Large N, Small T, Multiple P

A Causal Matrix Completion Method for CRM Panel Data

Zhongming Jiang[1,2]

Supervised by

Prof. Fred Feinberg[1,3]

Prof. Longxiu Tian[4]

May 18, 2024

In partial fulfillment of the requirements for the degree of

Bachelor of Science in Statistics (Honors)

[1]Department of Statistics, University of Michigan
[2]Department of Mathematics, University of Michigan
[3]Stephen M. Ross School of Business, University of Michigan
[4]Kenan-Flagler Business School, University of North Carolina

Abstract

The prototypical customer relationship management (CRM) panel structure is composed of many customers (large N), with short histories (small T), and multiple outcome metrics (multiple P). Our paper aims to tackle the challenges of causal inference that firms face in such CRM settings, which are additionally characterized by unobserved heterogeneity, time dynamics, and staggered adoption. Despite the success of synthetic control methods (SCM) in contemporary marketing applications, extant variants typically necessitate "small N, large T" data regimes to be performant – e.g., a handful of firm- or jurisdiction-level donor units, each with long time series.

To extend to the "large N, small T, multiple P" setting, we bridge SCM to the broader causal matrix completion (MC) paradigm and leverage the "multiple P" ubiquitous to contemporary CRM: the presence of multiple outcomes enables a shared matrix *singular value decomposition* (cf. SCM's *factorization*), which helps jointly identify individual-level latent factors to establish conditional ignorability, compensating for overall short time series at the customer level. We employ a Bayesian causal inference approach, specifying a joint posterior of the nonrandom missingness of potential outcomes, together with the likelihood of the observed outcomes. We introduce two distinct variants of Bayesian causal MC models, each estimated independently through the implementation of the Gibbs sampling (independent multiple P's) and the Hamiltonian Monte Carlo (concurrent multiple P's) -based data augmentation procedure. We empirically illustrate our approach through a comprehensive customer-level database of gift card purchases and redemptions from a U.S. hospitality startup. We compare the effectiveness with extant SCM under the German reunification empirical study and devise a generalized framework for marketing and statistics researchers applicable to a wide range of CRM panel structures.

Keywords: Customer-Base Analysis, Bayesian Causal Inference, Counterfactual Estimation, Synthetic Control Method, Panel Data, Latent Factor Model, Matrix Completion

Supplementary Materials: [Slide] and [RPubs]

Acknowledgements

In this honors thesis, I present some key findings from my ongoing research under my thesis advisor, Prof. Fred Feinberg, from January 2023 to the present. My research has also been supported by my thesis co-advisor, Prof. Longxiu Tian at UNC Kenan-Flagler, where I have worked as a research assistant from May 2023 to August 2023. I greatly thank Prof. Feinberg and Prof. Tian for guiding my research on Bayesian causal inference and probabilistic machine learning. In particular, I thank Prof. Tian for his idea of implementing the Bayesian causal MC model with concurrent multiple P's.

This honors thesis is a selection of my research outcomes that best represents my understanding and training in statistics and mathematics, with an external interest in its application in marketing and quantitative social science. I am grateful to the company for providing the CRM dataset used for performing longitudinal data analysis and statistical modeling, and to the audience who provided feedback on my presentation at the Complex Systems Advanced Academic Workshop in October 2023. My appreciation also goes to the Center for the Study of Complex Systems for nominating my work for the first-place prize of the Rick Riolo Memorial Fund Undergraduate Research. In addition, I want to thank my parents and sister a lot for their constant financial support and encouragement during my college time, especially when the pandemic made things hard. I also thank my thesis advisors, Prof. Feinberg and Prof. Tian, for helping me identify my research interests in quantitative marketing and setting a career goal in academia, and my statistics advisor, Ms. Gina Cornacchia, for her help in planning my statistics classes and introducing me to research opportunities.

Please note that while I adopt *we* as the first-person perspective, the entire thesis is solely authored by me. I thank my advisors for their suggestions and edits from weekly meetings since January 2023. No forms of generative AI tools have been used to complete any sections of this thesis. I take full responsibility for any errors that may remain in the thesis.

Contents

List of Figures

List of Tables

1

Introduction

In the past decade, business-to-consumer (B2C) firms across the globe have been at the forefront of embracing digital marketing, in an effort to reach a broader range of customers and audiences more effectively and quickly. This digitization, in turn, has led to an unprecedented drive by B2C firms towards more atomic and real-time customer relationship management (CRM), a function typically found within a firm's marketing arm that collectively encompasses the strategies and technologies of audience engagement, lifecycle marketing, and customer lifetime value (CLV). Of fundamental importance to accurately calculating CLV is accurately modeling and predicting customers' retention rate — a forward-looking expectation on the likelihood of an individual remaining as a customer over a given time period. In doing so, firms can proactively target customers who are most vulnerable to quitting, personalize marketing communications to upsell or cross-sell, and even use these predictions to segment customers who are of low- or negative- value to the firm. Conventional methods make use of metrics such as recency and demographics to address the *cold start* problem in CRM, which arises when firms are faced with the challenge of making inferences about customers based on limited data at the outset of the relationship. However, companies often encounter a situation where they observe a newly acquired customer on only one occasion (Padilla and Ascarza 2021). This challenge severely hinders their ability to track the behavior and impression of customers throughout their subsequent purchases.

So how can we differentiate between customers who have terminated their relationship

with the firm from those who are merely experiencing an extended pause in their purchasing activity (Fader and Hardie 2009)? In contractual settings (e.g., subscription or membership), we observe the time period at which customers churn (i.e., end their formal relationship with the firm), and thus the CLV models can be straightforward. On the other hand, in non-contractual settings, where firms do not explicitly observe customer churning, it presents a significant challenge for firms to tell if a customer — in particular, a newly acquired one — is going to be retained or churn in the next period.

A common solution in non-subscription settings is to construct the probabilistic models for CLV (Netzer et al. 2008; Fader and Hardie 2009; Fader et al. 2010) that often rely on three latent parameters: *lifetime* (how long the customer relationship lasts), *purchase rate* (how often purchases occur), and *monetary value* (the value of future transactions). These three measures, also known as recency (R), frequency (F), and monetary (M) value in RFM analysis, are unobserved in non-subscription settings, yet crucial for probabilistic models like "Pareto/Negative Binomial Distribution (NBD)" (Schmittlein et al. 1987), which seeks to predict future customer transactions and overall lifetime value. However, the rigid assumptions of such a model[1] have proven to be less applicable to a broader range of non-subscription customer-level observational CRM data (Fader et al. 2010).

In particular, such observational CRM data often contain many individual customers (large N) with jagged arrayed[2] time-series[3] cross-sectional[4] features. Known as panel data or longitudinal data, such CRM data also accompany many dimensions (multiple P) regarding transaction types, such as purchases or redemptions. With RFM analysis aforementioned, we can measure at least six dimensions for such CRM panel data (i.e., a combination of purchases or redemptions with recency, frequency, or monetary value). One may also consider incorporating additional important outcome metrics, such as the clumpiness (C), which can be extended by a metric-based approach in RFM framework (Zhang et al. 2015).

Unfortunately, existing marketing literature lacks explicit models that can accommodate the common data challenges encountered in CRM. Consequently, this study aims to address this gap by proposing a model suitable for CRM panel data often characterized by the challenge of large N, small T, and multiple P.

In Table 1.1, we illustrate a simulated individual-level transaction history for N customers and T periods, where $N \gg T$. If we observe each cross-sectional customer $n = 1, 2, \ldots, N$ at certain discrete-time periods $t = 1, 2, \ldots, 8$ (here, 8 indicates the last period we could

[1] A Poisson distribution assumes that transactions can occur at any time for customer purchasing while active.

[2] Customer-level transaction with various starting and ending period

[3] Often across small T due to the nature that only a tiny portion of *loyal* customers have frequent transactions

[4] Often across large N, a common pattern in CRM data

observe), then we mark a "✓" for that block. From Table 1.1, we can see that individual customers are not necessarily always observed with a transaction (either purchase or redemption). From a customer segmentation perspective (Ascarza et al. 2018), customer $n = 1$ is considered as a *loyal* customer who is *engaged* with the firm, so such category is not our primary target to retain as many customers as we can. Customers $n = 2$ and $n = 3$ are called *silently gone* customers, since they become inactive early on. Customers $n = 4$ and $n = 5$, in contrast, are those newly acquired customers. Notably, customers $n = 3$ and $n = 5$ are known as one-time purchasers (or one-time redeemers), characterized by a single purchase (or redemption). Lastly, $n = N$ signifies customers with sporadic transaction patterns. This category shares characteristics with *loyal* customers in terms of their time span (roughly the same T), but has a more complicated underlying mechanism. We will take an in-depth look at a real-world CRM panel data application in Section 5.2.

Table 1.1: Customer-Level Transaction History

	Small T (Periods)							
Large N (Customers)	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$
$n = 1$	✓		✓	✓	✓		✓	✓
$n = 2$	✓	✓		✓				
$n = 3$		✓						
$n = 4$					✓		✓	✓
$n = 5$								✓
⋮				⋮				
$n = N$	✓				✓			✓

Recognizing such a panel data challenge and noting the growing popularity of quasi-experiments that promote causality research in marketing (Goldfarb et al. 2022) beyond models in customer-base analysis, we propose a Bayesian causal matrix completion (MC) model (to be introduced in Section 4.2.3) that explicitly works with customer-level panel data featuring jagged arrays, large N, small T, and multiple P.

The remainder of the paper is organized as follows: In Chapter 2, we conduct a detailed literature review on quasi-experiments, specifying the considerations and motivations behind the selection of our model. In Chapter 3, we outline the mathematical derivations and the underlying assumptions adopted in our model. In Chapter 4, we introduce relevant models with a motivating example, specifically the German reunification. In Chapter 5, we apply our Bayesian causal MC model to real-world CRM panel data. In Chapter 6, we discuss our model's contributions in terms of methodological advances compared to a baseline model and suggest future work for further refinement.

2

Literature Review

With observational data being a prototypical marketing data setting in contrast to experimental data, quasi-experimental designs have been intensively applied in marketing causality research (Goldfarb et al. 2022). Their goal is to estimate the counterfactual of an object had the treatment not occurred, thereby enabling us to overcome the fundamental problem of causal inference[1] (Holland 1986). Several quasi-experimental designs appear promising for CRM causality research, including propensity score matching (PSM; Rosenbaum and Rubin 1983), difference-in-differences (DiD; Ashenfelter and Card 1985), and the synthetic control method (SCM; Abadie and Gardeazabal 2003; Abadie et al. 2010).

In observational studies, the assignment of treatment is often not random, so these quasi-experimental methods have various assumptions and/or specific data characteristics in order to estimate the causal effects out of unconfoundedness (Kim et al. 2020). In summary, PSM estimates the probability (propensity score) of a unit receiving the treatment, given observed characteristics (Rosenbaum and Rubin 1983). Then, like other matching methods (Abadie and Imbens 2006; Doudchenko and Imbens 2016), PSM matches the propensity score for control and treated units. However, we notice that panel data often have time-varying confounders. Traditional PSM (Rosenbaum and Rubin 1983) does not account for changes over time in the covariates, unfortunately. In addition, PSM assumes that the assignment of units to treatment and control groups, based on the propensity

[1] That is, we can compare the counterfactual outcome with the observed outcome for the same observational unit and, therefore, derive the causal effect.

score, is as good as random (conditional independence assumption). According to Kim et al. (2020), the estimated treatment effects will be biased if there are unobserved characteristics that affect assignment to treatment and are not orthogonal to the outcome.

DiD, on the other hand, compares the changes in outcomes over time between a treatment group and a control group. PSM can be used in conjunction with DiD to ensure that the treatment and control groups are comparable on baseline covariates. The conventional DiD method (Card and Krueger 1994) requires that the trends in outcomes for both groups would have been parallel in the absence of the treatment. However, the selection of comparison units to reduce biases in observational studies is ambiguous (Abadie et al. 2010).

The generalization of the DiD methods, SCM, has been developed and intensively used in comparative case studies in political science (Abadie and Gardeazabal 2003; Abadie et al. 2010), with an explicit data-driven control unit selection procedure (Kim et al. 2020). With a single treated unit, SCM creates a weighted convex combination of untreated (control) units to construct a synthetic counterfactual (Abadie et al. 2010). Even though we could relax the constraint that the standard SCM is not limited to only a single treated unit, it still relies on an assumption that all treated units receive the treatment at a single point in time, known as the static treatment assignment or static adoption assumption (Doudchenko and Imbens 2016; Ben-Michael et al. 2021).

SCM plays a significant role in recent marketing literature, notably in examining the causal effect of a soda tax on firms' and consumers' behaviors in Berkeley, CA (Rojas and Wang 2020; Kim et al. 2020), and in assessing the impact of offline TV advertising on various dimensions of online chatter (Tirunillai and Tellis 2017). In Figure 2.1, we outline a flowchart[2] for selecting different quasi-experimental estimators, based on various data characteristics and assumptions.

2.1 Limitations in Synthetic Control Method

SCM, despite being "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens 2017), has been shown to be very unlikely to hold in real-world applications due to its restrictive weighting constraints (Doudchenko and Imbens 2016). Furthermore, the limitation to convex combinations (non-negative weights that sum to one, without an intercept) biases the SCM estimator (Ferman and Pinto 2019; Carvalho et al. 2018).

In terms of statistical inference, SCM is also untenable (Kim et al. 2020). Abadie et

[2]The essential structure is inspired by a mind map of the taxonomy of causal inference, first introduced by Prof. Kathleen Li at a conference in May 2023.

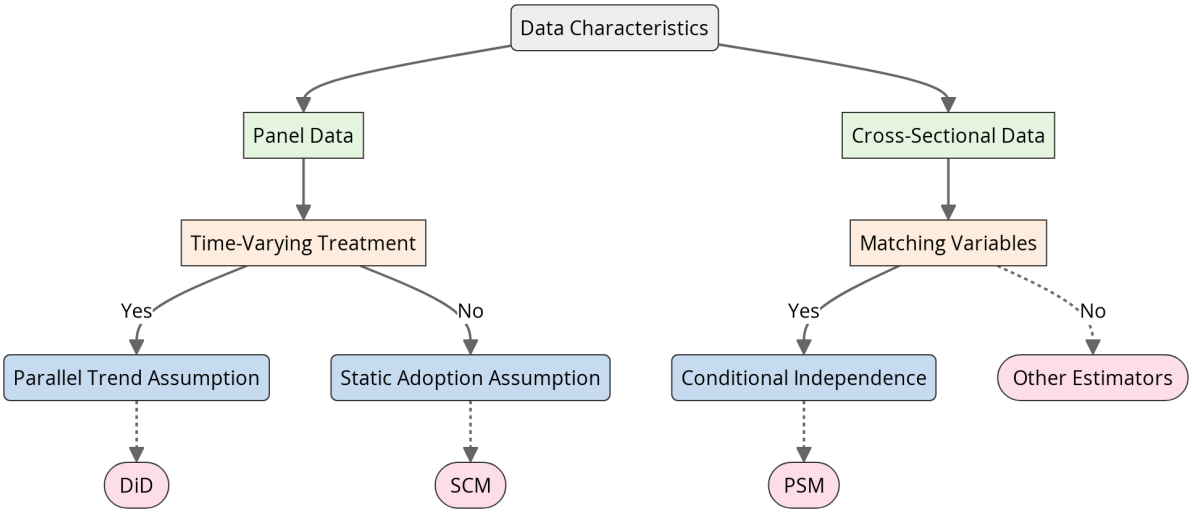

Figure 2.1: Choice of Quasi-Experimental Designs

al. (2010, 2015) adopt the placebo test, a form of permutation test, whose validity is challenged by Hahn and Shi (2017). They argue that the symmetry assumption is violated and that the current form of the permutation test cannot serve as a proper tool for inference with SCM.

In addition to its inherent limitations, SCM requires extensions to be applicable to panel data. In Section 2.2, we will discuss existing extensions of SCM and illustrate how our model integrates into the broader context.

2.2 Extensions on Synthetic Control Method

There are, broadly speaking, three categories of extensions to SCM (Pang et al. 2021). The first category involves extending standard matching or re-weighting methods to panel data settings. This includes best subset methods (Hsiao et al. 2011), which combine panel data methods that use observed data to construct counterfactuals; regularized weights (Doudchenko and Imbens 2016), which introduce a more flexible SCM estimator by allowing negative weights and additive differences; and panel matching (Imai et al. 2021), an extension of matching methods that incorporates treatment history matching and covariate balance into time-series cross-sectional data.

The second category is hybrid methods, also known as doubly robust methods. Some previous research on doubly robust estimators includes synthetic DiD (Arkhangelsky et al. 2021; a computational implementation of the synthetic DiD estimator for estimating treatment effects in various contexts with repeated observations over time), augmented SCM (Ben-Michael et al. 2021; an extension of SCM to settings with imperfect pre-treatment fit, using an outcome model to estimate and correct bias), and augmented DiD

(Li and Van den Bulte 2022; an estimator that extends over SCM by better handling heterogeneity between treatment and control units for estimating the average treatment effect on the treated, ATT).

