.1.2 Tree graphs and associated properties

Proof of Proposition 4.1. Any given tree graph is equivalent to a missing not at random assumption. Pick any missing data pattern r_ℓ with associated path $1_d \rightarrow r_1 \rightarrow r_2 \rightarrow \dots \rightarrow r_\ell$. By Theorem 4.1, the selection odds factorizes with respect to the tree graph, so we have the

following decomposition

$$\begin{aligned}
 \frac{P(R = r_\ell | X)}{P(R = r_{1_d} | X)} &= \prod_{i=1}^{\ell} \frac{P(R = r_i | X)}{P(R = r_{i-1} | X)} \\
 &= \prod_{i=1}^{\ell} \frac{P(R = r_i | X_{r_i})}{P(R = r_{i-1} | X_{r_i})} \\
 &= \prod_{i=1}^{\ell} f_{r_i}(X_{r_i})
 \end{aligned}$$

for functions $\{f_{r_i}\}_{r_i \in \mathcal{R}}$. Multiplying both sides by $P(R = 1_d | X)$ implies that $P(R = r_\ell | X)$ depends on $X_{\bar{r}_\ell}$, which implies it is MNAR assumption.

Additionally, from Theorem 4.1, we know that the pattern-mixture model factorization holds. Factor the full data distribution as

$$p(x, r) = p(x_{\bar{r}} | x_r, r) \cdot p(x_r | r) \cdot p(r).$$

The extrapolation distributions $\{p(x_{\bar{r}} | x_r, r)\}_{r \in \mathcal{R}}$ are the only distributions not identified from the observed data. However, the tree graph provides a way to identify each distribution from the observed data. As above, pick any missing data pattern r_ℓ with associated path $1_d \rightarrow r_1 \rightarrow r_2 \rightarrow \dots \rightarrow r_\ell$. Then, we have

$$\begin{aligned}
 p(x_{\bar{r}_\ell} | x_{r_\ell}, r_\ell) &\stackrel{T}{=} p(x_{\bar{r}_\ell} | x_{r_\ell}, r \in \text{PA}_T(r_\ell)) \\
 &= p(x_{\bar{r}_\ell} | x_{r_\ell}, r \in \text{PA}_T(r_\ell))
 \end{aligned}$$

Thus, the full data distribution is nonparametrically identified. $\square$

Proof of Proposition 4.2. To prove equivalence of all three statements, we prove in a cycle.

- (1 $\Rightarrow$ 2) Definition 4.2 implies the single parent property. We prove the contrapositive. Suppose there exists one pattern r with two parents, labeled s_1 and s_2 . Then, the path from $1_d \rightarrow \cdots \rightarrow s_1 \rightarrow r$ and $1_d \rightarrow \cdots \rightarrow s_2 \rightarrow r$ both exist in the graph, which means that there is more than one path to r from 1_d .
- (2 $\Rightarrow$ 3) Next, suppose that every pattern $r \neq 1_d$ in T has exactly one parent. There are exactly 2^d patterns in T with 1_d as the source, so there are $2^d - 1$ that require a parent. Thus, T must contain at least $2^d - 1$ edges, and since every pattern $r \neq 1_d$ has only parent, there are no more $2^d - 1$ total edges.
- (3 $\Rightarrow$ 1) Lastly, we prove by contradiction. Suppose that instead of $2^d - 1$ edges, there are 2^d total edges in T . By the Pigeonhole Principle, there exists one pattern with at least $\lceil 2^d / (2^d - 1) \rceil = 2$ parents. Denote this pattern by r . If r has at least 2 parents, then there are at least two paths from 1_d to r , which implies this is not a tree graph, thereby resulting in a contradiction.

□

Lemma 4.1. *The following combinatorial identities hold*

$$\sum_{m=0}^d \binom{d}{m} = 2^d, \quad \sum_{m=0}^d m \binom{d}{m} = d \cdot 2^{d-1}.$$

Proof of Lemma 4.1. We will prove both equations using combinatorial arguments. We prove the first equation first. Observe that the LHS counts the number of ways to form a committee of size 0 to d from d individuals. Alternatively, one can count the number of committees by noting that each of the d individuals has two choices: to be in the committee or not. We then obtain 2^d on the RHS. Thus, equality holds.

For the second equation, note that the LHS counts the number of ways to form a committee of any size with a leader. We can alternatively count this by selecting the leader first

from d individuals and then forming a committee from the remaining $d - 1$ individuals. This precisely gives us $d \cdot 2^{d-1}$ on the RHS, and equality holds. $\square$

Proof of Proposition 4.3. In a regular pattern graph, the observed variables that a pattern has is exactly a subset of its parents' observed variables. Therefore, every missing pattern r has $2^m - 1$ parents, where m is the number of 0s in r . There are $\binom{d}{m}$ patterns with m 0s. Thus, using Property 2 and since m ranges from 1 to d , we have

$$|\mathcal{T}_d| = \prod_{m=1}^d (2^m - 1) \binom{d}{m}.$$

Since $2^{m-2} < 2^m - 1$ when $m > \log_2(4/3) \approx 0.415$, we have the following lower bound

$$\begin{aligned} |\mathcal{T}_d| &\geq \prod_{m=1}^d (2^{m-2}) \binom{d}{m} \\ &= 2^{\sum_{m=1}^d (m-2) \binom{d}{m}}. \end{aligned}$$

Focusing on the term in the exponent, we have

$$\begin{aligned} \sum_{m=1}^d (m-2) \binom{d}{m} &= \sum_{m=1}^d m \binom{d}{m} - 2 \sum_{m=1}^d \binom{d}{m} \\ &= \sum_{m=0}^d m \binom{d}{m} - 2 \left(\sum_{m=0}^d \binom{d}{m} - 1 \right) \\ &= d \cdot 2^{d-1} - 2(2^d - 1) \\ &= (d/2 - 2) \cdot 2^d + 2, \end{aligned}$$

where the second to last inequality can be obtained via standard combinatorial arguments

(see Lemma 4.1 for completeness). This implies that

$$|\mathcal{T}_d| \geq 2^{(d/2-2) \cdot 2^d + 2} = 2^{\Omega(d \cdot 2^d)},$$

which is precisely super-exponential. $\square$

Proof of Proposition 4.7. First, we bound the size of $T \in \mathcal{T}_{\text{GNCMV}}$. For every pattern with m observed variables, there are a total of $\binom{d}{m}$ patterns. Thus, we have

$$|\mathcal{T}_{\text{GNCMV}}| = \prod_{m=0}^{d-1} \binom{d}{m+1}^{\binom{d}{m}}.$$