The third category is factor models, also known as the generalized SCM (Xu 2017). Bai (2009) first implements latent factor models (LFMs) that consider large N and large T panel data models with unobservable multiple interactive fixed effects (IFE). Pang (2010; 2014) proposes nonlinear IFE models with exogenous covariates in a Bayesian hierarchical framework. Gobillon and Magnac (2016) demonstrate that IFE models outperform SCM in DiD settings when the factor loadings of the treatment and control groups do not share common support. Xu (2017) then proposes a generalized SCM that unifies SCM with linear fixed effects models, under the framework of which DiD is a special case. More recently, Athey et al. (2021) propose a class of MC estimators that summarizes the IFE extension on SCM as a subset of MC methods.

The aforementioned extensions have somewhat relaxed the innate weighting constraints, accommodated multiple treated units, and enhanced the predictive performance and robustness of counterfactual estimation in SCM (Pang et al. 2021). However, these existing extensions still encounter challenges not only in inference but also in prediction. As previously mentioned, the interpretability of the SCM placebo test as a permutation test is compromised due to non-random treatment assignment (Hahn and Shi 2017). Additionally, Frequentist inferential methods necessitate a repeated sampling interpretation, such as a bootstrapping procedure, for quantifying uncertainties of a LFM (Xu 2017). Beyond inferential limitations, the rigid parametric assumptions of existing models restrict the full utilization of available panel data sources[3] for counterfactual predictions (Beck and Katz 2007; Pang 2010, 2014).

2.3 Bayesian Causal Inference

Given these existing challenges, we recognize that the Bayesian causal inference framework (Li et al. 2023) presents a viable alternative. First, the Bayesian approach comprehensively captures uncertainties from the data generation process (DGP), parameter estimation, and model selection (Pang et al. 2021). Second, Bayesian hierarchical modeling accommodates data heterogeneity and dynamics, enabling flexible functional forms and the use of shrinkage priors for model feature selection (Gelman 2005). Lastly, within the Bayesian causal inference framework, the counterfactual in SCM is treated as a missing data problem (Rubin 1976). This approach relies on the posterior predictive distribution of the treated counterfactuals to draw inferences about the treatment effects on the

[3]For example, time-series relationships among units based on their outcome trajectories, cross-sectional relationships among units based on their observed characteristics, and temporal relationships within units between their known past and unknown future.

treated, considering such missingness under the missing not at random (MNAR) framework since the assignment mechanism is allowed to correlate with unobserved potential outcomes (Pang et al. 2021).

Several pieces of literature have adopted the Bayesian approach as an extension to SCM. For example, Kim et al. (2020) propose two fully Bayesian SCM models with horseshoe and spike-and-slab priors that are designed for a single treated unit. Their models assume the availability of a sufficiently large number of control units to form a synthetic control unit. Pinkney (2021) offers an improved and extended Bayesian SCM that builds on the LFM with IFE, essentially providing a Bayesian perspective to Xu (2017). Pang et al. (2021) introduce the dynamic multi-level LFM and develop an estimation strategy using Markov chain Monte Carlo (MCMC). More recently, Martinez and Vives-i-Bastida (2023) propose the Bayesian SCM as an alternative method to perform inference for the family of SCM. They derive a Bernstein-von Mises (BvM) style result, outlining conditions under which the Bayesian SCM estimator and the maximum likelihood estimator (MLE) converge in the total variation sense.

This study, therefore, aims to continue the exploration of Bayesian SCM. In particular, we adopt and adapt the framework presented by Pang et al. (2021) to fit panel data. We propose a Bayesian causal MC model, drawing inspiration from Athey et al. (2021), and thereby generalize the family of SCMs to include more flexible forms. In Chapter 3, we will first re-examine the block structure of our working panel data, in line with Athey et al. (2021). Then, we will introduce our causal estimands and explicitly outline all necessary assumptions. We also aim to follow and enhance the posterior predictive inference approach of Pang et al. (2021).

3

Framework

In Chapter 3, we begin by reinvestigating the block structure as first developed by Athey et al. (2021). Then, we introduce the causal estimands that this study primarily focuses on. After reviewing the assignment mechanisms, we eventually derive the posterior predictive inference that fits into our Bayesian causal inference framework. In Chapter 3, we also present some interesting observations from previous research and propose them here so that readers may further consider these theoretical results for future work.

3.1 Block Structure

Consider a longitudinal study with N cross-sectional units observed over T time periods. We index the units by $i \in \{1, 2, \ldots, N\}$ and the time periods by $t \in \{1, 2, \ldots, T\}$. Within the potential outcomes framework, each unit i at each time t is associated with two potential outcomes: $Y_{it}(0)$ under control conditions, and $Y_{it}(1)$ under treatment conditions. Recalling the fundamental problem of causal inference (Holland 1986), the observable outcome for unit i at time t is $Y_{it} = Y_{it}(w_{it})$, where w_{it} is a binary indicator of treatment exposure.

3.1.1 Treatment Matrix

The matrix $\mathbf{W}$, with elements w_{it}, represents the treatment assignments for all units across all time periods in a binary fashion, with 1 indicating treatment exposure and 0

indicating no treatment. This can be formally represented as

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & \cdots & 1 & 1 \\ 0 & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}_{N \times T},$$

where each row corresponds to a cross-sectional unit i and each column to a time period t. The pattern of the above example matrix $\mathbf{W}$ follows the staggered nature of treatment adoption. We formally define the *staggered adoption* below, following Athey and Imbens (2022).

Definition 1 (Staggered Adoption)**.** Staggered adoption is defined by assigning each unit i in a longitudinal study an adoption time a_i from the set $\mathbb{A} = \{1, 2, \ldots, T, c\}$. For $a_i \leq T$, unit i is a treated unit, receiving treatment at time a_i; for $a_i = c > T$, unit i is a control unit, never receiving treatment within the study period. The treatment status of unit i at time t is denoted by $w_{it} = \mathbb{I}(t \geq a_i)$, where $\mathbb{I}$ is the indicator function.

3.1.2 Matrix Representation and Partitioning

Following the definition of the *staggered adoption* and the treatment assignment matrix $\mathbf{W}$, we now introduce the potential outcome matrix $\mathbf{Y}$. First, we define two sets, where $\mathcal{Y}$ stands for observed entries and $\mathcal{N}$ stands for missing entries in $\mathbf{Y}$, corresponding to the treatment exposure represented in $\mathbf{W}$. We define $\mathcal{Y}$ as the set of pairs (i, t) with $i \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$, such that $w_{it} = 0$, representing the observed entries. Conversely, $\mathcal{N}$ is the set of pairs (i, t) where $w_{it} = 1$, indicating the missing entries in the outcome matrix due to treatment exposure.

The potential outcome matrix $\mathbf{Y}$ is constructed to match the dimensions of $\mathbf{W}$, with each element Y_{it} corresponding to the observed outcome for unit i at time t. Formally, this can be represented as

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1,T-1} & Y_{1T} \\ Y_{21} & Y_{22} & \cdots & Y_{2,T-1} & Y_{2T} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{N-1,1} & Y_{N-1,2} & \cdots & Y_{N-1,T-1} & Y_{N-1,T} \\ Y_{N1} & Y_{N2} & \cdots & Y_{N,T-1} & Y_{NT} \end{bmatrix}_{N \times T},$$

where Y_{it} is observed if $Y_{it} = Y_{it}(0)$ with $(i, t) \in \mathcal{Y}$, and Y_{it} is missing if $Y_{it} = Y_{it}(1)$ with

$(i,t) \in \mathcal{N}$. The matched matrix $\mathbf{Y}$ for the above example matrix $\mathbf{W}$ is given by

$$\mathbf{Y} = \begin{bmatrix} Y_{11}(0) & Y_{12}(1) & \cdots & Y_{1,T-1}(1) & Y_{1T}(1) & \text{(Early Adopter)} \\ Y_{21}(0) & Y_{22}(0) & \cdots & Y_{2,T-1}(1) & Y_{2T}(1) & \text{(Progressive Adopter)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ Y_{N-1,1}(0) & Y_{N-1,2}(0) & \cdots & Y_{N-1,T-1}(0) & Y_{N-1,T}(1) & \text{(Late Adopter)} \\ Y_{N1}(0) & Y_{N2}(0) & \cdots & Y_{N,T-1}(0) & Y_{NT}(0) & \text{(Never Adopter)} \end{bmatrix}_{N \times T},$$

where $Y_{it}(0)$ and $Y_{it}(1)$ indicate the observed and the missing portions of the panel data for $(i,t) \in \mathcal{Y}$ and $(i,t) \in \mathcal{N}$, respectively. In this example, an early adopter has a long panel of missing data. On the other hand, a never adopter has observed data across the entire time span. We define the *matrix partitioning* below to split the observed and missing parts of $\mathbf{Y}$.

Definition 2 (Matrix Partitioning)**.** We partition the indices of $\mathbf{Y}$ into two sets

1. $S_{\text{obs}} \equiv \{(i,t) | w_{it} = 0\}$, where the outcome $Y_{it}(w_{it})$ is observed,
2. $S_{\text{mis}} \equiv \{(i,t) | w_{it} = 1\}$, where the outcome $Y_{it}(w_{it})$ is missing.

The union $S = S_{\text{obs}} \cup S_{\text{mis}}$ constitutes all indices. The observed and missing parts of $\mathbf{Y}$ are denoted as $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, respectively.

3.1.3 Covariate Matrix

Then, we introduce the covariate matrix $\mathbf{X}$ to characterize the block structure of any data characteristics. Let X_{it} be a $(p+1) \times 1$ vector of exogenous covariates for unit i at time t such that

$$X_{it} = \begin{bmatrix} X_{it1} \\ X_{it2} \\ \vdots \\ X_{it,p+1} \end{bmatrix}_{(p+1)\times 1},$$

where X_{itj} is the j-th covariate of unit i at time t. The covariate matrix $\mathbf{X}$ for unit i over T time periods, X_i, is a $T \times (p+1)$ matrix given by

$$X_i = \begin{bmatrix} X_{i1}^\top \\ X_{i2}^\top \\ \vdots \\ X_{iT}^\top \end{bmatrix} = \begin{bmatrix} X_{i11} & X_{i12} & \cdots & X_{i1,p+1} \\ X_{i21} & X_{i22} & \cdots & X_{i2,p+1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{iT1} & X_{iT2} & \cdots & X_{iT,p+1} \end{bmatrix}_{T\times(p+1)},$$

where each row $X_{it}^\top$ represents the transposed covariate vector for unit i at time t.

We define the full covariate matrix $\mathbf{X}$ for a population of N units as the collection $\{X_1, X_2, \ldots, X_N\}$, where each X_i is stacked vertically to form

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix},$$

creating a block diagonal matrix where each block is a $T \times (p+1)$ covariate matrix for each unit.

This summarizes the theoretical formulation of the panel data characteristics. At the beginning of Chapter 4, we present a treatment adoption plot for German reunification that visually demonstrates the degree and shape of data missingness.

3.2 Estimands

Following the construction of the treatment and outcome matrices $\mathbf{W}$ and $\mathbf{Y}$, we now introduce key assumptions and define the causal estimands for our study. The following assumptions are crucial for the validity of causal inference in the panel study setting. Building upon the framework established by Athey and Imbens (2022), we tailor the assumptions to the specific context of this study. In doing so, we introduce two key assumptions designed to exclude the possibility of cross-sectional spillover and anticipation effects.

Assumption 1 (Homogeneous Treatment Effect Across Units)**.** For all units i, j, time periods t, and adoption dates a and a', the effect of adopting treatment at time a relative to a' on the outcome in period t is the same for all units, such that

$$Y_{it}(a) - Y_{it}(a') = Y_{jt}(a) - Y_{jt}(a').$$

This first assumption, adapted from the fourth assumption made by Athey and Imbens (2022), implies a constant treatment effect across units, negating the presence of unit-specific treatment effect variations and cross-sectional spillover effects. It is also commonly referred to as the cross-sectional stable unit treatment value assumption (SUTVA).

Assumption 2 (No Anticipation)**.** For any unit i and for all time periods before its treatment adoption $t < a_i$,

$$Y_{it}(a_i) = Y_{it}(c),$$

where $Y_{it}(c)$ represents the potential outcome when the treatment vector is all zeros (i.e., under the *pure control* condition).

The above assumption, adapted from the second assumption made by Athey and Imbens (2022) by replacing $Y_{it}(\infty)$ with $Y_{it}(c)$ for notation clarity, implies that the current untreated potential outcomes are not impacted by future treatment. The violation of this assumption may occur if units anticipate certain policies or treatments prior to their implementation. After introducing these two assumptions, which empirical researchers often rely on without explicit acknowledgment, we introduce three important causal estimands for this study.

Definition 3 (Treatment Effect)**.** The (individual) treatment effect for a treated unit i, with adoption time $a_i \leq T$, at time $t \geq a_i$, is defined as

$$\delta_{it} = Y_{it}(a_i) - Y_{it}(c),$$

representing the difference between the observed post-treatment outcome and the counterfactual outcome, assuming the unit had never received treatment by period T.

Definition 4 (Sample Average Treatment Effect on the Treated, ATT)**.** The sample average treatment effect on the treated (ATT) for units under treatment for a duration of τ periods is

$$\delta_\tau = \frac{\sum_{i:T-\tau+1\leq a_i\leq T} \delta_{i,a_i+\tau-1}}{N_{\mathrm{tr},\tau}},$$

where $N_{\mathrm{tr},\tau}$ is the number of treated units in the sample that have been under treatment for τ periods.

Definition 5 (Root Mean Square Error, RMSE)**.** Given a longitudinal study with N cross-sectional units observed over T time periods, let $Y_{it}(w_{it})$ denote the observed outcome for unit i at time t, where $w_{it} \in \{0, 1\}$ indicates the absence or presence of treatment. Let $\hat{Y}_{it}(0)$ and $\hat{Y}_{it}(1)$ represent the predicted outcomes under control and treatment conditions, respectively. The root mean square error (RMSE), denoted as ρ, is defined as the square root of the average squared difference between the observed and predicted outcomes, adjusted for the treatment status, across all units and time periods. It is given by

$$\rho = \sqrt{\frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(w_{it} \cdot (Y_{it}(1) - \hat{Y}_{it}(1))^2 + (1 - w_{it}) \cdot (Y_{it}(0) - \hat{Y}_{it}(0))^2 \Big)},$$

where $w_{it} = 1$ if unit i is treated at time t, and $w_{it} = 0$ otherwise.

The treatment effect estimand is a critical indicator for testing the presence of causal effects. In longitudinal studies, the interest often extends to such effects over various periods, which implies the importance of examining the sample ATT. Lastly, the RMSE plays a vital role in causal inference placebo tests. It quantifies the discrepancy between

observed outcomes and those predicted by a model under the null hypothesis of no treatment effect, thereby measuring the effectiveness of control and treatment predictions. In Chapter 5, we derive these estimands through our model from the data, offering readers a comprehensive understanding of these causal estimands with practical implications at that stage.

3.3 Assignment Mechanisms

The subsequent assumption adopted in this study is related to the treatment assignment mechanism. First, we review the concept of the assignment mechanism. Among the three basic restrictions on assignment mechanisms outlined in Imbens and Rubin (2015), we adopt one as our forthcoming assumption.

Definition 6 (Assignment Mechanism)**.** Let there be a finite set of units indexed by $N = \{1, 2, \ldots, n\}$, and let $\mathbf{W}$ be an assignment matrix where w_i corresponds to the allocation of unit $i \in N$. The assignment mechanism, denoted as $\mathbb{P}(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$, is a function mapping the covariate space and potential outcomes to a probability distribution over the Cartesian product $\{0, 1\}^N$, the set of all possible assignments. Formally,

$$\sum_{\mathbf{W} \in \{0,1\}^N} \mathbb{P}(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = 1$$

for every possible realization of the covariate matrix $\mathbf{X}$ and potential outcomes $\mathbf{Y}(0), \mathbf{Y}(1)$. This implies that $\mathbb{P}(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ is row-exchangeable, as it is invariant under any permutation of its index set N.

We assume that the assignment mechanism in this study can be decomposed into individual probabilities for each unit, independent of the assignments of other units. This assumption, called *individualistic assignment*, states that each unit's likelihood of receiving treatment is unaffected by the treatment status of any other unit. For a rigorous definition of *individualistic assignment*, as well as the other two assignment mechanisms, readers are encouraged to read Chapter 3 of Imbens and Rubin (2015).

Assumption 3 (Individualistic Assignment)**.** Consider a population of N units, each denoted by $i \in \{1, 2, \ldots, N\}$. Let $\mathbf{W} = (w_1(a_1), w_2(a_2), \ldots, w_N(a_N))$ represent the vector of adoption times for treatment, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ the vector of covariates, $\mathbf{Y}(0) = (Y_1(0), Y_2(0), \ldots, Y_N(0))$ the vector of observed potential outcomes (under control) for each unit, and $\mathbf{Y}(1) = (Y_1(1), Y_2(1), \ldots, Y_N(1))$ the vector of missing potential outcomes (under treatment) for each unit. The adoption time of unit i, $w_i(a_i)$, is assumed to be independent of the covariates or potential outcomes of other units, and also independent

of their time of adoption, conditional on X_i, $Y_i(0)$, and $Y_i(1)$. Formally,

$$\mathbb{P}(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{i=1}^{N} \mathbb{P}(w_i(a_i)|X_i, Y_i(0), Y_i(1)).$$

Assumption 4 (Positivity)**.** We ensure that each unit has a non-zero probability of being treated, satisfying the condition

$$0 < \mathbb{P}(w_i(a_i)|X_i, Y_i(0), Y_i(1)) < 1 \quad \forall\, i.$$

The *positivity* assumption is essential for the validity of the *individualistic assignment* assumption.

3.4 Posterior Predictive Inference

We note that $\mathbb{P}(w_i(a_i)|X_i, Y_i) = \mathbb{P}(w_i(a_i)|X_i, Y_i(0), Y_i(1))$, indicating that the treatment assignment mechanism may be correlated with $Y_i(1)$, the counterfactual outcome, as discussed by Pang et al. (2021). To prevent potential confounding, it is common to adopt another assumption known as the *ignorability* assumption (Rubin 1978).

Assumption 5 (Ignorability of Treatment Assignment)**.** Let $\mathbf{X}$ represent pre-treatment covariates, $\mathbf{W}$ the treatment assignment, and $\mathbf{Y}(0)$, $\mathbf{Y}(1)$ the potential outcomes under control and treatment, respectively. The treatment assignment is said to be ignorable if it satisfies the following condition:

$$(\mathbf{W} \perp\!\!\!\perp \mathbf{Y}(0), \mathbf{Y}(1)) \,|\, \mathbf{X},$$

where $\perp\!\!\!\perp$ denotes statistical independence.

The assumption specified above indicates that, conditional on the covariates $\mathbf{X}$, the treatment assignment $\mathbf{W}$ is statistically independent of the potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$. This allows for the unbiased estimation of causal effects from observational data by adjusting for $\mathbf{X}$. However, under conditions where data are MNAR, this assumption may not hold. In MNAR scenarios, the relationship between the treatment assignment mechanism and unobserved (missing) outcomes could introduce bias that cannot be mitigated merely by conditioning on the observed covariates and outcomes. Therefore, following Pang et al. (2021), we propose a stricter assumption to address this challenge, which is stated below.

Assumption 6 (Latent Ignorability)**.** The assignment mechanism is independent of any missing or observed untreated outcomes for each unit i, conditional on the observed pre-

treatment covariates X_i and a vector of latent variables $U_i = (u_{i1}, u_{i2}, \ldots, u_{iT})$. That is,

$$\mathbb{P}(w_i(a_i)|X_i, Y_i, U_i) = \mathbb{P}(w_i(a_i)|X_i, Y_i(0), Y_i(1), U_i) = \mathbb{P}(w_i(a_i)|X_i, U_i),$$

where X_i may include both time-varying and time-invariant pre-treatment covariates, and U_i captures unit-level heterogeneity, such as unit fixed effects and unit-specific time trends.

The above assumption is considered an extension of the *strict exogeneity* assumption often assumed in fixed effects (FE) models (Xu 2017). Once we condition on X_i and U_i, the entire time series of Y_i is assumed to be independent of $w_i(a_i)$. This result is analogous to the uncorrelatedness of error terms and covariates in the *strict exogeneity* assumption. We present this finding as a proposition below. It precludes dynamic feedback from past outcomes on current and future treatment assignments, conditional on U_i (Pang et al. 2021).

Proposition 1 (Latent Ignorability and Strict Exogeneity)**.** Latent ignorability extends the concept of strict exogeneity by incorporating latent variables that capture unobserved heterogeneity. For a treatment assignment mechanism $w_i(a_i)$, latent ignorability can be formalized as

$$\mathbb{P}(w_i(a_i)|X_i, U_i) = \mathbb{P}(w_i(a_i)|X_i, Y_i, U_i),$$

where U_i represents the latent variables that are potentially correlated with the unobserved components of the outcome.

Proof. See Appendix A.1. □

Below we state another proposition that connects the concept of *latent ignorability* with the *parallel trends* assumption[1] (Pang et al. 2021). The *latent ignorability* assumption enhances this by considering not only observable covariates but also unobserved factors through latent variables.

Proposition 2 (Latent Ignorability and Parallel Trends)**.** Under the latent ignorability assumption, if the latent variable U_i is a unit-specific constant such that $u_{i1} = u_{i2} = \cdots = u_{iT} = u_i$ for all i, then the parallel trends assumption is satisfied. Specifically, latent ignorability implies that, in the absence of treatment, the untreated potential outcomes for all units would follow a parallel path over time.

Proof. See Appendix A.2. □

[1] In the absence of treatment, the potential outcomes for treated and untreated units would exhibit similar trends over time.

We make an additional assumption, called the *feasible data extraction* assumption, to allow the factorization of unit-specific time trends into multiple common trends with heterogeneous impacts, as discussed in Xu (2017), Athey et al. (2018), Bai and Ng (2021), and Pang et al. (2021). This assumption is fundamental to the factor-augmented approach upon which our model is constructed.

Assumption 7 (Feasibility)**.** For each unit i, it is assumed there exists an unobserved covariate vector U_i, such that for the entire population of N units over T time periods, the stacked $(N \times T)$ matrix $\mathbf{U} = (U_1, \ldots, U_N)$ can be approximated by the product of two lower-rank matrices:

$$\mathbf{U} \approx \mathbf{\Gamma}^\top \mathbf{f},$$

where $\mathbf{f} = (f_1, \ldots, f_T)$ represents a $(r \times T)$ matrix of common factors and $\mathbf{\Gamma} = (\gamma_1, \ldots, \gamma_N)$ denotes a $(r \times N)$ matrix of factor loadings, with the rank $r \ll \min\{N, T\}$.

This approximation suggests that the complex structure of unobserved covariates across units and times can be effectively represented by a limited set of underlying factors (f_t) and their loadings on each unit (γ_i). This mechanism is akin to matrix factorization (MF) and demonstrates a connection to SCM, which is further detailed in Appendix B. However, it is important to note that the *feasibility* assumption might be compromised if unit-specific time trends are highly idiosyncratic.

Before we can fully derive the posterior predictive inference, we further assume that the *exchangeability* assumption is met. This assumption states that the statistical properties of $(X_{it}^{\perp}, Y_{it}(c))$ remain invariant regardless of the observation order. Additionally, we revisit de Finetti's theorem (de Finetti 1963) to provide readers with the necessary background to understand our derivation of the posterior predictive distribution at the end of Chapter 3.

Assumption 8 (Exchangeability)**.** Given a vector of latent variables $\mathbf{U}$, the sequence $\{(X_{it}, Y_{it}(c))\}_{i=1,\ldots,N}^{t=1,\ldots,T}$ is exchangeable. That is, the joint distribution of $\{(X_{it}, Y_{it}(c))\}$ remains invariant to permutations in the indices i and t. Formally, for any permutation π over the set $\mathcal{I} = \{1, \ldots, N\} \times \{1, \ldots, T\}$, it holds that

$$(X_{\pi(i)t}, Y_{\pi(i)t}(c)) \overset{d}{=} (X_{it}, Y_{it}(c)),$$

where $\overset{d}{=}$ denotes equality in distribution.

Theorem 1 (de Finetti 1963)**.** For an infinite sequence of exchangeable binary random variables $(X_1, X_2, \ldots)$, there exists a probability measure μ on $[0, 1]$ such that the joint distribution of any finite subsequence $(X_1, \ldots, X_n)$ is a mixture of independent and identically distributed (i.i.d.) Bernoulli distributions. Specifically, for any n and any particular

sequence $(x_1, \ldots, x_n)$ in $\{0,1\}^n$, we have

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^s (1-\theta)^{n-s} d\mu(\theta),$$

where $s = \sum_{i=1}^n x_i$ is the number of 1's in the sequence $(x_1, \ldots, x_n)$.

Proof. See Kirsch (2019). □

Following the approach in Pang et al. (2021), we derive the posterior predictive distribution of the counterfactual outcome $\mathbf{Y}(1)$ as

$$\begin{aligned}
&\mathbb{P}(\mathbf{Y}(1)|\mathbf{X}, \mathbf{U}, \mathbf{Y}(0), \mathbf{W}) \\
&\propto \mathbb{P}(\mathbf{X}, \mathbf{U}, \mathbf{Y}(0), \mathbf{Y}(1)) \cdot \mathbb{P}(\mathbf{W}|\mathbf{X}, \mathbf{U}, \mathbf{Y}(0), \mathbf{Y}(1)) && \text{(Bayes' Theorem)} \\
&\propto \mathbb{P}(\mathbf{X}, \mathbf{U}, \mathbf{Y}) \cdot \mathbb{P}(\mathbf{W}|\mathbf{X}, \mathbf{U}, \mathbf{Y}) && \text{(Latent Ignorability)} \\
&\propto \mathbb{P}(\mathbf{X}^*, \mathbf{Y}) \cdot \mathbb{P}(\mathbf{W}|\mathbf{X}^*, \mathbf{Y}) && (\mathbf{X}^* = (\mathbf{X}, \mathbf{U})) \\
&\propto \mathbb{P}(\mathbf{X}^*, \mathbf{Y}) && \text{(Normalizing Constant)} \\
&\propto \mathbb{P}\left(\{X_{it}^*, Y_{it}\}\right) && \text{(Exchangeability)} \\
&\propto \int \prod_{it \in S_{\text{mis}}} f(Y_{it}(1)|X_{it}^*, \theta^*) \prod_{it \in S_{\text{obs}}} f(Y_{it}(0)|X_{it}^*, \theta^*) \pi(\theta^*)\, d\theta^*. && \text{(de Finetti's Theorem)}
\end{aligned}$$

We apply Bayes' theorem in the second line. Then, we apply our *latent ignorability* assumption and proceed to the third line. In the fourth line, we consider $\mathbf{X}^* = (\mathbf{X}, \mathbf{U})$, which is a collection of covariates and latent variables. The fifth line omits the normalizing constant term $\mathbb{P}(\mathbf{W}|\mathbf{X}^*, \mathbf{Y})$ since this treatment assignment mechanism does not depend on $\mathbf{Y}(1)$. The penultimate line applies the *exchangeability* assumption, where each $\mathbb{P}\left(\{X_{it}^*, Y_{it}\}\right)$ is assumed to be i.i.d., given some parameters and their prior distributions. We apply de Finetti's theorem to arrive at the last line, deriving that θ^* is the parameter governing the DGP of Y_{it}, conditioned on X_{it}^* and $\theta^* = (\theta, \mathbf{U})$. We present this development as a proposition, with a rigorous proof available in Appendix A.1 of the Supplementary Materials of Pang et al. (2021).

Proposition 3 (Posterior Predictive Distribution)**.** Given covariates $\mathbf{X}$, latent variables $\mathbf{U}$, observed outcomes $\mathbf{Y}(0)$, and treatment assignment $\mathbf{W}$, the posterior predictive distribution of the counterfactual outcome $\mathbf{Y}(1)$ is derived as

$$\mathbb{P}(\mathbf{Y}(1)|\mathbf{X}, \mathbf{U}, \mathbf{Y}(0), \mathbf{W}) \propto \int \underbrace{\prod_{it \in S_{\text{mis}}} f(Y_{it}(1)|X_{it}^*, \theta^*)}_{\text{posterior predictive distribution}} \underbrace{\prod_{it \in S_{\text{obs}}} f(Y_{it}(0)|X_{it}^*, \theta^*)}_{\text{likelihood}} \pi(\theta^*)\, d\theta^*,$$

where S_{mis} and S_{obs} denote the partitioning sets of missing and observed data indices, respectively, and θ^* are the parameters governing the DGP of Y_{it}, conditioned on X_{it}^* and latent parameters θ^*.

Proof. See Appendix A.1 of the Supplementary Materials of Pang et al. (2021). □

This concludes Chapter 3. In Chapters 4 and 5, we define and later implement our model in two empirical applications to demonstrate the core essence of our model and assess its performance against existing extensions in SCM.

4

Methodology

Chapter 4 can be divided into two parts. In the first part, we present a well-known application of SCM — the German reunification — as motivation. This discussion will cover the problem of interest and how it aligns with our paradigm of block structure, as introduced in Section 3.1. The second part of Chapter 4 then explores previous methods that have attempted to address this problem. We will reintroduce these methods using consistent notation and ultimately derive the final functional form needed to implement our model.

4.1 German Reunification

The event of German reunification unfolded between November 9, 1989, and March 15, 1991. The German Democratic Republic (East Germany) joined the Federal Republic of Germany (West Germany), marking the end of a division that had been in place since the end of World War II. The reunification of East and West Germany in 1991 is often considered an important social science quasi-natural experiment (Redding and Sturm 2008), where, for instance, West Germany serves as our unique treatment unit. We have data on GDP per capita for West Germany and other countries. Assuming we also have the ability to collect covariates that could potentially influence GDP per capita growth, the question arises: Can we leverage the existing data to estimate what the GDP per capita of West Germany would have been had it not united with East Germany in 1991?

The immediate answer to this question is straightforward. We consider the other countries as our control units, or a donor pool (Abadie et al. 2014). Thanks to Hainmueller (2014), replicated data for German reunification are available. This dataset includes 17 OECD member countries (including West Germany, the USA, the UK, Switzerland, and others) with annual data from 1960 to 2003. The data contain a single outcome variable $\mathbf{Y}$, GDP per capita for West Germany, which is adjusted for Purchasing Power Parity (PPP) and measured in 2002 USD. Additionally, the dataset includes a set of standard economic predictors $\mathbf{X}$, such as average trade openness, average inflation rate, average industry share of value added from 1981 to 1990, average percentage of secondary school attainment in the total population aged 25 and older from 1980 to 1985, and average investment rate from 1975 to 1980. For simplicity, we treat German reunification as a non-duration time event that occurred in 1991. Hence, the pre-intervention period spans from 1960 to 1990 (inclusive), and the post-intervention period is from 1991 to 2003. Table 4.1 shows the pre-reunification characteristics of West Germany alongside the population-weighted average of the other 16 OECD countries in the donor pool.

Table 4.1: Economic Indicators for West Germany and OECD Sample

Indicator (Units)	**West Germany**	**OECD Sample**
GDP per capita (USD)	15808.9	13669.4
Trade openness (%)	56.8	59.8
Inflation rate (%)	2.6	7.6
Industry share (%)	34.5	33.8
Schooling (%)	55.5	38.7
Investment rate (%)	27.0	25.9

Clearly, we see from Table 4.1 that the pre-reunification characteristics do not align well if we simply consider the population-weighted average. The essence of computing the counterfactual GDP per capita for West Germany lies in aligning the pre-treatment characteristics ($\mathbf{X}$) and outcome ($\mathbf{Y}$) effectively. By assigning different weights to each OECD member country, where the weights can be obtained via a convex optimization algorithm (Abadie et al. 2014), we construct a counterfactual West Germany sample that matches the pre-treatment data, including both the outcome and covariates. The core idea here is to turn the observational data into a quasi-natural experiment, as long as this process can control for those unobserved variables.

In Figure 4.1, we display the outcome variable ($\mathbf{Y}$) in a time-series plot. There is no missingness across $\mathbf{Y}$; however, GDP per capita after 1990 for West Germany is fundamentally different from its former regime. The observed data for West Germany from 1991 to 2003 should instead be considered as MNAR, where we could apply our model to impute the missingness and compute the GDP per capita of the counterfactual West Germany had reunification not occurred. Figure 4.2 illustrates that West Germany, the

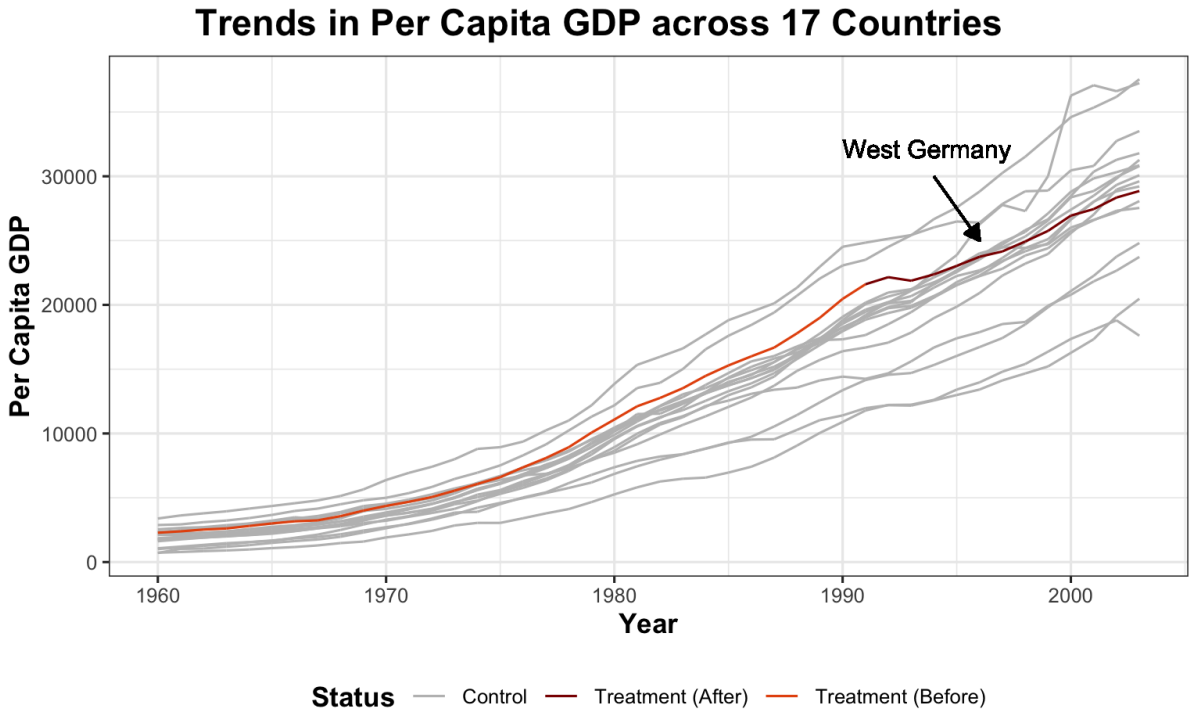

Figure 4.1: Trends in Per Capita GDP across 17 Countries

treated unit, is considered to have missing data after the intervention (colored dark red).