For sufficiently large d , we further obtain

$$\log |\mathcal{T}_{\text{GNCMV}}| = \sum_{m=0}^{d-1} \binom{d}{m} \log \binom{d}{m+1} \geq \sum_{m=0}^{d-1} \binom{d}{m} = \Omega(2^d).$$

So, $\mathcal{T}_{\text{GNCMV}}$ forms a large class. Next, for every $T \in \mathcal{T}_{\text{GNCMV}}$, we prove the following two properties:

1. It achieves the maximum possible depth of d .

Since every pattern is the farthest it can be from 1_d , this implies that the longest chain in the graph is formed via the path from 1_d to 0_d . This chain has length d , which implies that this NCMV graph has the maximum possible depth, in contrast to CCMV.

2. Every pattern r in T is positioned at the maximum possible distance from the source node 1_d , thereby corresponding to the most information flow.

By construction, every pattern $r \neq 1_d$ has a parent s such that s contains exactly one more observed variable than r . Therefore, this implies that length of the path from

the source node 1_d to any pattern that contains k 0s is exactly k , which is also the maximum distance away from the source node it can be.

□

C.1.3 Conjugate odds properties

Proof of Proposition 4.4. Suppose that $p(x|A = a)$ belongs to the K -mixture model

$$\mathcal{M}_K(\mathcal{P}) := \left\{ p = \sum_{j=1}^K w_j p_j \mid p_j \in \mathcal{P}, \sum_{j=1}^K w_j = 1, w_j > 0 \forall j \right\}$$

such that every component is an element of $\mathcal{P}$, and the odds $\mathcal{O}(\mathcal{P})$ is a conjugate odds for $\mathcal{P}$.

To be precise, suppose we can write $p(x|A = a)$ as

$$p(x|A = a) = \sum_{j=1}^K w_j p_j(x|A = a)$$

for positive weights w_j that sum to 1. Now, suppose that $P(A = a'|x)/P(A = a|x)$ is conjugate for $p_j(x|A = a)$ for all j . Then, we have

$$\begin{aligned} p(x|A = a') &\propto p(x|A = a) \cdot \frac{P(A = a'|x)}{P(A = a|x)} \\ &= \sum_{j=1}^K w_j p_j(x|A = a) \cdot \frac{P(A = a'|x)}{P(A = a|x)} \\ &= \sum_{j=1}^K w_j \cdot \zeta_j \cdot p_j(x|A = a') \end{aligned}$$

for some $\{\zeta_j\}_j$ that are all positive due to the conjugacy of the odds model. Finally, this

implies that

$$p(x|A = a') = \sum_{j=1}^K \tilde{w}_j p_j(x|A = a')$$

for some set of perturbed weights $\{\tilde{w}_j\}_j$. So, we have that $p(x|A = a')$ is also K -mixture model with components belonging to $\mathcal{P}$, and the result follows. $\square$

Proof of Proposition 4.5. We assume the $p(x|A = a)$ has the following exponential family parameterization

$$p(x|A = a) = h(x)g(\eta) \exp(\eta^\top T(x)).$$

We also further assume that the odds is a logistic regression and linear in the sufficient statistic

$$\log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma^\top T(x).$$

We have the following equality

$$p(x|A = a') = \frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) \Big/ \int_{-\infty}^{\infty} \frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) \, dx.$$

Focusing on the unnormalized distribution, we have

$$\begin{aligned} \frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) &= \exp(\gamma_0 + \gamma^\top T(x)) \cdot h(x)g(\eta) \exp(\eta^\top T(x)) \\ &= \exp(\gamma_0)h(x)g(\eta) \exp((\eta + \gamma)^\top T(x)). \end{aligned} \quad (4.2)$$

As $h(x) \exp((\eta + \gamma)^\top T(x))$ is an unnormalized exponential family distribution with natural parameter $\eta + \gamma$, it follows that

$$\int_{-\infty}^{\infty} h(x) \exp((\eta + \gamma)^\top T(x)) \, dx = \frac{1}{g(\eta + \gamma)}.$$

Returning to equation (4.2), we see that

$$\begin{aligned}
 p(x|A = a') &= \frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) \Big/ \int_{-\infty}^{\infty} \frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) \, dx \\
 &= \frac{\exp(\gamma_0) \cdot h(x) \cdot g(\eta) \cdot \exp((\eta + \gamma)^\top T(x))}{\exp(\gamma_0) \cdot g(\eta) \cdot 1/g(\eta + \gamma)} \\
 &= h(x)g(\eta + \gamma) \exp((\eta + \gamma)^\top T(x)), \\
 &= h(x)g(\eta') \exp((\eta')^\top T(x))
 \end{aligned}$$

with $\eta' := \eta + \gamma$. Finally, solving for γ_0 , we have

$$\exp(\gamma_0) := \frac{g(\eta')}{g(\eta)} \cdot \frac{P(A = a')}{P(A = a)} \iff \gamma_0 := \log \frac{P(A = a')}{P(A = a)} + \log \frac{g(\eta')}{g(\eta)},$$

as desired. □

Proof of Corollary 4.1. Suppose that $P(A = a'|x)/P(A = a|x)$ is modeled using a logistic regression. That is, we model the log-odds like

$$\log \frac{P(A = a'|x)}{P(A = a|x)} = g(x)$$

for some function g . Then, we have

$$p(x|A = a') \propto p(x|A = a) \cdot \frac{P(A = a'|x)}{P(A = a|x)} = p(x|A = a) \cdot \exp(g(x)),$$

and this is precisely an exponential tilting, as desired. □

Proof of Corollary 4.2. By definition, we can write $p(x|A = a')$ as

$$p(x|A = a') \propto p(x|A = a) \cdot \frac{P(A = a'|x)}{P(A = a|x)}$$

$$\begin{aligned}
&= \exp(\gamma_0 + \gamma^\top T(x)) \sum_{k=1}^K w_k \cdot h(x) g(\eta_k) \exp(\eta_k^\top T(x)) \\
&= \sum_{k=1}^K w_k \cdot \exp(\gamma_0) h(x) g(\eta_k) \exp((\eta_k + \gamma)^\top T(x)).
\end{aligned}$$