Recall Definition 3, the treatment effect can then be estimated via

$$\delta_{\text{West Germany},t} = Y_{\text{West Germany},t}(1) - Y_{\text{West Germany},t}(0),$$

where $\delta_{\text{West Germany},t}$ is minimized to 0 before the intervention (year 1990, inclusive). Many researchers have attempted to demonstrate a negative treatment effect resulting from the German reunification (Abadie et al. 2014; Pinkney 2021; Pang et al. 2021). How confident are their claims? In Section 4.2, we review and replicate the methods that researchers have employed to address this question.

4.2 Modeling

Now, we review two models: the standard SCM as proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2014), along with the Bayesian alternative to the standard SCM, complemented by the IFE model as proposed by Pinkney (2021). Then, we introduce our Bayesian causal MC model. Broadly speaking, we demonstrate how our model integrates into the interdisciplinary area of Bayesian causal inference, with applications in econometric modeling, and probabilistic machine learning, with applications

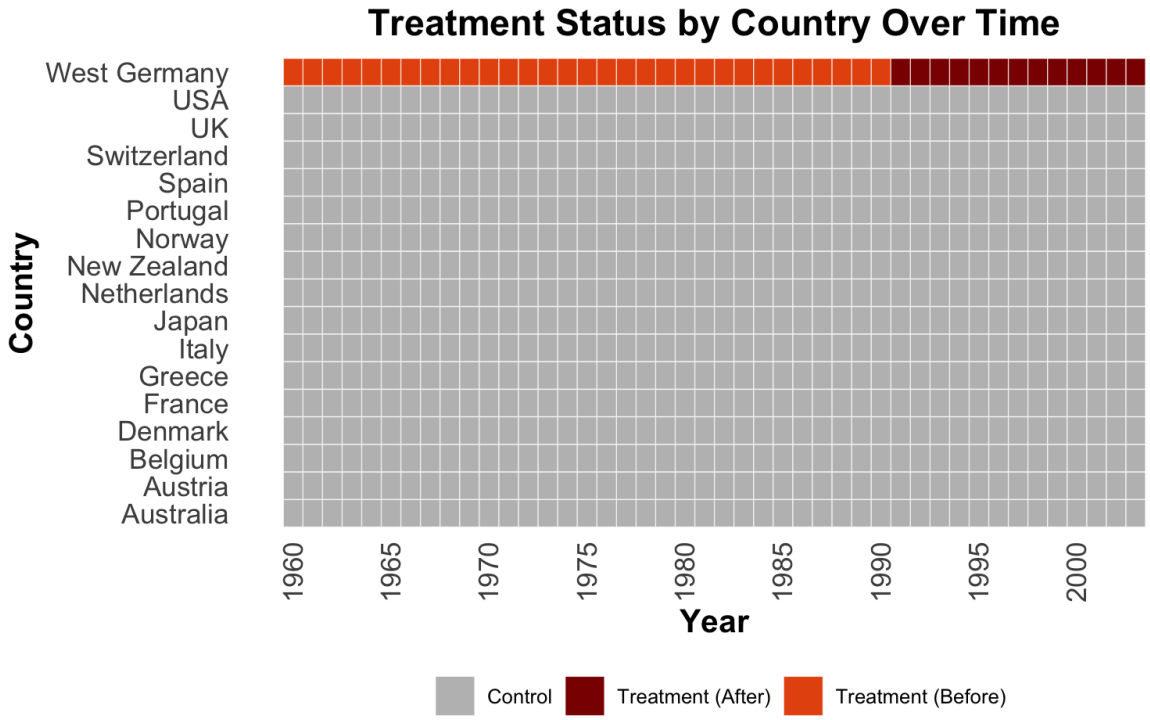

Figure 4.2: Treatment Status by Country Over Time

in recommender systems[1].

We regard Bayesian causal MC as the most generalizable variant within the SCM family. To demonstrate its dual advantages, we analyze its performance against other existing SCM extensions (standard SCM and Bayesian SCM with IFE) in terms of efficiency and illustrate how it fits our paradigm of large N, small T, and multiple P for enhanced generalizability.

4.2.1 Underlying Factor Model of Standard SCM

SCM, initially introduced by Abadie and Gardeazabal (2003) and further developed by Abadie et al. (2010), serves as a fundamental approach in causal inference studies. In a longitudinal study with $N+1$ units observed over T time periods, where one unit ($j = 0$) is treated and N units act as potential controls within a donor pool, SCM constructs a *synthetic control* to estimate counterfactuals. For unit i at time t, let Y_{it} denote the observed outcome, with $Y_{it}(1)$ and $Y_{it}(0)$ representing potential outcomes under treatment and control, respectively. The treatment assignment occurs at time T_0+1, differentiating pre-treatment periods ($t = 1, \ldots, T_0$) from post-treatment periods ($t = T_0 + 1, \ldots, T$).

[1]Due to the length, we only cover the first perspective on Bayesian causal inference in Chapter 4. For readers interested in gaining an understanding of our proposed model from a probabilistic machine learning perspective, please visit Appendix B.

The synthetic control's outcome for the treated unit, $\hat{Y}_{0t}$, is computed as

$$\hat{Y}_{0t} = \sum_{j=1}^{N} \hat{\omega}_j Y_{jt}(0),$$

where the weights $\hat{\omega}$ are optimized to minimize the squared difference between the treated unit's pre-treatment outcomes and the weighted average of the control units, subject to the constraints that the weights are non-negative and sum to one.

Model 1 (Synthetic Control Method, SCM)**.** The underlying factor model of SCM is

$$Y_{it} = w_{it}^{\top}\delta_{it} + X_i^{\top}\xi_t + \alpha_t + \Gamma_i^{\top} f_t^p + \varepsilon_{it},$$

and its matrix representation is

$$\mathbf{Y} = \mathbf{W}^{\top}\boldsymbol{\delta} + \mathbf{X}^{\top}\boldsymbol{\xi} + \boldsymbol{\alpha} + \boldsymbol{\Gamma}^{\top}\mathbf{f} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is the matrix of potential outcomes for all units across times, including both control and treatment cases. $\mathbf{W}$, the binary treatment indicator matrix, assigns treatment status across units and times. δ_{it}, represented by $\boldsymbol{\delta}$ in matrix form, is the heterogeneous treatment effect (HTE) for unit i at time t. $\mathbf{X}$ aggregates observed covariates into a matrix, with columns for specific covariates and rows corresponding to units at different times. Hence, X_i is the vector of covariates specific to unit i, structured within $\mathbf{X}$. ξ_t, denoted by $\boldsymbol{\xi}$, captures time-varying effects. α_t, represented by $\boldsymbol{\alpha}$, denotes fixed effects associated with time t. f_t, constituting the matrix $\mathbf{f}$, is the $(1 \times L)$ vector of unobserved common factors. Γ_i, constituting the matrix $\boldsymbol{\Gamma}$, is the $(L \times 1)$ vector of unknown factor loadings. Lastly, ε_{it}, the error term for unit i at time t, is compiled in the error matrix $\boldsymbol{\varepsilon}$.

In Model 1, we combine the factor model discussed by Abadie et al. (2010) with the SCM component from Abadie et al. (2014), specifically, the synthetic control's outcome part. Here, we introduce the matrix representation and rearrange the matrix multiplication order as in Abadie et al. (2010). Model 1 enables the estimation of the causal effect of the intervention on the treated unit during post-intervention periods, denoted as $\delta_{0t} = Y_{0t}(1) - \hat{Y}_{0t}$ for $t > T_0$. This estimator represents a more generalizable form of the DiD estimator, given the relaxation of underlying assumptions. The relationship between them is presented as a proposition below.

Proposition 4 (SCM as a Generalized Form of DiD)**.** SCM can be viewed as a generalized form of the DiD estimator when SCM assigns equal weights to control units that satisfy the parallel trends assumption with the treated unit. Under these conditions, and assuming additive effects, SCM could replicate the DiD estimator.

Proof. See Appendix A.3. □

4.2.2 Functional Form of Bayesian SCM with IFE

To understand Pinkney's (2021) Bayesian SCM with IFE model, we divide it into two components: the pure Bayesian version of the standard SCM (Tuomaala 2019; Kim et al. 2020) and the IFE model in LFM (Bai 2009; Xu 2017), which belongs to the family of panel data models in econometrics. Below, we reintroduce Pinkney's (2021) Bayesian SCM IFE model by correcting the notation used in their proposed functional forms.

Model 2 (Bayesian Synthetic Control Method with Interactive Fixed Effects, Bayesian SCM IFE)**.** We define the following functional form to accurately incorporate interactive fixed effects

$$Y_{it} = w_{it}^{\top}\delta_{it} + X_{it}^{\top}\xi + \Gamma_i^{\top} f_t + \varepsilon_{it},$$

where Y_{it} denotes the potential outcome for unit i at time t, $w_{it}^{\top}\delta_{it}$ represents the treatment effect for unit i at time t, $X_{it}^{\top}\xi$ captures the effects of observed covariates, $\Gamma_i^{\top} f_t$ describes the interaction between unit-specific factor loadings and common latent factors, and ε_{it} accounts for the idiosyncratic error term.

Its matrix representation is

$$\mathbf{Y} = \mathbf{W}^{\top}\boldsymbol{\delta} + \mathbf{X}^{\top}\boldsymbol{\xi} + \boldsymbol{\Gamma}^{\top}\mathbf{f} + \boldsymbol{\varepsilon},$$

with $\mathbf{Y}$ representing the matrix of potential outcomes across all units and time periods, $\mathbf{W}^{\top}\boldsymbol{\delta}$ capturing the matrix of treatment effects across all units and time periods, $\mathbf{X}$ as the covariate matrix with coefficients $\boldsymbol{\xi}$, $\boldsymbol{\Gamma}$ signifying the matrix of unit-specific factor loadings, $\mathbf{f}$ as the matrix of common latent factors, and $\boldsymbol{\varepsilon}$ compiling the error terms.

Model 2 enhances Bayesian SCM by integrating the IFE model, initially proposed by Bai (2009), and further elaborated by Xu (2017) in a generalized SCM. Model 2 overcomes the limitations of generalized SCM through a two-step estimation process: initially estimating IFEs for the control group and subsequently capturing the treated unit's latent factors via their factor loadings in the pre-treatment phase. However, Pinkney (2021) critiques this methodology for potentially reducing estimation efficiency due to the separate fitting of latent factors and loadings. By estimating latent factors concurrently while preserving the treated unit's data in the treatment phase, Model 2 effectively utilizes more data for estimation, yielding comprehensive uncertainty distributions for each parameter.

Bayesian SCM IFE incorporates the components $X_{it}^{\top}$, $\Gamma_i^{\top}$, and f_t, alongside an idiosyncratic error term ε_{it}. Adhering to the methodological underpinnings suggested by Farouni (2015), Model 2 employs a simplified approach for estimating Bayesian latent factor loadings and weights. This approach ensures the factors f_t are uncorrelated, and applies constraints on the factor loading matrix $\Gamma_i^{\top}$ to set upper-triangular elements to zero

and ensure positivity in the diagonal elements, therefore enhancing interpretability and estimation efficiency.

4.2.3 Functional Form of Bayesian Causal MC

In this thesis, we present two versions of Bayesian causal MC models, each with its advantages and preferred applications. The first model, Bayesian causal MC with independent multiple P's, extends the functional forms of Bayesian SCM as discussed in Pang et al. (2021) and MC in Athey et al. (2021). This model enables multi-outcome modeling through either an iterative or a separate Bayesian hierarchical modeling process with dynamic factors. The second model, Bayesian causal MC with concurrent multiple P's, employs an original approach that allows for simultaneous multi-outcome modeling by utilizing a shared matrix *singular value decomposition* (SVD), comparable to SCM's *factorization*. This approach helps to jointly identify individual-level latent factors to establish conditional ignorability. Although these Bayesian causal MC models differ in their functional forms and specific implementations, they lead to similar expected outcomes. Both models are well-suited for addressing challenges with large N, small T, and multiple P, and we aim to demonstrate this in later empirical applications.

Model 3 (Bayesian Causal Matrix Completion with Independent Multiple P's)**.** [2] Following the functional form proposed by Pang et al. (2021), we introduce a linear model that estimates the counterfactual outcome for unit i at time t and outcome dimension p, such that

$$Y_{it}^{p} = \underbrace{X_{it}^{\top}\xi^{p}}_{\text{Constant Effects}} + \underbrace{Z_{it}^{\top}\zeta_{i}^{p}}_{\text{Unit-level Effects}} + \underbrace{A_{it}^{\top}\alpha_{t}^{p}}_{\text{Time-level Effects}} + \underbrace{\Gamma_{i}^{\top}f_{t}^{p}}_{\text{Latent Factors}} + \underbrace{\varepsilon_{it}^{p}}_{\text{Error Term}},$$

for $p = 1, 2, \ldots, P$. The matrix representation of the model is

$$\mathbf{Y}^{p} = \mathbf{X}^{\top}\boldsymbol{\xi}^{p} + \mathbf{Z}^{\top}\boldsymbol{\zeta}^{p} + \mathbf{A}^{\top}\boldsymbol{\alpha}^{p} + \boldsymbol{\Gamma}^{\top}\mathbf{f}^{p} + \boldsymbol{\varepsilon}^{p},$$

where $\mathbf{Y}^{p}$ is the matrix of potential outcomes for all units and times under outcome dimension p. $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{A}$ are matrices of covariates with constant effects, unit-level random effects, and time-level random effects, respectively, each associated with their coefficient vectors $\boldsymbol{\xi}^{p}$, $\boldsymbol{\zeta}^{p}$, and $\boldsymbol{\alpha}^{p}$. The term $\boldsymbol{\Gamma}^{\top}\mathbf{f}^{p}$ captures the contribution of latent factors, with $\boldsymbol{\Gamma}$ being the matrix of unit-specific factor loadings and $\mathbf{f}^{p}$ the matrix of latent factors. $\boldsymbol{\varepsilon}^{p}$ represents the matrix of error terms, assumed to be normally distributed with mean zero and variance σ^{2}.

[2] We have revised certain components in `bpCausal`, an R software developed by Pang et al. (2021). We introduce a new framework, called `BCMC`, designed to be effectively implemented in scenarios with large N, small T, and multiple P. Detailed replication codes, including the revised parts, can be found in Appendix C.3.

Model 4 (Bayesian Causal Matrix Completion with Concurrent Multiple P's)**.** [3] We define the jointly encoding multi-output SCM latent factor specification as a 3-way factorization

$$Y_{it}^p = f_t \Sigma^p \gamma_i + \varepsilon_{it}^p,$$

or, in its matrix representation,

$$\mathbf{Y}^p = \mathbf{f} \boldsymbol{\Sigma}^p \boldsymbol{\Gamma} + \boldsymbol{\varepsilon}^p,$$

where $\mathbf{Y}^p$ is the matrix of potential outcomes for all units across all times under outcome dimension p. The matrix $\mathbf{f}$ is $(T \times L)$, with the row vector f_t corresponding to the L-dimensional latent factors at time t. $\boldsymbol{\Sigma}^p$ is the $(L \times L)$ diagonal scaling matrix, unique to each outcome p, scaling the impact of the latent factors. The matrix $\boldsymbol{\Gamma}$ is $(L \times N)$ and contains the factor loadings for each unit, with γ_i being the column vector for unit i. The matrix $\boldsymbol{\varepsilon}^p$ is the $(T \times N)$ matrix of error terms for outcome dimension p, with each element ε_{it}^p assumed to be normally distributed with mean zero and variance σ^2.

4.3 Estimation and Inference

Algorithm 1 employs Gibbs sampling, adapted from Pang et al. (2021), to iteratively estimate parameters of the Bayesian causal MC model for multiple outcomes. Starting with robust parameter estimation, the algorithm leverages Gibbs sampling to sequentially update posterior distributions. This method makes use of untreated observational data, applying Bayesian shrinkage to minimize parameter uncertainty. Following parameter estimation, the algorithm systematically generates predictive draws, which are then used to construct counterfactuals iteratively.

Algorithm 2 outlines another Bayesian causal MC model that manages multiple outcomes simultaneously, rather than iteratively. Beginning with a shared set of input data, we apply Hamiltonian Monte Carlo (HMC) with the No-U-Turn Sampler (NUTS) to efficiently explore the posterior distribution and estimate model parameters. Predictions for counterfactuals are then generated for all outcomes simultaneously. This simultaneous prediction phase improves the computation of counterfactuals, potentially enhancing consistency and correlated accuracy across multiple outcome dimensions. The third step involves summarizing the posterior distribution of the predicted outcomes and conducting diagnostic tests to ensure the model's convergence and the validity of its inferences. Compared to the iterative approach, this simultaneous method may offer a more cohesive understanding of the outcomes.

[3] The algorithms for concurrent multiple P's are implemented in `JAX` and `NumPyro`. The core functions for Model 4 can be found in Appendix C.4.

Algorithm 1 Bayesian Causal Matrix Completion with Independent Multiple P's

Input: Observed data $\{(X_{it}, Y_{it}(0))\}$ for all units i and times t in the control period, set of untreated observations S_{obs}, number of draws G, number of outcomes P
Output: Posterior samples for parameters, counterfactual estimates, inference diagnostics

▷ Step 1: Model Parameter Estimation

for $g \leftarrow 1$ **to** G **do**
    Estimate Bayesian causal MC model parameters using Bayesian shrinkage
    Obtain posterior samples for parameters conditional on Θ
end for

▷ Step 2: Prediction and Integration

for each treated unit i in the interval $a_i \leq t \leq T$ **do**
    Generate posterior predictive draws of $Y_{it}(1)$
    Construct empirical integration for counterfactuals
end for

▷ Step 3: Inference and Diagnostics

for each treated unit i **do**
    Summarize the empirical posterior distribution of δ_{it}
    Calculate the posterior mean, variance, and 95% credible intervals
end for
Perform Bayesian diagnostic tests on posterior distributions

Algorithm 2 Bayesian Causal Matrix Completion with Concurrent Multiple P's

Input: Observed data $\{(X_{it}, Y_{it}(0))\}$ for all units i and times t in the control period, number of latent factors L, number of outcomes P
Output: Estimated parameters $\mathbf{f}$, $\{\mathbf{\Sigma}^p\}_{p=1}^P$, $\mathbf{\Gamma}$, counterfactual outcomes $\{\mathbf{Y}^p(1)\}_{p=1}^P$, posterior samples, and diagnostics

▷ Step 1: Model Parameter Estimation

for $p \leftarrow 1$ **to** P **do**
    Define Bayesian causal MC model for outcome p with Bayesian shrinkage
    Data augment $Y_{it} = \{Y_{it}(0), Y_{it}^p(1)\}$
    Execute HMC with NUTS to sample from posterior distributions
end for

▷ Step 2: Prediction and Integration

for $p \leftarrow 1$ **to** P **do**
    for all treated units i and times t **do**
        Generate posterior predictive draws of $Y_{it}^p(1)$
        Aggregate posterior predictions to form $\mathbf{Y}^p(1)$
    end for
end for

▷ Step 3: Inference and Diagnostics

for $p \leftarrow 1$ **to** P **do**
    Summarize the posterior distribution of $\mathbf{Y}^p(1)$
    Calculate the mean, variance, and 95% credible intervals
end for
Conduct diagnostic tests on MCMC convergence and mixing

4.4 Generalization

Building on Pang et al.'s (2021) generalization of existing SCM extensions, we also present our Bayesian causal MC framework, in both its independent and concurrent versions, as a Bayesian alternative generalized method for SCM. Our work draws inspiration from Athey et al. (2021) to bridge the SCM literature in econometrics with the MC literature in recommender systems. Specifically, we illustrate how our Bayesian causal MC method with independent multiple P's can be viewed as the most generalizable form within the SCM family. Following the claims made by Pang et al. (2021) and Athey et al. (2021), we present Proposition 5. In a similar way, we discuss how our model could be seen as a generalized form of Pinkney's (2021) Bayesian SCM IFE in the originally stated Proposition 6.

Proposition 5 (Bayesian Causal MC as a Generalized Form of Standard SCM)**.** Consider the Bayesian causal MC model with independent multiple P's for unit i at time t and outcome dimension p

$$Y_{it}^{p} = w_{it}^{\top}\delta_{it}^{p} + X_{it}^{\top}\xi^{p} + Z_{it}^{\top}\zeta_{i}^{p} + A_{it}^{\top}\alpha_{t}^{p} + \Gamma_{i}^{\top}f_{t}^{p} + \varepsilon_{it}^{p},$$

where we incorporate the HTE δ_{it}^{p} for unit i at time t with outcome dimension p and binary treatment indicator w_{it}, without loss of generality. This model generalizes the underlying factor model of Abadie et al.'s (2010) SCM

$$Y_{it} = w_{it}^{\top}\delta_{it} + X_{i}^{\top}\xi_{t} + \alpha_{t} + \Gamma_{i}^{\top}f_{t}^{p} + \varepsilon_{it},$$

by setting $Z_{it} = \varnothing$ and $X_i = A_i$, which disallows $A_{it}^{\top}\alpha_{t}^{p}$ to vary over time, as well as considering only a single outcome dimension. Hence, we can recover Abadie et al.'s (2010) SCM via our Bayesian causal MC model.

Proof. See Appendix A.4. □

Proposition 6 (Bayesian Causal MC as a Generalized Form of Bayesian SCM IFE)**.** Consider the Bayesian causal MC model as defined in Proposition 5. This model also generalizes Pinkney's (2021) Bayesian SCM IFE

$$Y_{it} = w_{it}^{\top}\delta_{it} + X_{it}^{\top}\xi + \Gamma_{i}^{\top}f_{t} + \varepsilon_{it},$$

by setting $Z_{it} = A_{it} = \varnothing$ and considering only a single outcome dimension. Hence, we can recover Pinkney's (2021) model, as well as other latent factor models (e.g., Xu 2017), via our Bayesian causal MC.

Proof. See Appendix A.5. □

5

Empirical Application

In Chapter 5, we reinvestigate the German reunification empirical example by applying Abadie et al.'s (2014) SCM, Pinkney's (2021) Bayesian SCM IFE, and our Bayesian causal MC model. The first empirical application primarily serves as an effectiveness comparison since it involves only one treated unit and has less complicated structures than panel data characterized by large N, small T, and multiple P. In the second application, we test our proposed model on a specific CRM panel dataset. We begin by analyzing its longitudinal structure, fitting it into our modeling framework, and eventually discussing key findings as well as the model's performance and diagnostics.

5.1 Implementation in German Reunification

5.1.1 Replication of Standard SCM

We first re-implement Abadie et al.'s (2014) SCM by directly modifying their `Synth` package in `R` (Abadie et al. 2011) based on specified regulations. To successfully replicate the German reunification study, we follow the exact steps described in Abadie et al. (2014), where we include a set of time-invariant covariates (trade openness, inflation rate, industry share, schooling, and investment rate, all summarized by a sufficient statistic, mean, across time periods; recall Table 4.1) on the side of $\mathbf{X}$. To be consistent with Abadie et al.'s (2014) SCM, we do not include fixed effects in this study, although Model

1 does allow the addition of such effects. Abadie et al.'s (2014) model then estimates the synthetic control weight for the rest of the 16 OECD countries efficiently by minimizing a constrained minimization problem

$$\hat{\beta} = \arg\min_{\beta \in \Lambda} \sum_{t=1}^{T_0} \left(Y_{0t} - \beta_0 - \sum_{j=1}^{J} \beta_j Y_{jt} \right)^2 ,$$

where β_0 is the intercept, and the constraints imposed on β are defined as

$$\Lambda = \left\{ \beta \in \mathbb{R}^{J+1} : \beta_0 = 0, \beta_j \geq 0 \text{ for } j = 1, \ldots, J \text{ and } \sum_{j=1}^{J} \beta_j = 1 \right\}.$$

The algorithm produces two sets of weights, where we denote w_{var} (weights for 6 predictive indicators in $\mathbf{X}$) and w_{ctr} (weights for the rest of the 16 OECD countries), all subject to the constraints that the weights are non-negative and sum to one (Abadie et al. 2014). We present the computed weights for both sets in Table 5.1 below.

Table 5.1: Weights for Economic Indicators and OECD Countries

Indicator	w_{var}	**Country**	w_{ctr}
GDP per capita (USD)	0.442	USA	0.219
Trade openness (%)	0.134	UK	0.001
Inflation rate (%)	0.072	Austria	0.418
Industry share (%)	0.001	Belgium	0.001
Schooling (%)	0.107	Denmark	0.001
Investment rate (%)	0.245	France	0.001
		Italy	0.001
		Netherlands	0.090
		Norway	0.001
		Switzerland	0.111
		Japan	0.155
		Greece	0.000
		Portugal	0.000
		Spain	0.001
		Australia	0.000
		New Zealand	0.000

Due to data missingness in several covariates, Abadie et al. (2014) took an approach to only consider the average rate for industry share between 1981 and 1990, average schooling between 1980 and 1985, and investment rate in 1980. Many researchers who have replicated the German reunification study (Pinkney 2021; Pang et al. 2021) have followed the same procedure. However, we apply multivariate imputation by chained equations (MICE) via the `mice` package in `R` (van Buuren and Groothuis-Oudshoorn 2011) to impute the missing data in covariates. In particular, we evaluate average rates

across industry share, schooling, and investment for the entire T, rather than selecting specific periods as chosen by Abadie et al. (2014). The computed w_{var} and w_{ctr} are slightly different from those in the exact replication by Abadie et al. (2014). With these two sets of weights computed, we apply the equation

$$\hat{Y}_{0t} = \sum_{j=1}^{16} \hat{\omega}_{\text{var},j} Y_{jt}(0)$$

to obtain Table 5.2 below[1].

Table 5.2: Construction of Synthetic West Germany in Comparison with West Germany

Indicator	**West Germany**	**Synthetic West Germany**
GDP per capita (USD)	15808.9	15802.2
Trade openness (%)	56.8	56.9
Inflation rate (%)	2.6	3.5
Industry share (%)	34.5	34.3
Schooling (%)	55.5	55.2
Investment rate (%)	27.0	27.0

We now see that synthetic West Germany, constructed from a set of 16 OECD countries with synthetic weights, aligns well with West Germany in pre-treatment characteristics. This quasi-experimental design helps us discover a counterfactual West Germany, which allows us to perform counterfactual estimation and also obtain treatment effects and various other causal estimands, as we first introduced in Section 3.2. The time-series trend for counterfactual West Germany is represented by the solid yellow line in Figure 5.1. We observe that the outcome of interest, GDP per capita (USD), aligns almost exactly the same as that of observed West Germany (in brown solid line) in pre-treatment periods (before 1990).

5.1.2 Replication of Bayesian SCM IFE

We then replicate Pinkney's (2021) Bayesian SCM IFE from the provided `Stan` codes. We import exactly the same dataset that we used in replicating Abadie et al.'s (2014) SCM. Recall Model 2, the additional component incorporated latent factors. To stress the sparsity-inducing horseshoe+ prior, we follow exactly Pinkney's (2021) choice on $L = 8$, which doubles Tuomaala's (2019) choice, so that we could closely replicate Pinkney's (2021) German reunification study. However, with multiple testings on the number of latent factors, any choice between $L = 6$ and $L = 10$ is reasonable.

After successfully revising and compiling the `Stan` codes and data list, we then set the

[1]The exact replicated codes can be found in Appendix C.1. I also detail a tutorial for re-implementing Abadie et al.'s (2014) SCM, and the link can also be accessed in Appendix C.

initial value to 0.1 with `max_treedepth` at 14 and `adapt_delta` set to 0.95, exactly based on Pinkney's (2021) selection. Also, the fit is performed using `Stan` with 4 parallel chains, using 250 warm-up iterations and 250 post-warmup iterations. I only pick half the number of Pinkney's (2021) choice in order to speed up the long fitting process. We analyze the fit on MCMC diagnostics in Appendix C.2, and the detailed replication codes and additional supplemental implementation findings can also be found there.

To access the posterior distribution for counterfactual West Germany, i.e., the posterior sample, which is in dimension 250, 4, and 2552, where 250 implies that there are 250 iterations, 4 implies that there are 4 parallel chains, and 2552 implies that there are a total of 2552 parameters. The dimension for our desired samples for counterfactual West Germany has 748 parameters, which is expected since $748 = 17 \times 44$. Now, we create an empty array with dimensions 250, 4, 17, and 44. With a nested loop, we enter each country first and then access the GDP per capita (USD) in each year. Then, we report this matrix back to our giant array. The giant array is indeed composed of nested matrices. Think of this array as a 17×44 matrix. Then, in each singular cell, it is again a 250×4 matrix, displaying GDP per capita (USD) under all iterations and chains for this specific country in the specific year. With this giant array set up, we eventually enter the nested loop again to extract some useful information for Bayesian inference. We extract the mean, 2.5%, 97.5%, and mid-50% GDP per capita (USD) for each country under each year. The time-series trend for counterfactual West Germany is represented by the dashed blue line in Figure 5.1. For better demonstration purposes, we exclude the mid-50% credible intervals (CI) in Figure 5.1. However, one interested in this replication may refer to Appendix C and gain a deeper understanding of Pinkney's (2021) Bayesian SCM IFE through our detailed replication tutorial.

5.1.3 Implementation of Bayesian Causal MC

Finally, we implement our proposed Bayesian causal MC model in the German reunification study as our first example. Given the length of constraints in this honors thesis, we only implement Model 3, which involves the use of independent multiple P's rather than concurrent multiple P's. Since the German reunification study only has one outcome of interest, i.e., GDP per capita (USD), either choice of our Bayesian causal MC models works exactly the same.

We also import exactly the same dataset that we used in replicating Abadie et al.'s (2014). Although our model could enable the addition of unit fixed effects and unit-varying coefficients, we do not consider them in the German reunification replication since our replications for Abadie et al.'s (2014) SCM and Pinkney's (2021) Bayesian SCM IFE do not include them as well. However, our model incorporates time-varying coefficients, where the other two models could not handle such effects, even though we

test and show that all the covariate coefficients are almost constant over time, according to Pang et al. (2021). For latent factor selections, our model produces a rather different implication for the suggestion of L. We test for the posterior distribution of a scaling parameter to capture the importance of the corresponding factor, and it indicates that any factors between 4 and 6 should work perfectly since they exhibit bimodal posteriors, while several other factors show mixed posteriors (Pang et al. 2021). However, we still pick $L = 8$ to serve as a baseline comparison with Pinkney's (2021) Bayesian SCM IFE.

In implementing our Bayesian causal MC model, we import Pang et al.'s (2021) `bpCausal` and pre-specify the following parameters. We enable the time-level random effects to follow an AR(1) autoregressive process. We assume the covariates to exhibit time-level random effects, but not constant (fixed) effects or unit-level random effects. Our pre-specification involves setting up an MCMC model with 15,000 iterations, including a 5,000 iteration burn-in phase to ensure stability before recording results. LASSO regularization, directly applied from Pang et al. (2021), is used across various model components: constant coefficients (`xlasso`), unit-level random coefficients (`zlasso`), time-level coefficients (`alasso`), and factor loadings (`flasso`), all set to 1 for enabling shrinkage. Hyper-prior parameters for these components are set to diffuse values (0.001) to indicate broad, non-informative priors, supporting a flexible approach to regularization. This setup aims to balance computational efficiency with accuracy and interpretability of the model, leveraging LASSO for sparsity and improved prediction accuracy. We obtain the empirical posterior distribution for counterfactual West Germany, represented by the dashed red line in Figure 5.1. Its 95% CI is also computed and visualized in Figure 5.1.

5.1.4 Counterfactual Estimation

With two replication models and our Bayesian causal MC model being successfully implemented in the German reunification case, we have produced three distinct counterfactual estimations for West Germany. In Figure 5.1, to verify the goodness-of-fit, we clearly observe that the counterfactual GDP per capita (USD) for West Germany across all three models matches well in the pre-treatment periods (before 1990). From a quasi-experimental design perspective, we claim that all three models have effectively removed confounders, allowing us to further derive a causal relationship.

For the post-treatment periods (after 1990), we note that the counterfactual GDP per capita (USD) for West Germany in all three models has grown more rapidly than that of the observed West Germany. This suggests that the German reunification has indeed reduced the GDP per capita for West Germany. Our Bayesian causal MC model, depicted with dashed red lines and pink shaded areas for the 95% CI, appears to be more extreme and uncertain in its predictions. This might be attributable to the choice of a not-so-accurate latent factor specification (e.g., $L = 8$). Overall, all three models have produced

similar findings and align well with each other. We will take a closer look in Section 5.1.5 and discuss any evidence of treatment effects.