Then, simplifying with algebra and renormalizing yields

$$p(x|A = a') = \sum_{k=1}^K \tilde{w}_k \cdot h(x) g(\eta_k + \gamma) \exp((\eta_k + \gamma)^\top T(x)),$$

where

$$\tilde{w}_k := \frac{w_k \cdot g(\eta_k)}{g(\eta_k + \gamma)} \bigg/ \sum_{k'=1}^K \frac{w_{k'} \cdot g(\eta_{k'})}{g(\eta_{k'} + \gamma)}.$$

This concludes the proof. Additionally, we note that in this proof we assumed that each mixture was the same distribution, but this argument generalizes to other distributions. For example, instead of just a mixture of Gaussians, one could have a mixture of a Gaussian and a Binomial. $\square$

Proof of Proposition 4.6. We assume the $p(x|A = a)$ has the following Pareto distribution parameterization

$$p(x|A = a; \alpha, \beta) = \begin{cases} \frac{\alpha \beta^\alpha}{x^{\alpha+1}} & x \geq \beta \\ 0 & \text{o.w.} \end{cases}.$$

We also further assume that the odds obeys the following parametric model

$$\frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 x^\gamma, \quad \gamma_0 := \frac{P(A = a')}{P(A = a)} \cdot \frac{\alpha'}{\alpha} \cdot \beta^{\alpha' - \alpha}.$$

We have the following equality

$$p(x|A = a') = \frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) \bigg/ \int_{\beta}^{\infty} \frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) \, dx.$$

Focusing on the unnormalized distribution, we have

$$\frac{P(A = a'|x)}{P(A = a|x)} \cdot p(x|A = a) = \frac{\alpha\beta^\alpha\gamma_0}{x^{\alpha-\gamma+1}}.$$

The normalizing constant must be

$$\int_{\beta}^{\infty} \frac{\alpha\beta^\alpha\gamma_0}{x^{\alpha-\gamma+1}} dx = \frac{\alpha\beta^\alpha\gamma_0}{(\alpha-\gamma)\beta^{\alpha-\gamma}} = \frac{\alpha\beta^\gamma\gamma_0}{\alpha-\gamma}.$$

Thus, we have

$$p(x|A = a'; \alpha', \beta) = \begin{cases} \frac{\alpha'\beta^{\alpha'}}{x^{\alpha'+1}} & x \geq \beta \\ 0 & \text{o.w.} \end{cases}$$

for $\alpha' := \alpha - \gamma$. Finally, solving for γ_0 , we have

$$\gamma_0 := \frac{P(A = a')}{P(A = a)} \cdot \frac{\alpha'}{\alpha} \cdot \frac{\beta^{\alpha'}}{\beta^\alpha},$$

as desired. □

Proof of Corollary 4.4. Suppose we have the decomposition $X := (X_1, X_2)$, where $X_1 \perp X_2 \mid A$. Let $p(x_1|A = a)$ and $p(x_2|A = a)$ be two exponential family distributions such that

$$p(x_1|A = a) = h_1(x_1)g_1(\eta_1) \exp(\eta_1^\top T_1(x_1)),$$

$$p(x_2|A = a) = h_2(x_2)g_2(\eta_2) \exp(\eta_2^\top T_2(x_2)).$$

Then, we have

$$p(x|A = a) = p(x_1|A = a) \cdot p(x_2|A = a)$$

$$= h_1(x_1)h_2(x_2)g_1(\eta_1)g_2(\eta_2)\exp(\eta_1^\top T_1(x_1) + \eta_2^\top T_2(x_2)).$$

Finally, it follows that

$$\begin{aligned} p(x|A = a') &\propto p(x|A = a) \cdot P(A = a'|x)/P(A = a|x) \\ &= h_1(x_1)h_2(x_2)g_1(\eta_1)g_2(\eta_2)\exp((\eta_1 + \gamma_1)^\top T_1(x_1) + (\eta_2 + \gamma_2)^\top T_2(x_2)). \end{aligned}$$

Thus, the natural parameter is $\zeta := (\eta_1 + \gamma_1, \eta_2 + \gamma_2)$ and $T(x) := (T_1(x_1), T_2(x_2))$, as desired. $\square$

C.1.4 Tree graphs and conjugate odds

Proof of Theorem 4.2. By assumption, the missigness mechanism can be specified using a tree graph. Then, by Proposition 4.1, we obtain an identification formula for the selection odds

$$\frac{P(R = r_\ell|X)}{P(R = 1_d|X)} \stackrel{\text{tree graph}}{=} \prod_{i=1}^{\ell} \frac{P(R = r_i|X_{r_i})}{P(R = r_{i-1}|X_{r_i})},$$

where $1_d =: r_0 \rightarrow r_1 \rightarrow r_2 \rightarrow \cdots \rightarrow r_\ell$ is the unique path in the tree graph from the source 1_d to pattern r_ℓ .

We proceed with the proof inductively. First, partition the set of patterns $\mathcal{R}$ into $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_d$, where $\mathcal{R}_k$ denotes the set of patterns in the tree graph that are exactly k edges away from the source node 1_d . Here, $\mathcal{R}_0$ is trivially the set $\{1_d\}$.

For the base case, it is sufficient to consider the set $\mathcal{R}_1$, and note that for any $r \in \mathcal{R}_1$, we have

$$p(x|R = r) \propto p(x|R = 1_d) \cdot \frac{P(R = r|x)}{P(R = 1_d|x)}.$$

Therefore, by conjugacy of the odds, $p(x|R = r)$ is the same probability family as $p(x|R = 1_d)$ for any $r \in \mathcal{R}_1$.

Next, fix k , and suppose that for all $r \in \mathcal{R}_k$, $p(x|R = r)$ is the same probability family as $p(x|R = 1_d)$. Then, for any $s \in \mathcal{R}_{k+1}$, there exists $r' \in \mathcal{R}_k$ such that $r' \rightarrow s$ (r' is the unique parent of s). It follows that

$$p(x|R = s) \propto p(x|R = r) \cdot \frac{P(R = s|x)}{P(R = r|x)}.$$

Again, by the inductive hypothesis and conjugacy of the selection odds, it must follow that $p(x|R = s)$ also belongs to the same probability family as $p(x|R = r)$ and thus, also $p(x|R = 1_d)$ by transitivity. $\square$

C.1.5 Asymptotic theory

Proof of Proposition 4.8. We prove this directly. Suppose that the patterns do not have a direct parent-child relationship.