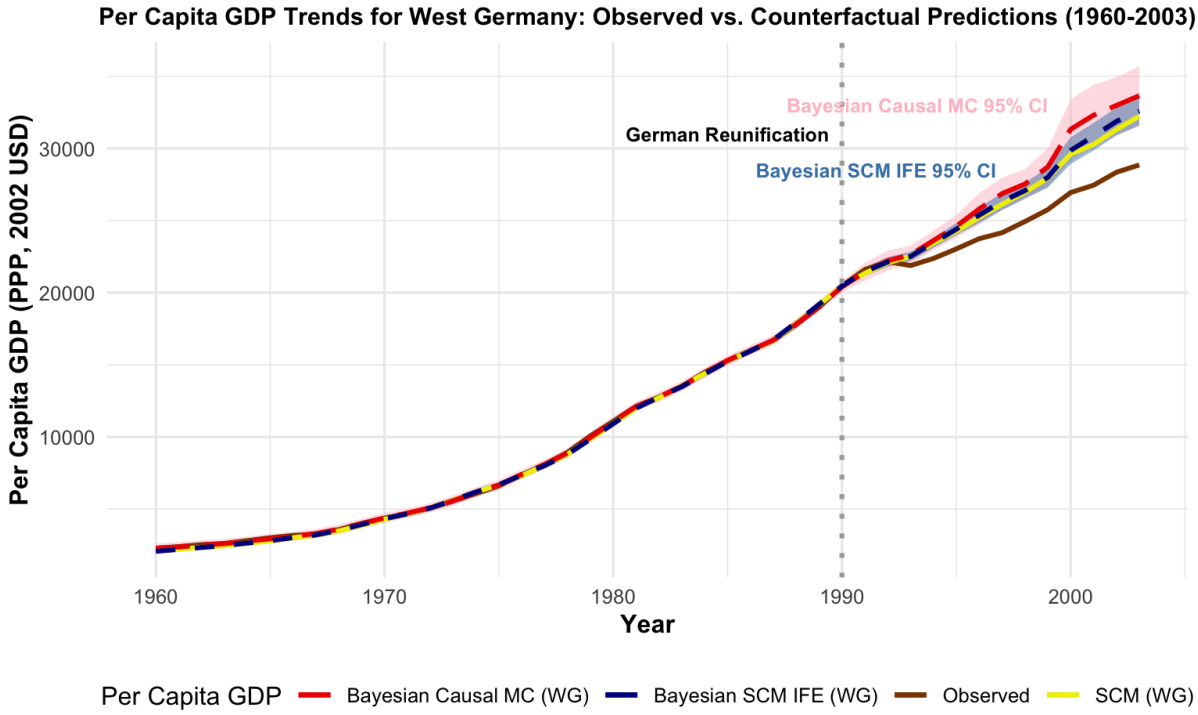

Figure 5.1: Trends in Per Capita GDP: West Germany under Counterfactual Predictions

5.1.5 Evidence of Treatment Effects

We draw inferences regarding the treatment effects of German reunification from Figure 5.2 in a time-series manner. Before reunification, we observe absolutely no treatment effects, as there is no treatment or intervention happening, which aligns with our assumptions. After immediate reunification and lasting for less than three years (namely, 1990 to 1993), we observe that the counterfactual West Germany, had the reunification not occurred, exhibits a positive treatment effect, with a local peak around 1991. This temporal effect suggests that if West Germany had not reunified with East Germany, the GDP per capita (USD) for West Germany might have experienced a temporal increase. However, such temporal effects quickly fade away and even transform into long-term negative treatment effects. This implies that after reunification, the GDP per capita (USD) for West Germany has declined compared to a scenario where West Germany did not undergo reunification.

All three models have provided similar evidence, indicating that there is indeed a negative treatment effect. Both Pinkney's (2021) Bayesian SCM IFE and our Bayesian causal MC model have clearly indicated a negative treatment effect after 1993, as the 95% CI does not include 0. For Abadie et al.'s (2014) SCM, making a confident statement is

challenging without knowing the uncertainty ranges. However, given very similar outcome predictions after reunification compared to the two Bayesian models, we would consider the SCM to also suggest the existence of treatment effects. In particular, we notice that our Bayesian causal MC model exhibits a much wider uncertainty range than Pinkney's (2021) Bayesian SCM IFE; it also demonstrates an overall more significant treatment effect (evidenced by a steeper decline) compared to the Bayesian SCM IFE. This may imply that incorporating time-varying coefficients results in a more uncertain posterior distribution, thereby enhancing confidence that our model more accurately represents the real counterfactual scenario compared to Abadie et al.'s (2014) SCM and Pinkney's (2021) Bayesian SCM IFE.

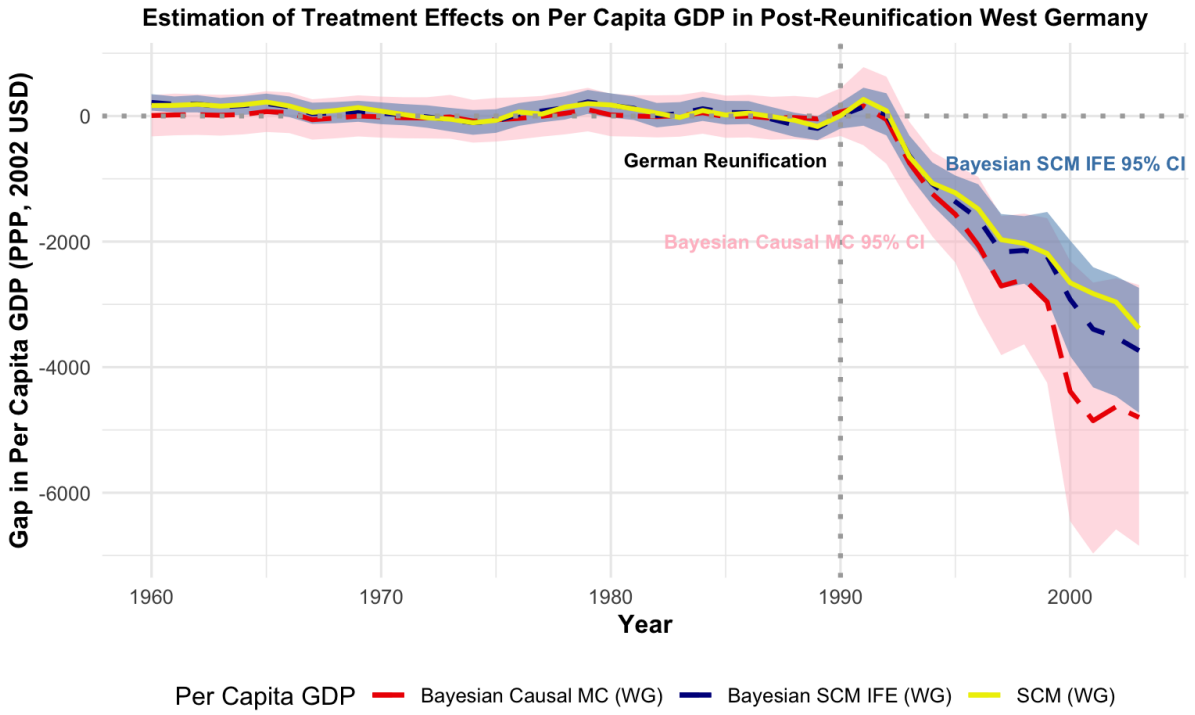

Figure 5.2: Estimated Treatment Effects in Per Capita GDP for West Germany

Following Definition 4, we compute the ATT for West Germany for a duration of $\tau = 13$ years (after 1990). The corresponding ATT values are presented in Table 5.3, where we also include the ATT values at both lower (2.5%) and upper (97.5%) credible intervals. We observe similar patterns for the ATT values, with our Bayesian causal MC model producing a stronger treatment effect in magnitude compared to the other two methods. The overall average uncertainty range is also higher than that of Pinkney's (2021) model.

5.1.6 Effectiveness

To conclude our discussion on this empirical application regarding German reunification, the key takeaway is that our Bayesian causal MC model has demonstrated the capability

Table 5.3: Average Treatment Effect on the Treated West Germany

Model	**ATT** δ_τ
SCM (50%)	-1699.7
Bayesian SCM IFE (2.5%)	-2498.2
Bayesian SCM IFE (50%)	-1897.6
Bayesian SCM IFE (97.5%)	-1279.8
Bayesian Causal MC (2.5%)	-3213.3
Bayesian Causal MC (50%)	-2012.4
Bayesian Causal MC (97.5%)	-809.7

to more efficiently uncover the treatment effect, offering a more reliable CI compared to other SCM extensions. The flexibility of our proposed model enables us to leverage detailed pre-specifications on prior selections and to set reasonable effects on covariates (whether they be constant effects, unit-level random effects, or time-level random effects). This approach allows us to utilize the LFM component to establish the *conditional ignorability* assumption, accurately identify the causal relationship, and derive the correct treatment effect. To further illustrate the superior performance of our Bayesian causal MC model, we extend our analysis beyond the single outcome problem of German reunification to address a more complex panel data structure in CRM. In the challenging context of large N, small T, and multiple P scenario, neither Abadie et al.'s (2010) SCM nor Pinkney's (2021) Bayesian SCM IFE can directly resolve the issue, as the data characteristics could immediately prove problematic in those circumstances.

5.2 CRM Panel Data

We begin our analysis by detailing the raw format of the data and outlining a method to efficiently transform any non-structural data into a panel data framework. This study leverages a comprehensive customer-level database of gift card purchases and redemptions from a U.S. hospitality startup. Our primary data source consists of a collection of raw `JSON` files, including cross-sectional data on brand tags, projects, and users' information, as well as time-series data on redemption history and revenue views spanning the years 2021, 2022, and the first month of 2023 (covering a duration of 25 months), extracted from the firm's CRM database.

5.2.1 Longitudinal Data Analysis

Before performing any data manipulation and wrangling, we first explore what the data looks like. Although we have a total of 9 `JSON` files, for easier access to our data, we initially convert all of them into tabular formats in `SQL`. We then import the results into `R` and merge multiple tables together using the `full_join` function from the `dplyr`

package. We present some significant findings from the data descriptions and include a link at the beginning of Appendix C for readers interested in familiarizing themselves with additional aspects of the data.

Figure 5.3 provides a preview of all variables in our dataset after a series of data manipulations and wrangling. Figure 5.3 displays a total of 34 variables, each with different degrees of missingness, marked by a red dashed line at 75% implemented arbitrarily. This is mainly due to two reasons. First, during the tabular merging process, we set some common variables, in this case, `user_id` (a unique label for various users), `project_id` (a unique label for various projects offered by a restaurant), `created_at` (the date a project was consumed by a user, in YYYY-MM-DD format), and `account_created_at` (the date of user registration on the platform, in YYYY-MM-DD format), as the joining keys. The four columns at the bottom have no missing data for this reason. However, in the merging process, several other variables may not contain a particular row of these common keys, thereby causing missingness. The second reason is more straightforward: the provided raw `JSON` files initially have different degrees of missingness.

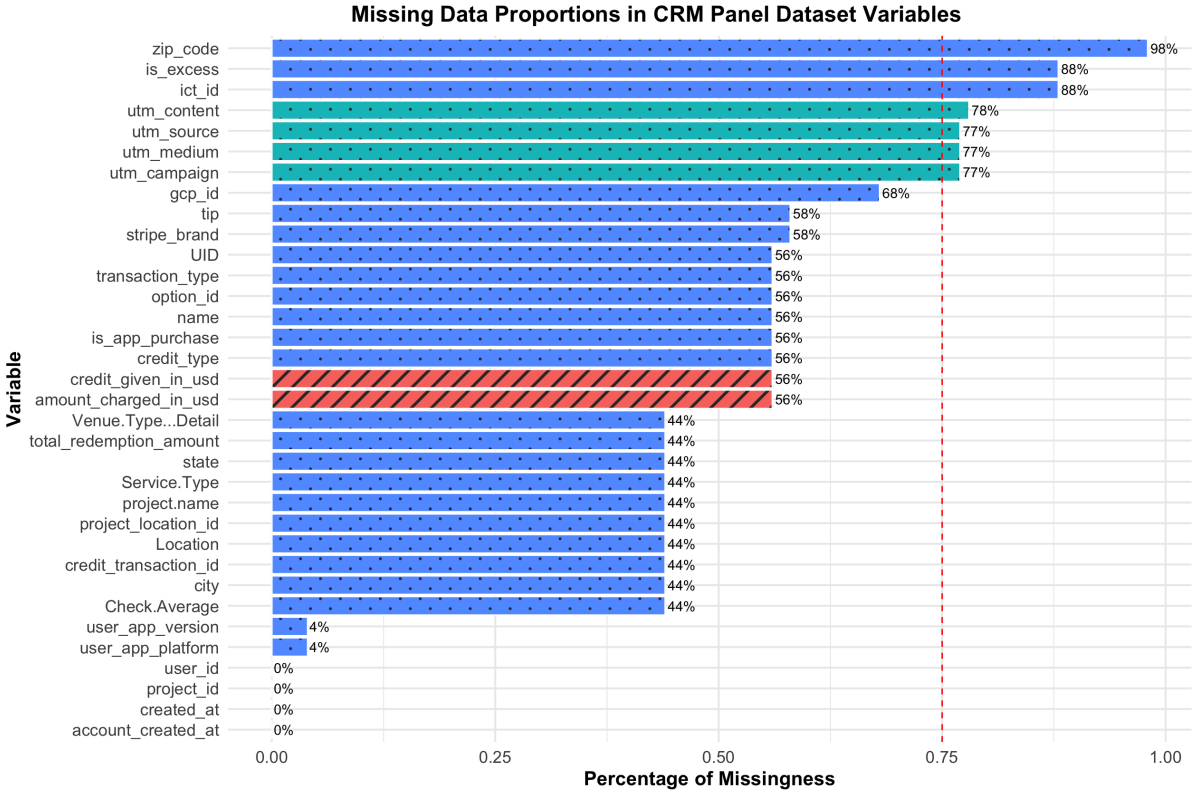

Figure 5.3: Missing Data Proportions in CRM Panel Dataset Variables

To transform the merged data into a panel (longitudinal) data frame, we initially create a unique identifier by combining `user_id` and `project_id`. We explore different time granularities (e.g., daily, weekly, bi-weekly, monthly, and quarterly) for the time index. Considering the concerns regarding data sparsity and the level of granular clarity we aim

to achieve, we opt for a bi-weekly time index as a balanced choice, constructed from `created_at`. This data wrangling process is efficiently facilitated by the `pdata.frame` function from the `plm` package (Croissant and Millo 2018). With the panel data frame now featuring bi-weekly granularity, we proceed to examine other variables and articulate our research statement below.

5.2.2 Model-Free Evidence

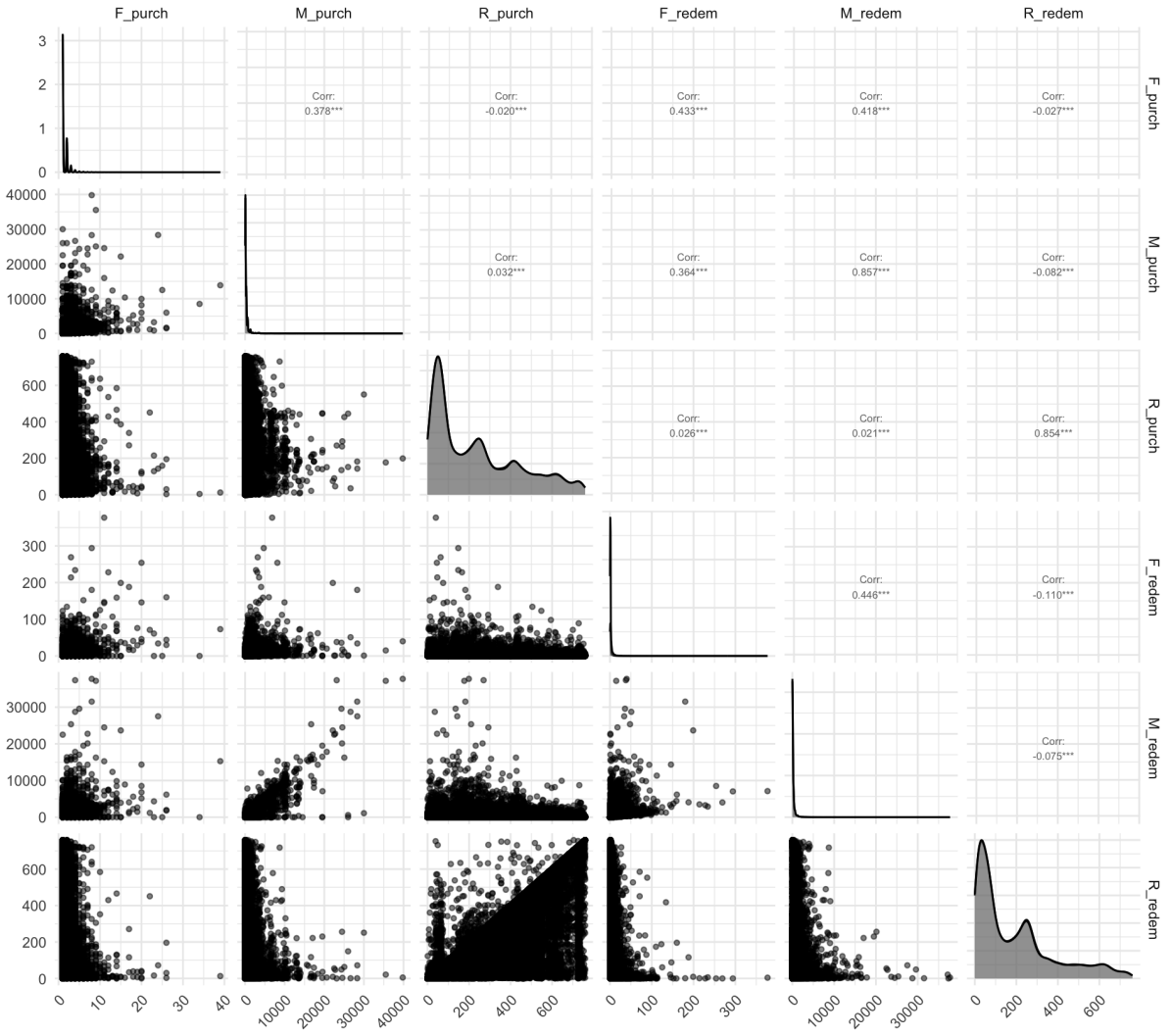

Figure 5.4: Correlation Analysis for RFM of Purchases and Redemptions

Among the remaining 30 variables depicted in Figure 5.3, we identify two particularly important outcome variables for our study: `amount_charged_in_usd` (the amount of purchase per user/project) and `total_redemption_amount` (the amount of redemption per user/project). In Figure 5.4, we visualize the distribution of the frequency (F) of purchases and redemptions made by each user per project at their most recent record in columns 1 and 4, respectively. Similarly, the distribution of the monetary (M) values of purchases and redemptions is presented in columns 2 and 5 of Figure 5.4. We observe an

extremely long right tail for these four dimensions, which further suggests the sparsity of data due to the accumulation of one-time purchasers/redeemers, indicative of the *cold start* problem.

We further apply feature engineering techniques to expand our CRM panel data frame. Considering the two important monetary (M) value outcomes of purchase and redemption, we can apply the `group_by` and `summarize` functions in `dplyr` to quickly gain the frequency (F) of purchases and redemptions. We can also apply the same functions to obtain the recency (R), which can be computed by the difference between the current time index and the end of the data recording time index. We convert the bi-weekly indexes that our data inputs into the unit of days in Figure 5.4.

The distribution of recency differences for both purchases and redemptions exhibits similar peaks and troughs, indicating the influence of seasonal promotions or holiday effects. There is a notable increase in users making purchases and redemptions in recent days, which could be attributed either to the growing popularity of the platform or a significantly impactful holiday effect (e.g., Christmas and New Year, as inferred from the detailed date information in our data). In Figure 5.4, we present another correlation plot between the recency of both dimensions. Without any significantly extreme outliers, we observe that the correlation is significantly positive ($r = 0.854$).

5.2.3 RFM Framework

Through model-free evidence, we have identified and derived our six outcomes of interest: the recency (R), frequency (F), and monetary (M) value of both purchases and redemptions. This constitutes the RFM framework for addressing our problem of interest. This approach, informed by marketing domain knowledge, is adaptable across a broad spectrum of disciplines. It offers a method to extract multiple dimensions from a single P problem. As previously mentioned, accommodating multiple P's helps offset the limitations of short time series (small T) at the unit level.

We then perform transformations on our six-outcome P's. In Figure 5.5, we visualize the distributions of RFM across purchases and redemptions with no transformation (column 1), square root transformation (column 2), and log transformation (column 3). The log transformation proves to be more effective in rendering the distribution of each variable more symmetric, especially for the frequency (F) and monetary (M) value variables. Consequently, we apply a log transformation to every outcome variable.

After implementing the log transformation, we proceed with an adjusted min-max normalization

$$\hat{v} = \frac{10(v - v_{\min})}{v_{\max} - v_{\min}}$$

to expand the range from $[0, 1]$ to $[0, 10]$, where v represents any value before the min-max normalization is applied. This adjusted min-max normalization process is reversible, as the function is surjective.

Figure 5.5: Distribution of Multi-Outcome P's with Transformations

Section 5.2.3 concludes our discussion on the final $\mathbf{Y}$ component, comprising a list of six outcome variables. To gain a comprehensive understanding of how the CRM panel data integrates with our Bayesian causal MC model, it is important to address two other types of variables, $\mathbf{W}$ and $\mathbf{X}$. These will be addressed in the subsequent subsections (Sections 5.2.4 and 5.2.5).

5.2.4 Treatment Staggered Adoption

Recall the German reunification study, where we have 16 OECD countries as control units and West Germany as the single treated unit. In this CRM panel data structure, all customers are essentially in a "treated" status. Rather than defining real treatment through promotional activities (despite having such information, labeled by `utm_`) or other clear interventions, we conceptualize treatment here as a binary status, indicating whether a data entry at a block is missing ($w_{it} = 1$, in treatment) or observed ($w_{it} = 0$, in control).

Referring back to Table 1.1, the customer-level transaction history exhibits a pattern resembling a sparse matrix. Although our Bayesian causal MC model can accommodate this back-and-forth switch, we adopt an assumption by focusing only on a cumulative outcome measure. Consequently, our six dimensions of outcome variables follow a non-decreasing pattern. Should any missing entry occur at period t, we sum up its previous entries and allocate this cumulative value to period $t + 1$. Under this assumption, our "treatment" status exhibits a staggered adoption pattern (refer to Definition 1).

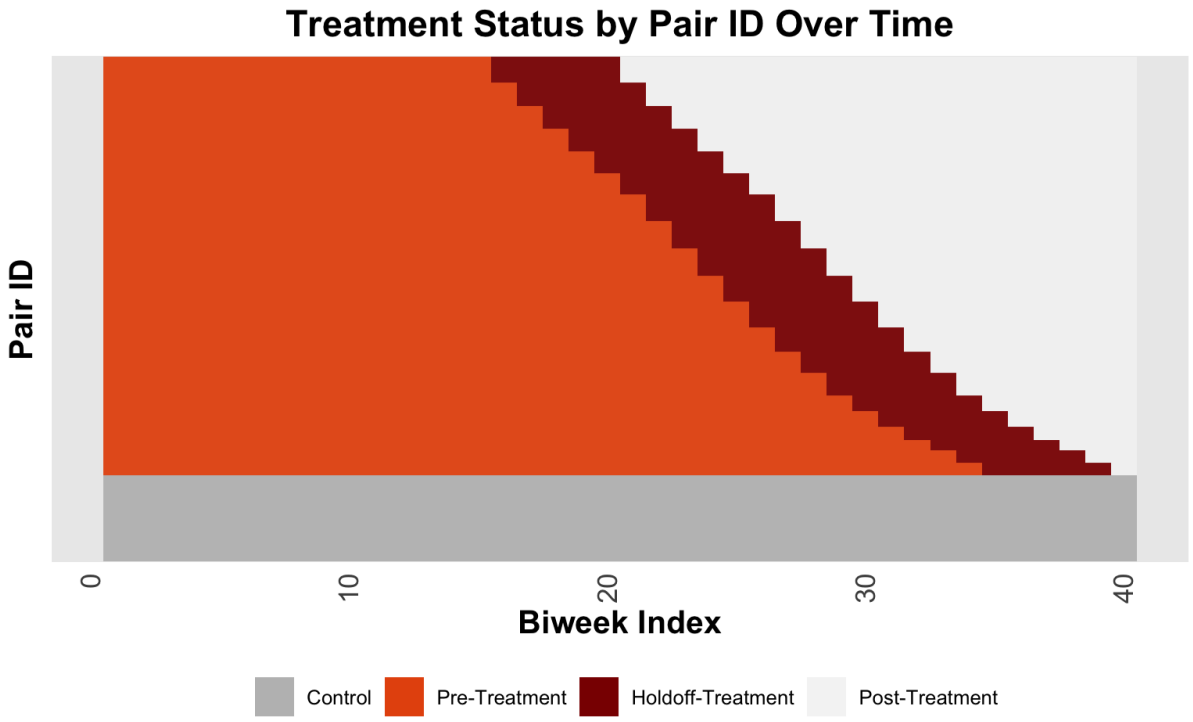

Figure 5.6: Treatment Status by Pair ID Over Time

In Figure 5.6, we plot a selection of users who make at least two purchases throughout their entire CLV, representing a third of all users. Including one-time purchasers would significantly compromise the dataset due to the presence of flat values across their entire CLV. Given that the data span 25 months, equivalent to 55 bi-weeks, Figure 5.6 displays

only bi-week indices up to 40. This means for any customers who joined our platform very early, a portion of their data is obscured, allowing them to have at most 40 observed periods (grey blocks, indicating control status). For those customers whose CLV length is originally under 40, we withhold 5 periods and exclude them from our algorithms. The hidden 5 periods of observed data (dark red blocks, indicating treated missing status) are stored separately so that we can later evaluate our counterfactual predictions against them and assess the treatment effect. The orange blocks denote those treated customers in observed status.

To understand why the idea of SCM is applicable to this study, imagine thousands of West Germanys (large N) paired with hundreds of OECD countries in control. Each of these West Germanys possesses an additional 5 periods of observed data. Our Bayesian causal MC model is designed to effectively compute the counterfactual estimation for both the dark red and white blocks, imputing the missingness with these values. While we cannot directly assess how our counterfactual estimation diverges from the multiple outcome data at those white blocks (purely missing without any pre-holdoff data), we can gauge the overall performance of the model by comparing each customer's 5 holdoff[2] periods' observed outcomes with our imputed counterfactual outcomes.

Therefore, Figure 5.6 illustrates the structure of staggered adoption treatment, which presents a more complex structure compared to Figure 4.1. Additionally, the y-axis is labeled as `Pair ID`, which combines both `user_id` and `project_id`.

5.2.5 Covariates

We have now completed the discussion on $\mathbf{Y}$ and $\mathbf{W}$. The remaining component in our data is $\mathbf{X}$. In fact, the covariate side of our model is not necessary. The Bayesian causal MC model can be implemented in this CRM panel data study directly without inputting any covariates. However, we still present some selections here to help readers gain a better understanding of our data.

Our covariates mainly come from two sources. One source is depicted in Figure 5.3, where we classify the remainder of non-outcome, non-treatment variables as covariates. The other source involves utilizing Yelp's Fusion API to web-scrape additional data related to projects/restaurants.

However, several variables are in text format, e.g., `utm_` (promotion-related covariates, including promotion content, source, medium, and campaign). We adopt two approaches. If the character types of data are countable and small in size (rule of thumb: ≤ 10), we convert such variables into categorical levels and assign different factors, similar to one-hot

[2] Also known as the holdout period, i.e. the period during which data are withheld for testing a model.

encoding. If the text data contains significantly varied texts, such as promotion contents which are unique to each user, we employ natural language processing (NLP) models. This includes NLP-based FastText word embeddings (Joulin et al. 2016) with principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and iterative imputer (`sklearn.impute.IterativeImputer`), as well as Clark et al.'s (2019) efficiently learning an encoder that classifies token replacements accurately (ELECTRA, a faster transformer model than Devlin's (2018) bidirectional encoder representations from transformers, BERT model), to convert them into vectors. The detailed implementation of Yelp's Fusion API, tag embeddings based on FastText (Joulin et al. 2016), and ELECTRA (Clark et al. 2019) can be found in Appendix C.5, C.6, and C.7.

In Table 5.4, we present the description of the enriched attribute that we obtain from applying Yelp's Fusion API. These additionally retrieved data are eventually merged with our dataset shown in Figure 5.3.

Table 5.4: Description of Yelp Enriched Data Attributes

Attribute	**Description**
yelp_tag	The categories the business falls under (e.g., "Restaurant", "Cafe").
rating	The average Yelp rating of the business.
review_num	The total number of reviews the business has received on Yelp.
price_level	The price level of the business, represented by number of dollar signs (e.g., $$$).
transactions	Types of transactions the business offers (e.g., "pickup", "delivery").
yelp_url	The Yelp URL directing to the business's Yelp page that allows us to check manually.

5.3 Implementation in CRM Panel Data

Similar to Section 5.1.3, we now implement our proposed Bayesian causal MC model in CRM panel data as discussed in Section 5.2. The algorithms of `BCMC` are implemented within the Bayesian causal MC framework, accommodating independent multiple P's under two scenarios: one without covariates and one with covariates. We implement both examples separately below and compare the impact of including covariates.

Our defined `BCMC` function (see Appendix C.3) operates on a given dataset with specified unit and time indices (`index`, in this case, pair ID), assessing the impact of a treatment variable (`Dname`, in this case, a binary treatment status `D_holdoff` with $w_{it} = 1$ indicating treatment and $w_{it} = 0$ indicating control) on a vector of outcomes (`Yname_vector`, in this case, a vector of six outcomes including the recency, frequency, and monetary value of purchase and redemption). The parameter `re` specifies the structure of the random effects, incorporating two-way random effects selected from pair ID and bi-weekly index.

The parameter `ar1` indicates whether the time-level random effects adhere to an AR(1) process, and we assume that the time-level random effects follow an AR(1) process, rather than being multi-level and independent. The parameter `r` denotes the number of latent factors in the model, for which we specify $L = 8$ in this study. The MCMC settings are determined by the number of iterations (`niter`, where we set $15,000$) and the number of burn-in steps (`burn`, where we set $5,000$). Regularization is applied to the coefficients through LASSO, controlled by `xlasso`, `zlasso`, `alasso`, and `flasso`, where we pick default values of 1, with hyper-prior parameters set to diffuse priors (0.001). In Figure 5.7, the counterfactual multi-outcomes are depicted with red lines, and their 95% CIs are illustrated by shaded pink regions.

`BCMC` also allows for the inclusion of covariates with fixed effects (`Xname`), unit-level random effects (`Zname`), and time-level random effects (`Aname`), relying on Pang et al.'s (2021) Bayesian LFM. Since we follow the assumption that all outcome measures rely on cumulative distributions, we should not perceive any temporal effects, therefore setting $\mathbf{A} = \varnothing$. Hence, we pick essential covariates that we believe will potentially affect six outcomes from Figure 5.3 and Table 5.4, setting them as parts of $\mathbf{Z}$ and $\mathbf{X}$. In Figure 5.8, the counterfactual multi-outcomes are similarly depicted in red lines with their 95% CIs shown by shaded pink regions.

5.3.1 Counterfactual Estimation

From both Figure 5.7 and Figure 5.8, the observed cumulative functions for each of the six dimensions are recorded up until the 0 index in `Relative Time`, where the computed ATT across staggered adoption patterns finds an average treatment adoption time in the algorithms. For time indices from 0 to 20, we depict them as missing data, shown in dashed brown lines. The 5 periods of holdoff are depicted by the region between two dashed blue lines. For time indices from -5 to 0, we can compare the counterfactual outcome value with the observed holdoff counterpart. We present the treatment effect over time on the y-axis of Figure 5.9. While our Bayesian causal MC model continues predicting the missing components after 5 periods, we observe them deviating from the pre-assumptions quickly. For example, we notice that the monetary dimensions for both purchase and redemption, as well as the recency of purchase without covariates, begin to decline at certain time indices after the holdoff period. This violates the assumption of cumulative distributions across all six outcomes. We also observe the uncertainty range increases at later periods. The computational power of our Bayesian causal MC model is therefore optimized within a few future steps, given the case that we deal with a short T problem. By withholding periods more than 5, we can test the model's accuracy over a longer timeline.

From Figure 5.7 and Figure 5.8, we observe that on average, a customer takes approx-

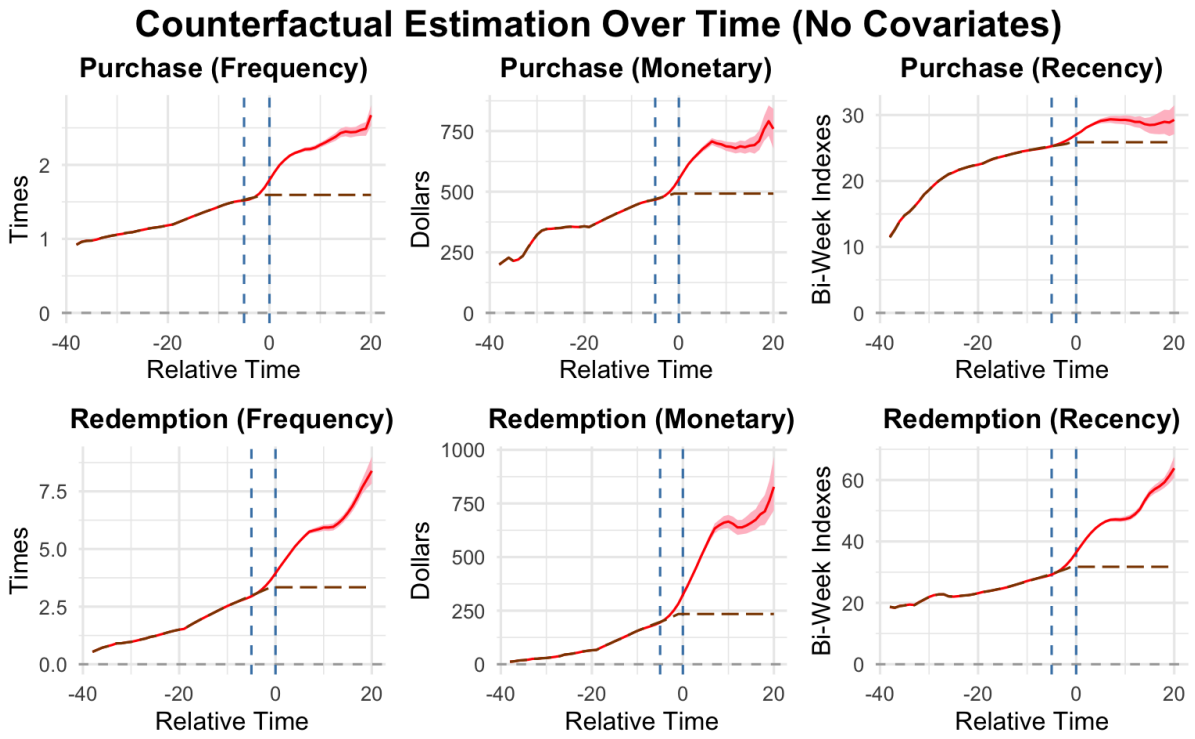

Figure 5.7: Counterfactual Estimation Over Time (No Covariates)