Let $\mathcal{X}_r := \{X_{i,R_i} \mid R_i = r\}$ be the observed data under pattern r , so by definition, $\mathcal{X}_a \cap \mathcal{X}_{a'} = \emptyset$ for $a \neq a'$. There are two types of models fit on the data. The first model is the complete case model, which is only using data $\mathcal{X}_{1_d}$. The pattern 1_d contains no siblings. Any odds based on the patterns with 1_d .

The conjugate odds model $O_r(x) := P(R = r|x)/P(R = r'|x)$ is fit using the data $\mathcal{X}_r \cup \mathcal{X}_{r'}$. If two conjugate odds models are fit using completely separate data, then their resulting parameter estimates will be independent (this can be viewed as a form of sampling splitting).

Now, consider two distinct patterns r and s such that $r' \rightarrow r$ and $s' \rightarrow s$. Suppose we fit two conjugate odds models $O_r(x)$ and $O_s(x)$ using the data $\mathcal{X}_r \cup \mathcal{X}_{r'}$ and $\mathcal{X}_s \cup \mathcal{X}_{s'}$. Then,

failing to satisfy the first property necessarily implies that $s \neq r'$ and $r \neq s'$. Failing to satisfy the second property implies that $r' \neq s'$. All together, we have

$$\begin{aligned} (\mathcal{X}_r \cup \mathcal{X}_{r'}) \cap (\mathcal{X}_s \cup \mathcal{X}_{s'}) &= ((\mathcal{X}_r \cup \mathcal{X}_{r'}) \cap \mathcal{X}_s) \cup ((\mathcal{X}_r \cup \mathcal{X}_{r'}) \cap \mathcal{X}_{s'}) \\ &= \emptyset \cup \emptyset \\ &= \emptyset. \end{aligned}$$

So, the two models are fit on separate data sets. This implies that the estimated model parameters are independent, and therefore, have covariance 0. $\square$

Proof of Corollary 4.3. Consider every maximal clique in the undirected graph. If there is a path between two patterns in the undirected graph, then the estimated parameters for each of the models are correlated.

Moreover, for every submatrix that is determined by the maximal clique, the submatrix is full; that is, it contains only nonzero elements. $\square$

Proof of Proposition 4.9. We start by proving the first claim. In the CCMV case, the associated undirected graph is a clique, and the complete case data is used to fit every conjugate odds model $O_r(x) = O_r(x_r) := P(R = r|x_r)/P(R = 1_d|x_r)$. Therefore, the estimated parameters for all the conjugate odds models and the complete case model are all dependent. Thus, the correlation is nonzero, and CCMV provides the densest covariance matrix.

Now, we consider the second claim: any GNCMV assumption provides the sparsest asymptotic covariance matrix. The undirected graph associated with any GNCMV tree graph contains only maximal cliques of size 2, thereby leading to the sparsest possible matrix. $\square$

C.1.6 Proofs of results in the appendix

Proof of Theorem 4.3. Let $\mathcal{B} = \{j_1, j_2, \dots, j_B\} \subseteq [d]$ be a set of indices that correspond to several variables, and $\mathcal{R}_{\mathcal{B}} := \{s : s_j = 0, j \in \mathcal{B}\}$. Further, suppose that for any $r \in \mathcal{R}_{\mathcal{B}}$, $\text{Anc}_{G_1}(r) = \text{Anc}_{G_2}(r)$. To prove sufficiency, it suffices to show that the implied distribution $p(x_{j_1}, x_{j_2}, \dots, x_{j_B})$ is the same. We approach this from a pattern-mixture model standpoint.

Note that in a tree graph, a pattern's set of ancestors determines exactly the path from 1_d to that pattern. A given pattern's set of ancestors has a total ordering, and this total ordering uniquely determines the path from 1_d , as there is only a single path (in a tree graph). Therefore, since the patterns share the same ancestors in the two tree graphs, the patterns all have the same implied distributions. $\square$

Proof of Proposition 4.12. To show both claims, it suffices to show that the set of tree graphs is the smallest generating set for the set of pattern graphs. We do this by showing two facts: every generating set must be a superset of the set of tree graphs, and the set of tree graphs is a generating set.

First, note that any generating set must contain the set of tree graphs because every tree graph is minimal (see Proposition 4.2). Next, the set of tree graphs is a generating set. For any pattern graph G , construct an associated set of tree graphs $\mathcal{T}_G$, where each tree graph $T \in \mathcal{T}_G$ is made by choosing a single element in each parent set of G .

Then, note that performing this operation over all possible pattern graphs G and taking a union forms a generating set. More precisely, if we let $\mathcal{G}$ to be the set of all pattern graphs, we have

$$\bigcup_{G \in \mathcal{G}} \mathcal{T}_G \subseteq \mathcal{T}.$$

But also since $\bigcup_{G \in \mathcal{G}} \mathcal{T}_G$ is a generating set, we have

$$\bigcup_{G \in \mathcal{G}} \mathcal{T}_G \supseteq \mathcal{T}.$$

So, equality holds, and we are done. $\square$

C.2 Further Conjugate Odds Examples

C.2.1 Further logistic odds examples

Example 15 (Negative binomial). *The negative binomial distribution is widely used in modeling discrete data with overdispersion with one notable example in single cell RNA data. Suppose that $X|A = a \sim \text{NegBin}(r, p)$ with r known. The sufficient statistic is $T(x) = x$ with natural parameter $\eta = \log p$. Suppose that*

$$\log \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 + \gamma_1 x$$

where $\gamma = \gamma_1$. Then, via Proposition 4.5, $X|A = a'$ is negative binomially distributed with natural parameter

$$\eta' := \eta + \gamma_1 = \log p + \gamma_1.$$

Translating this back to the standard parameterization, we have

$$p' := \exp(\log p + \gamma_1).$$

Fact 1 (Product distribution of exponential family is still exponential family). *Let X_1 and X_2 be two independent exponential family distributions, and let $X = (X_1, X_2)$. This implies the following form*

$$p(x_1) = h_1(x_1)g_1(\eta_1)\exp(\eta_1^\top T_1(x_1)),$$

$$p(x_2) = h_2(x_2)g_2(\eta_2) \exp(\eta_2^\top T_2(x_2)).$$

Then,

$$p(x) = \underbrace{h_1(x_1)h_2(x_2)}_{h(x)} \underbrace{g_1(\eta_1)g_2(\eta_2)}_{g(\eta)} \exp(\eta^\top T(x))$$

is an exponential family distribution, where $\eta := (\eta_1, \eta_2)$, $g(\eta) := \prod_{j=1}^d g_j(\eta_j)$, and $T(x) := (T_1(x_1), T_2(x_2))$.

Corollary 4.4 (Product distribution). *Suppose we have the decomposition $X := (X_1, X_2)$, where $X_1 \perp X_2 \mid A$. Let $p(x_1|A = a)$ and $p(x_2|A = a)$ be two exponential family distributions such that*

$$p(x_1|A = a) = h_1(x_1)g_1(\eta_1) \exp(\eta_1^\top T_1(x_1)),$$

$$p(x_2|A = a) = h_2(x_2)g_2(\eta_2) \exp(\eta_2^\top T_2(x_2)).$$

Suppose that

$$\log P(A = a'|x)/P(A = a|x) = \gamma_0 + \gamma_1^\top T_1(x_1) + \gamma_2^\top T_2(x_2).$$

It follows that

$$\begin{aligned} p(x|A = a') &= \underbrace{h_1(x_1)h_2(x_2)}_{h(x)} \underbrace{g_1(\eta_1 + \gamma_1)g_2(\eta_2 + \gamma_2)}_{g(\zeta)} \exp((\eta_1 + \gamma_1)^\top T_1(x_1) + (\eta_2 + \gamma_2)^\top T_2(x_2)) \\ &= h(x)g(\zeta) \exp(\zeta^\top T(x)) \end{aligned}$$

with natural parameter $\zeta := (\eta_1 + \gamma_1, \eta_2 + \gamma_2)$ and $T(x) := (T_1(x_1), T_2(x_2))$.

C.2.2 Further power law odds examples

Proposition 4.10 (Power function family, Beta distribution). *Suppose that $p(x|A = a)$ is a Beta distribution*

$$p(x|A = a; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}.$$

Then, the associated odds model

$$O_{a'}(x; \gamma) := \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 x^{\gamma_1} (1-x)^{\gamma_2}, \quad \gamma_0 := \frac{P(A = a')}{P(A = a)} \cdot \frac{B(\alpha + \gamma_1, \beta + \gamma_2)}{B(\alpha, \beta)}$$

holds if and only if

$$p(x|A = a'; \alpha', \beta) = \frac{x^{\alpha'-1}(1-x)^{\beta-1}}{B(\alpha', \beta)},$$

where $\alpha' = \alpha + \gamma_1 \in \mathbb{R}^+$ and $\beta' = \beta + \gamma_2 \in \mathbb{R}^+$.

Proposition 4.11 (Power function family, Dirichlet distribution). *Suppose that X is a random variable belonging to the $K - 1$ simplex such that*

$$p(x|A = a; \alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^K x_j^{\alpha_j-1}$$

is a Dirichlet distribution. Then, the associated odds model

$$O_{a'}(x; \gamma) := \frac{P(A = a'|x)}{P(A = a|x)} = \gamma_0 \prod_{j=1}^K x_j^{\gamma_j}, \quad \gamma_0 := \frac{P(A = a')}{P(A = a)} \cdot \frac{B(\alpha + \gamma)}{B(\alpha)}$$

holds if and only if

$$p(x|A = a'; \alpha') = \frac{1}{B(\alpha')} \prod_{j=1}^K x_j^{\alpha'_j-1},$$

where $\alpha'_j = \alpha_j + \gamma_j \in \mathbb{R}^+$ for $j > 0$.

C.3 Additional Comments on Tree Graphs

C.3.1 Congruency

We now discuss the concept of congruency. When conducting real data analysis, we only need to construct a tree graph using patterns that are observed in the data and can ignore any pattern that is not observed in the real data. We formalize this notation in the following section, utilizing the fact that patterns that are not observed in the real data have Lebesgue measure 0.

Definition 4.8 (Congruency). Two tree graphs G_1 and G_2 are said to be *congruent* with respect to the observed data if G_1 and G_2 are identical after removing any patterns that do not appear in the observed data. We omit the phrase “with respect to the observed data” when it is clear from context.

In essence, two tree graphs G_1 and G_2 being congruent with respect to the observed data means that any statistical functional of the full-data distribution is the same regardless if the assumption G_1 or G_2 was made. We now introduce the idea of a representor graph to represent a set of congruent tree graphs.

Definition 4.9 (Representor). A representor of a set of tree graphs $\mathcal{T}$ is the tree graph that comprises only the patterns observed in the data and is congruent to every tree graph in $\mathcal{T}$.

The graph comprising of only the patterns that are observed in the real data implies that the graph does not contain superfluous information that is ignored by the observed data. Moreover, if the graph is congruent to every graph in $\mathcal{T}$, it is the minimal graph that represents all of the patterns.

Example 16. Suppose the observed data has $d = 3$ variables with only the following patterns: 111, 101, 011, and 001. Consider the following tree graphs in Figure 4.12, labeled from left to right as G_1 , G_2 , and G_3 . We see that G_1 and G_2 are congruent with respect to the observed

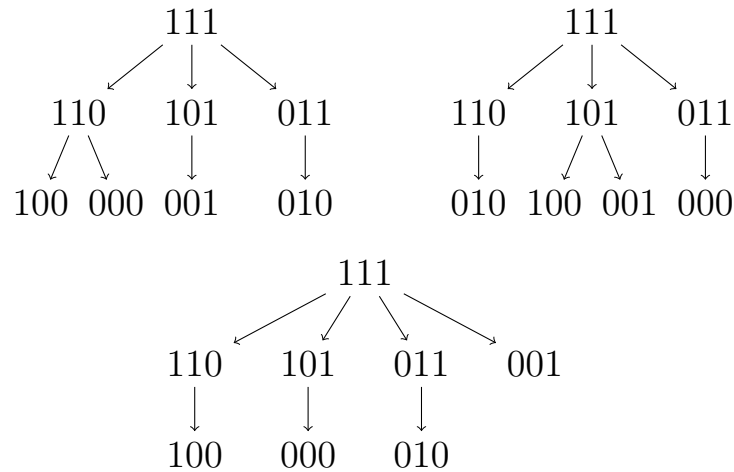


Figure 4.12: We provide three examples of tree graphs for $d = 3$ variables.

data, but G_3 is not congruent to G_1 or G_2 . Moreover, the respective representor graphs of $\{G_1, G_2\}$ and $\{G_3\}$ are found in Figure 4.13.

Remark 10 (Selecting a threshold). In practice, choosing the patterns that should appear in the representor can be done in various ways. A straightforward way would be to only consider the patterns that are observed in the data. More generally, one can consider thresholding based on the number of observations and only include missing patterns with a number of observations at least the threshold.

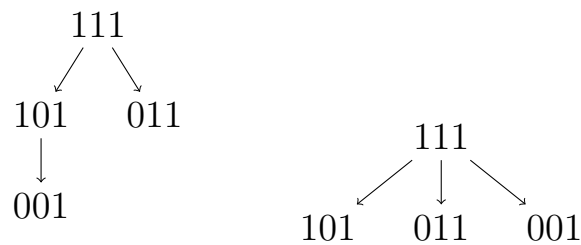


Figure 4.13: We provide examples of representor graphs, corresponding to the graphs in Figure 4.12.

The threshold can be selected to be any constant $C > 0$. For example, if we only keep patterns such that the number of observations is at least $C = 1$, this corresponds to selecting the patterns that are observed in the data. On the other hand, we may choose to keep patterns such that the number of observations is at least $C = 100$, which implies that we are seeking a sufficiently large enough effective sample size. One advantage to choosing $C > 1$ is to avoid potential problems with model fitting.

C.3.2 Invariance for a specific statistical functional

One may hypothesize that for a given parameter of interest, certain tree graphs may lead to the same identification formula for that parameter. More formally, we now consider invariance for a specific statistical functional.

Theorem 4.3 (Sufficient conditions for marginal distribution invariance). *Let $\mathcal{B} = \{j_1, j_2, \dots, j_B\} \subseteq [d]$ be a set of indices that correspond to several variables. Define the set $\mathcal{R}_{\mathcal{B}} := \{s : s_j = 0, j \in \mathcal{B}\}$ to contain exactly the patterns that have a 0 in each index in $\mathcal{B}$. Suppose that there are two distinct tree graphs G_1 and G_2 such that for each $r \in \mathcal{R}_{\mathcal{B}}$, r has the same ancestors in G_1 and G_2 . Then, any statistical functional of the distribution $p(x_{j_1}, x_{j_2}, \dots, x_{j_B})$ is the same regardless of the graph G_1 or G_2 .*

A consequence of this theorem is that all tree graphs for $d = 2$ induce unique marginal distributions.

Next, we now describe a way to combine tree graphs into a single pattern graph assumption using a merge operation. Graphically, the merge operation is very simple. Consider an example in Figure 4.14. We formalize the merge operation in mathematical notation as follows.

Definition 4.10 (Merge operation). Consider the pattern graph $G := G_1 \cup G_2$, where $\cup$ between two pattern graphs denotes the merge operation. The resulting graph $G := G_1 \cup G_2$

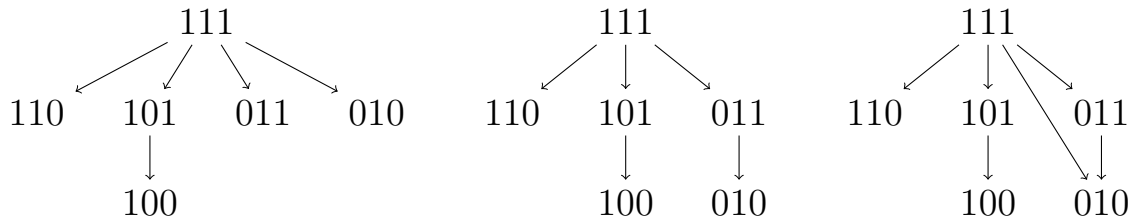


Figure 4.14: Merge property. Label the graphs from left to right as G_1 , G_2 , and G_3 .

is constructed such that $\text{PA}_G(r) = \text{PA}_{G_1}(r) \cup \text{PA}_{G_2}(r)$ for every pattern r . It satisfies the following properties:

- G is a pattern graph.
- G has at least the same number of edges as G_1 and G_2 .

Proposition 4.12 (Tree graphs generate pattern graphs). *The closure of the set of tree graphs under the merge operation is the set of pattern graphs. Moreover, the set of all tree graphs is the smallest set that generates pattern graphs.*

BIBLIOGRAPHY

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, B. F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.
- Alashwal, H., El Halaby, M., Crouse, J., Abdalla, A., and Moustafa, A. (2019). The application of unsupervised clustering methods to alzheimer’s disease. *Frontiers in Computational Neuroscience*, 13:31.
- Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099 – 3132.
- Anderson, M. J., de Valpine, P., Punnett, A., and Miller, A. E. (2019). A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution*, 9(6):3276–3294.
- Beesley, L. and Taylor, J. (2020). A stacked approach for chained equations multiple imputation incorporating the substantive model. *Biometrics*.
- Beesley, L. J., Bartlett, J. W., Wolf, G. T., and Taylor, J. M. G. (2016). Multiple imputation of missing covariates for the cox proportional hazards cure model. *Statistics in Medicine*, 35(26):4701–4717.
- Berg, E. and Yu, C. (2021). Estimation for nonignorable missing response or covariate us-

- ing semi-parametric quantile regression imputation and a parametric response probability model. *Statistica Sinica*.
- Bishop, C. M. and Svenskn, M. (2002). Bayesian hierarchical mixtures of experts. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 57–64, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Bock, H. H. (1994). *Information and Entropy in Cluster Analysis*, pages 115–147. Springer Netherlands, Dordrecht.
- Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Boyle, P. A., Yu, L., Wilson, R. S., Leurgans, S. E., Schneider, J. A., and Bennett, D. A. (2018). Person-specific contribution of neuropathologies to cognitive loss in old age. *Annals of Neurology*, 83(1):74–83.
- Brenowitz, W. D., Hubbard, R. A., Keene, C. D., Hawes, S. E., Longstreth Jr, W., Woltjer, R. L., and Kukull, W. A. (2017). Mixed neuropathologies and estimated rates of clinical progression in a large autopsy sample. *Alzheimer's & Dementia*, 13(6):654–662.
- Brintz, B., Fuentes, C., and Madsen, L. (2018). An asymptotic approximation to the n-mixture model for the estimation of disease prevalence. *Biometrics*, 74(4):1512–1518.
- Carreira-Perpinan, M. A. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12(1):141–152.
- Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm de-

- rived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82.
- Celeux, G. and Diebolt, J. (1987). The EM and SEM algorithms for mixtures :Statistical and numerical aspects. Technical Report RR-0641, INRIA.
- Chen, T., Zhang, N. L., Liu, T., Poon, K. M., and Wang, Y. (2012). Model-based multidimensional clustering of categorical data. *Artificial Intelligence*, 176(1):2246–2269.
- Chen, Y.-C. (2022a). Pattern graphs: A graphical approach to nonmonotone missing data. *The Annals of Statistics*, 50(1):129 – 146.
- Chen, Y.-C. (2022b). Statistical inference with local optima. *Journal of the American Statistical Association*, 0(0):1–13.
- Chen, Y.-C. and Sadinle, M. (2019). Nonparametric pattern-mixture models for inference with missing data. *arXiv preprint arXiv:1904.11085*.
- Cheng, G., Chen, Y.-C., Smith, M. A., and Zhao, Y.-Q. (2022). Handling nonmonotone missing data with available complete-case missing value assumption. *arXiv preprint arXiv:2207.02289*.
- Crowe, B. J., Lipkovich, I. A., and Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical statistics*, 9(4):269–279.
- De Meyer, G., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., De Deyn, P. P., Coart, E., Hansson, O., Minthon, L., Zetterberg, H., Blennow, K., Shaw, L., Trojanowski, J. Q., and Initiative, A. D. N. (2010). Diagnosis-independent alzheimer disease

- biomarker signature in cognitively normal elderly people. *Archives of Neurology*, 67(8):949–956.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.
- Diggle, P. J. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):49–93.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.
- Escudero, J., Zajicek, J. P., and Ifeachor, E. (2011). Early detection and characterization of alzheimer’s disease in clinical scenarios using bioprofile concepts and k-means. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6470–6473.
- Esscher, F. (1932). On the probability function in the collective theory of risk. *Scandinavian Actuarial Journal*, 1932(3):175–195.
- Franks, A. M., Airoidi, E. M., and Rubin, D. B. (2020). Nonstandard conditionally specified models for nonignorable missing data. *Proceedings of the National Academy of Sciences*, 117(32):19045–19053.
- Fruhwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26:78–89.
- Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., and Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical alzheimer’s

- disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):103–111.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.
- Gormley, I. C. and Frühwirth-Schnatter, S. (2018). Handbook of mixture analysis. *arXiv: Methodology*.
- Gormley, I. C. and Murphy, T. B. (2008). Exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027.
- Gu, Y. and Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*, 48(4):2082 – 2107.
- Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31(2):542–548.
- Haines, L. M. (2016). Maximum likelihood estimation for n-mixture models. *Biometrics*, 72(4):1235–1245.
- Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.
- Huerta, G., Jiang, W., and Tanner, M. A. (2003). Time series modeling via hierarchical mixtures. *Statistica Sinica*, 13(4):1097–1118.
- Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–347.

- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.
- Jordan, M. and Jacobs, R. (1993). Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pages 1339–1344 vol.2.
- Kang, H. (2013). The prevention and handling of missing data. *Korean Journal of Anesthesiology*, 64(5):402–406.
- Katzman, R., Aronson, M., Fuld, P., Kawas, C., Brown, T., Morgenstern, H., Frishman, W., Gidez, L., Eder, H., and Ooi, W. L. (1989). Development of dementing illnesses in an 80-year-old volunteer cohort. *Ann Neurol*, 25(4):317–324.
- Kéry, M., Royle, J. A., and Schmid, H. (2005). Modeling avian abundance from replicated counts using binomial mixture models. *Ecological Applications*, 15(4):1450–1461.
- Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106(493):157–165.
- Kitching, D. (2015). Depression in dementia. *Australian Prescriber*, 38:209–211.
- Kouw, W. M. and Loog, M. (2021). A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785.
- Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293.

- Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18:95–138.
- Lee, J. W. and Harel, O. (2022). Incomplete clustering analysis via multiple imputation. *Journal of Applied Statistics*, 50(9):1962–1979.
- Levine, R. A. and Casella, G. (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439.
- Linzer, D. A. and Lewis, J. B. (2011). polca: An r package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10):1–29.
- Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, Hoboken, New Jersey, 2nd edition.
- Malinsky, D., Shpitser, I., and Tchetgen, E. J. T. (2020). Semiparametric inference for non-monotone missing-not-at-random data: the no self-censoring model.
- Maslen, B., Popovic, G., Lim, M., Marzinelli, E., and Warton, D. (2023). How many sites? methods to assist design decisions when collecting multivariate data in ecology. *Methods in Ecology and Evolution*, 14(6):1564–1573.
- Maydeu-Olivares, A. and Liu, Y. (2015). Item diagnostics in multivariate discrete data. *Psychological Methods*, 20(2):276–292.
- Mayer, I., Josse, J., and Group, T. (2021). Transporting treatment effects with incomplete attributes.

- Mayer, I., Sverdrup, E., Gauss, T., Moyer, J.-D., Wager, S., and Josse, J. (2020). Doubly robust treatment effect estimation with missing attributes. *The Annals of Applied Statistics*, 14(3):1409 – 1431.
- McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2nd Edition. CRC Press, 2 edition.
- McParland, D. and Gormley, I. C. (2013). Clustering ordinal data via latent variable models. In Lausen, B., Van den Poel, D., and Ultsch, A., editors, *Algorithms from and for Nature and Life*, pages 127–135, Cham. Springer International Publishing.
- Meng, X. and D’Arcy, C. (2012). Education and dementia in the context of the cognitive reserve hypothesis: A systematic review with meta-analyses and qualitative analyses. *PLoS ONE*, 7.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558.
- Miao, W. and Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482.
- Mjørud, M., Røsvik, J., Rokstad, A. M., Kirkevold, M., and Engedal, K. (2014). Variables associated with change in quality of life among persons with dementia in nursing homes: A 10 months follow-up study. *PLoS One*, 9(12):e115248.
- Mohan, K. and Pearl, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems*, pages 1520–1528.

- Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.
- Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In *Advances in neural information processing systems*, pages 1277–1285.
- Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2):153–161.
- Nabi, R., Bhattacharya, R., and Shpitser, I. (2020). Full law identification in graphical models of missing data: Completeness results. *arXiv preprint arXiv:2004.04872*.
- O’Hagan, A., Murphy, T. B., Scrucca, L., and Gormley, I. C. (2019). Investigation of parameter uncertainty in clustering using a gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Computational Statistics*, 34(4):1779–1813.
- Ouyang, J. and Xu, G. (2022). Identifiability of latent class models with covariates. *Psychometrika*, 87(4):1343–1360.
- Paik, M. (1997). Multiple imputation for the cox proportional hazards model with missing covariates. *Lifetime data analysis*, 3:289–98.
- Phung, T., Reese, K., Shpitser, I., and Bhattacharya, R. (2025). Recursive equations for imputation of missing not at random data with sparse pattern support. *arXiv preprint arXiv:2507.16107*.
- Qi, L., Wang, C., and Prentice, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 100(472):1250–1263.

- Qiu, Y., Jacobs, D. M., Messer, K., Salmon, D. P., and Feldman, H. H. (2019). Cognitive heterogeneity in probable alzheimer disease: Clinical and neuropathologic features. *Neurology*, 93(8):e778–e790.
- Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (mimp) approach. *Statistics in medicine*, 28(9):1402–1414.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. *Copenhagen, Danish Institute for Educational Research*.
- Robins, J. M. and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in medicine*, 16(1):39–56.
- Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.
- Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87(1):113–124.
- Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

- Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3):271–282.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.
- Sadinle, M. and Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.
- Serafini, A., Murphy, T. B., and Scrucca, L. (2020). Handling missing data in model-based clustering.
- Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187.
- Shpitser, I. (2016). Consistent estimation of functions of data missing non-monotonically and not at random. In *Advances in Neural Information Processing Systems*, pages 3144–3152.
- Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

- Sportisse, A., Marbac, M., Biernacki, C., Boyer, C., Celeux, G., Josse, J., and Laporte, F. (2023). Model-based clustering with missing not at random data.
- Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.
- Stern, Y. (2012). Cognitive reserve in ageing and alzheimer’s disease. *Lancet Neurology*, 11(11):1006–1012.
- Suen, D. and Chen, Y.-C. (2021). Multistage estimators for missing covariates and incomplete outcomes.
- Suen, D. and Chen, Y.-C. (2023). Modeling missing at random neuropsychological test scores using a mixture of binomial product experts.
- Sugiyama, M. and Kawanabe, M. (2012). *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press.
- Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press.
- Tang, X. and Qu, A. (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, 25(4):1117–1137.
- Tchetgen Tchetgen, E. J., Wang, L., and Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069–2088.
- Teicher, H. (1961). Identifiability of Mixtures. *The Annals of Mathematical Statistics*, 32(1):244 – 248.

- Thomas, H., Lohaus, A., and Brainerd, C. J. (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development*, 58(9):i–190.
- Thow, M. E., Summers, M. J., Saunders, N. L., Summers, J. J., Ritchie, K., and Vickers, J. C. (2018). Further education improves cognitive reserve and triggers improvement in selective cognitive functions in older adults: The tasmanian healthy brain project. *Alzheimer's & dementia : diagnosis, assessment & disease monitoring*, 10(1):22–30.
- Tian, J. (2015). Missing at random in graphical models. In *Artificial Intelligence and Statistics*, pages 977–985.
- Tosto, G., Monsell, S. E., Hawes, S. E., Bruno, G., and Mayeux, R. (2016). Progression of extrapyramidal signs in alzheimer's disease: Clinical and neuropathological correlates. *Journal of Alzheimer's Disease*, 49(4):1085–1093.
- Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.
- Tukey, J. W. (1986). *Discussion 4: Mixture Modeling Versus Selection Modeling with Non-ignorable Nonresponse*, pages 143–148. Springer New York, New York, NY.
- van Buuren, S. and Groothuis-Oudshoorn, K. (2011a). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.
- van Buuren, S. and Groothuis-Oudshoorn, K. (2011b). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.
- van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence*. Springer.

- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4):450–469.
- Wang, C. Y. and Chen, H. Y. (2001). Augmented inverse probability weighted estimator for cox missing covariate regression. *Biometrics*, 57(2):414–419.
- Wang, H. and Kim, J. K. (2021). Propensity score estimation using density ratio model under item nonresponse.
- Wang, N. and Robins, J. M. (1998a). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4):935–948.
- Wang, N. and Robins, J. M. (1998b). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4):935–948.
- Waterhouse, S., MacKay, D., and Robinson, A. (1995). Bayesian methods for mixtures of experts. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press.
- White, A. and Murphy, T. B. (2014). Bayeslca: An r package for bayesian latent class analysis. *Journal of Statistical Software*, 61(13):1–28.
- White, I. R. and Royston, P. (2009). Imputing missing covariate values for the cox model. *Statistics in Medicine*, 28(15):1982–1998.
- Yesavage, J. A., Taylor, J. L., Friedman, L., Rosenberg, P. B., Lazzeroni, L. C., Leoutsakos, J. S., Kinoshita, L. M., Perlow, M. J., Munro, C. A., Devanand, D. P., Drye, L. T., Mintzer, J. E., Pollock, B. G., Porsteinsson, A. P., Schneider, L. S., Shade, D. M., Weintraub, D., Lyketsos, C. G., Noda, A., and Group, C. R. (2016). Principal components analysis of agitation outcomes in alzheimer’s disease. *Journal of Psychiatric Research*, 79:4–7.

- Yuksel, S., Wilson, J., and Gader, P. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1177–1193.
- Zamanian, A., Ahmidi, N., and Drton, M. (2023). Assessable and interpretable sensitivity analysis in the pattern graph framework for nonignorable missingness mechanisms. *Statistics in Medicine*, 42(29):5419–5450.
- Zhang, M., Katzman, R., Salmon, D., Jin, H., Cai, G., Wang, Z., Qu, G., Grant, I., Yu, E., Levy, P., et al. (1990). The prevalence of dementia and alzheimer’s disease in shanghai, china: impact of age, gender, and education. *Annals of Neurology*, 27(4):428–437.
- Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723.
- Zhao, P., Tang, N., Qu, A., and Jiang, D. (2017). Semiparametric estimating equations inference with nonignorable missing data. *Statistica Sinica*, pages 89–113.

VITA

Daniel Suen received dual B.S. degrees in Statistics and Electrical Engineering from the University of Florida in 2019. He earned his M.S. and Ph.D. in Statistics from the University of Washington in 2022 and 2025, respectively.