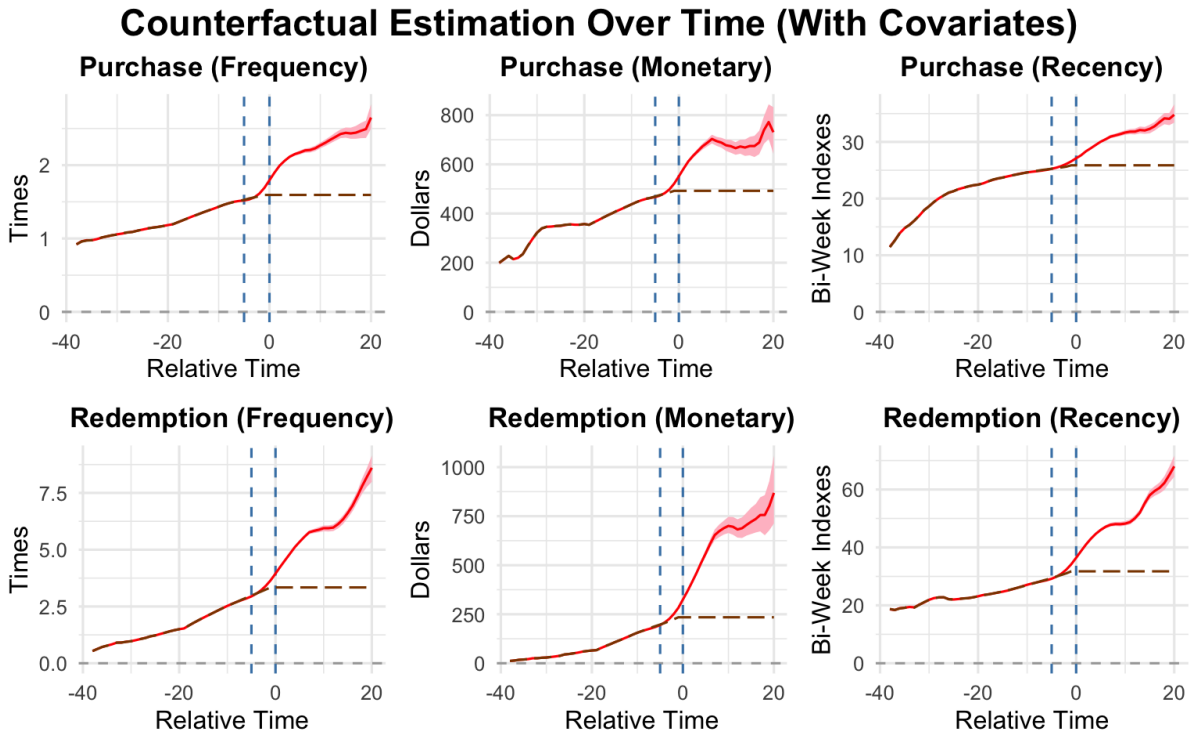

Figure 5.8: Counterfactual Estimation Over Time (With Covariates)

imately 40 units of time index to make another purchase, whereas only 5 units of time index are required for another redemption. The monetary columns suggest similar findings, where on average, an additional purchase is around $250, while an additional redemption amounts to less than $100. Analyzing the third columns of recency, we note that on average, an additional purchase occurs about 15 bi-weeks (210 days), whereas an additional redemption occurs in less than 10 bi-weeks (140 days). The advantage of multi-outcome modeling ensures that data prediction aligns well. With the posterior distribution generated for all sample users on average, we can also examine each individual user and see how their purchase and redemption might unfold in the next few periods (for instance, 5).

By comparing Figure 5.7 and Figure 5.8, we do not observe much difference when incorporating additional covariates as predictors. This implies that our Bayesian causal MC model can perform well as long as the number of latent factors is roughly specified correctly. However, the addition of covariates reduces the uncertainty levels in several dimensions, including the recency dimension for both purchases and redemptions.

5.3.2 Model Performance Evaluation

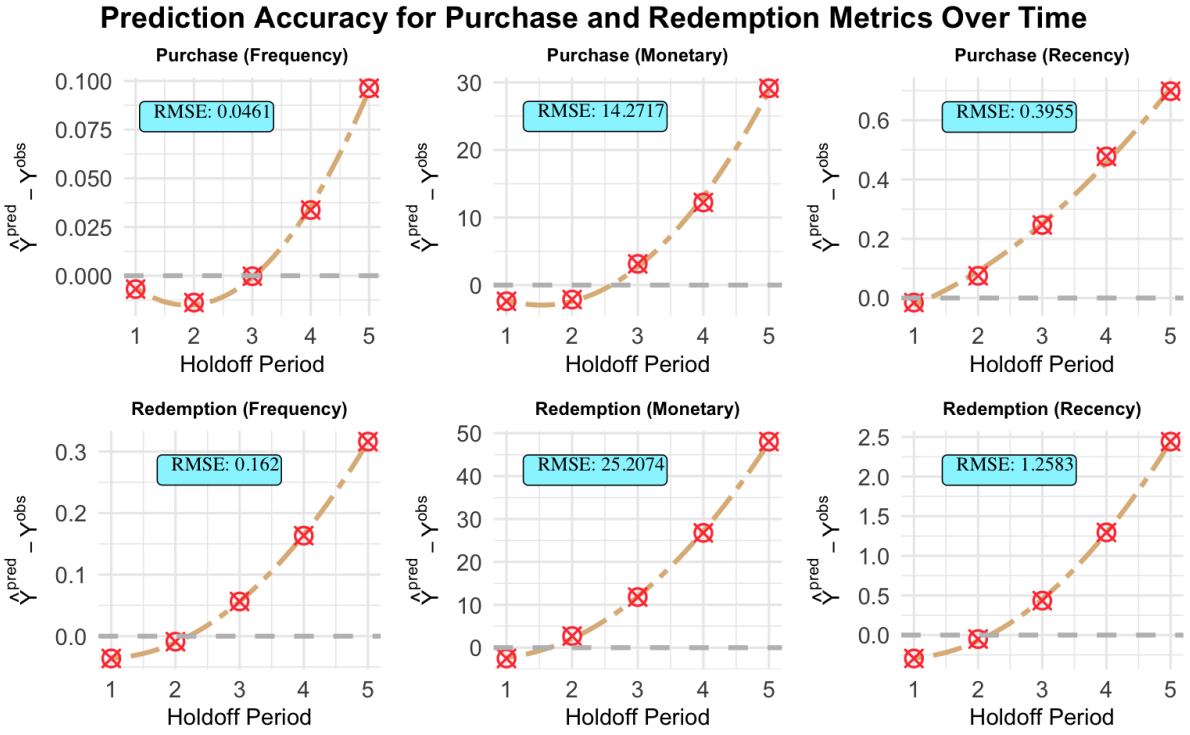

Figure 5.9: Prediction Accuracy for Purchase and Redemption Metrics Over Time

We compute the treatment effect for the 5 holdoff periods, as we have observed data serving as a baseline for comparison. Since the treatment concerns time and status change from missing to observed, we should, in theory, observe no treatment effect, implying that

$\hat{Y}_{\text{pred}} - Y_{\text{obs}} = 0$. In Figure 5.9, we note that in the frequency dimension of purchase and redemption, the counterfactual estimation performs best in the subsequent 3 periods and 2 periods, respectively. Overall, for frequency in the next 3 periods, the prediction is sufficiently accurate as the y-axis value is close to 0. Similar findings can be observed in the monetary value dimension and the recency dimension, where any counterfactual predictions within 2 periods are close enough to 0.

The results suggest that our Bayesian causal MC model effectively tackles the large N, small T, and multiple P challenge by delivering accurate predictions for up to 2 or 3 forthcoming periods, although its predictive strength diminishes over the long term, evidenced by a significantly wider uncertainty range. From an inferential perspective, the Bayesian causal MC model excels at accurately predicting the counterfactual counterpart, transforming missing data into imputed values. This model's ability to predict multiple outcomes accurately over the next few periods could be leveraged to develop a causal recommender system.

Furthermore, an examination of the RMSE values across the six dimensions reveals that for frequency and recency, the relatively small RMSEs signify the model's precision and low error rate. However, the monetary value dimension exhibits a larger error margin, attributable to the inherent data variability in monetary amounts compared to the more stable frequency and recency metrics measured in bi-weeks. With this, we conclude our discussion on CRM panel data and the effectiveness of our proposed Bayesian causal MC model.

6

Discussion

Due to the length limit of the honors thesis, we discuss the conditions under which our Bayesian causal MC model excels and where it faces limitations. We highlight 13 advantages and the flexibility of our model compared to Abadie et al.'s (2014) SCM in the following colored framed text box, with their model serving as a baseline in this study.

However, we acknowledge that our model differs in purpose from the idea of SCM. Reiterating a point made by Pang et al. (2021), we recognize that both Abadie et al.'s (2014) SCM and Ashenfelter and Card's (1985) DiD are design-based models with more transparent identification assumptions, which gain wider acceptance among researchers due to their relatively weak assumptions. While our model critiques Abadie et al.'s (2014) SCM for its constraints on weights (summing-to-one and non-negativity) and Ashenfelter and Card's (1985) DiD for its uniform weights constraint, we value the concept of their directly interpretable weights. Conversely, our Bayesian causal MC model, which adopts a model-based approach, often does not provide an intuitive interpretation of weights.

In particular, our Bayesian causal MC model with independent multiple P's faces several shortcomings. These are noted in another colored framed text box below.

Advantages of Bayesian Causal MC Method Over SCM

1. Relaxes constraint of non-negativity of weights.
2. Relaxes constraint of summing-to-one of weights.
3. Relaxes constraint of no intercept.
4. Accommodates multiple treated units.
5. Accommodates multiple outcomes.
6. Incorporates time-varying covariates.
7. Allows time-specific coefficients.
8. Allows unit-specific coefficients.
9. Allows model averaging.
10. Infers average treatment effects (ATE, ATT).
11. Infers individual treatment effects (ITE).
12. Incorporates interpretable Bayesian uncertainty measures.
13. Performs well when $N \gg T$.

This list details the comparative benefits of using the Bayesian causal MC model with independent multiple P's over Abadie et al.'s (2010) standard SCM, highlighting advancements in generalizability, flexibility, modeling capabilities, and inference.

Limitations of Bayesian Causal MC Method

1. Multiple P is not concurrently resolved.
2. Panel data characteristics have to follow staggered adoption assumption.
3. Bayesian approach is computationally expensive for extremely large N.
4. Number of pre-treatment periods for treated units needs to satisfy $T_0 > 20$.
5. Weights are not directly interpretable.
6. There may be a violation of potential SUTVA assumption.

This list identifies the main limitations when applying the Bayesian causal MC model with independent multiple P's. The constraints listed above should be carefully considered in practice.

However, the initial four limitations can be readily addressed by adopting Bayesian causal MC with concurrent multiple P's, as demonstrated in Model 4 (see its implementation in Appendix C.4). Specifically, Model 4 accommodates concurrent multiple P's by incorporating them into a $(L \times L)$ scaling matrix $\mathbf{\Sigma}^p$.

Notably, Bayesian causal MC with concurrent multiple P's does not rely on the assumption of staggered adoption, effectively overcoming the second limitation. Model 4 utilizes a factorization method to model counterfactuals directly via latent factors and scaling matrices, independent of any specific sequence of treatment adoption. This allows for the generation of each unit's counterfactual outcomes independently from the treatment timings across units, thereby eliminating the need for assumptions regarding the temporal sequence of treatment exposure. Through the interplay of the matrices $\mathbf{f}$, $\mathbf{\Sigma}^p$, and $\mathbf{\Gamma}$, Model 4 enables the simultaneous estimation of counterfactuals across multiple outcome dimensions p, sidestepping the constraints imposed by staggered treatment patterns.

The third limitation, concerning the computational expense associated with large datasets, is efficiently addressed by our `NumPyro` program, which significantly enhances processing efficiency. Built upon `JAX`, `NumPyro` supports automatic differentiation and GPU acceleration, enabling substantial Bayesian computation speedups through vectorized operations and parallel processing. This advancement effectively reduces the computational load typical of large N Bayesian models.

Regarding the fourth limitation, the rationale behind our approach stems from addressing the challenge posed by short T and large N, particularly the difficulty in accurately identifying γ_i. When individual-level time series data are limited, γ_i may remain undetermined. We claim that by generating multiple concurrent outputs, we inherently apply constraints through the functional form of $\mathbf{Y}^p$. This is because the same γ_i must optimally apply to several outputs for the same donor i, thereby improving identification.

The last two limitations present more complex challenges that may not be readily addressed by our Bayesian causal MC model. According to Pang et al. (2021), addressing the fifth limitation is contingent upon the weights carrying specific policy implications. As for the sixth limitation, we encounter a distinct challenge in the presence of policy diffusion or spillover effects, as discussed by Athey and Imbens (2018). These open questions are left for future researchers to explore.

Acronyms

1. **ALS**: **A**lternating **L**east **S**quares
2. **API**: **A**pplication **P**rogramming **I**nterface
3. **AR**: **A**uto**r**egressive Model
4. **ATE**: **A**verage **T**reatment **E**ffect
5. **ATT**: **A**verage **T**reatment Effect on the **T**reated
6. **BERT**: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
7. **B2C**: **B**usiness-**To**-**C**onsumer
8. **BvM**: **B**ernstein-**v**on **M**ises
9. **CI**: **C**redible **I**nterval
10. **CLV**: **C**ustomer **L**ifetime **V**alue
11. **CRM**: **C**ustomer **R**elationship **M**anagement
12. **DiD**: **D**ifference-**i**n-**D**ifferences
13. **DGP**: **D**ata **G**enerating **P**rocess
14. **ELECTRA**: **E**fficiently **L**earning an **E**ncoder that **C**lassifies **T**oken **R**eplacements **A**ccurately
15. **FE**: **F**ixed **E**ffects Model
16. **HTE**: **H**eterogeneous **T**reatment **E**ffect
17. **GDP**: **G**ross **D**omestic **P**roduct
18. **GPU**: **G**raphics **P**rocessing **U**nit
19. **HMC**: **H**amiltonian **M**onte **C**arlo
20. **IFE**: **I**nteractive **F**ixed **E**ffects Model

21. **ITE**: **I**ndividual **T**reatment **E**ffect
22. **LASSO**: **L**east **A**bsolute **S**hrinkage and **S**election **O**perator
23. **LFM**: **L**atent **F**actor **M**odel
24. **MC**: **M**atrix **C**ompletion
25. **MCMC**: **M**arkov **C**hain **M**onte **C**arlo
26. **MF**: **M**atrix **F**actorization
27. **MICE**: **M**ultivariate **I**mputation by **C**hained **E**quations
28. **MLE**: **M**aximum **L**ikelihood **E**stimator
29. **MNAR**: **M**issing **N**ot **A**t **R**andom
30. **NBD**: **N**egative **B**inomial **D**istribution
31. **NLP**: **N**atural **L**anguage **P**rocessing
32. **NUTS**: **N**o-**U**-**T**urn **S**ampler
33. **OECD**: **O**rganization for **E**conomic **C**o-operation and **D**evelopment
34. **PCA**: **P**rincipal **C**omponent **A**nalysis
35. **PPP**: **P**urchasing **P**ower **P**arity
36. **PSM**: **P**ropensity **S**core **M**atching
37. **RFM**: **R**ecency, **F**requency, and **M**onetary Value Model
38. **RMSE**: **R**oot **M**ean **S**quared **E**rror
39. **SCM**: **S**ynthetic **C**ontrol **M**ethod
40. **SGD**: **S**tochastic **G**radient **D**escent
41. **SUTVA**: **S**table **U**nit **T**reatment **V**alues **A**ssumption
42. **SVD**: **S**ingular **V**alue **D**ecomposition
43. **t-SNE**: **t**-distributed **S**tochastic **N**eighbor **E**mbedding
44. **USD**: **U**nited **S**tates **D**ollar

Bibliography

Abadie A, Diamond A, Hainmueller J (2010) Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* 105(490):493–505.

Abadie A, Diamond A, Hainmueller J (2011) Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. *Journal of Statistical Software* 42(13).

Abadie A, Diamond A, Hainmueller J (2014) Comparative Politics and the Synthetic Control Method. *American Journal of Political Science* 59(2):495–510.

Abadie A, Gardeazabal J (2003) The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review* 93(1):113–132.

Abadie A, Imbens GW (2006) Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica* 74(1):235–267.

Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S (2021) Synthetic Difference-in-Differences. *American Economic Review* 111(12):4088–4118.

Ascarza E, Netzer O, Hardie BGS (2018) Some Customers Would Rather Leave Without Saying Goodbye. *Marketing Science* 37(1):54–77.

Ashenfelter O, Card D (1985) Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *The Review of Economics and Statistics* 67(4):648.

Athey S, Bayati M, Doudchenko N, Imbens G, Khosravi K (2021) Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association* 116(536):1–15.

Athey S, Blei D, Donnelly R, Ruiz F, Schmidt T (2018) Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data. *AEA Papers and Proceedings* 108:64–67.

Athey S, Imbens GW (2017) The State of Applied Econometrics: Causality and Policy

Evaluation. *Journal of Economic Perspectives* 31(2):3–32.

Athey S, Imbens GW (2022) Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. *Journal of Econometrics* 226(1):62–79.

Bai J (2009) Panel Data Models with Interactive Fixed Effects. *Econometrica* 77(4):1229–1279.

Bai J, Ng S (2021) Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data. arXiv.org. Retrieved https://arxiv.org/abs/1910.06677.

Ben-Michael E, Feller A, Rothstein J (2021) The Augmented Synthetic Control Method. *Journal of the American Statistical Association* 116(536):1789–1803.

Buuren S van, Groothuis-Oudshoorn K (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45(3).

Card D, Krueger A (2000) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment. *American Economic Review* 90(5):1362–1396.

Carvalho C, Masini R, Medeiros MC (2018) ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics* 207(2):352–380.

Clark K, Luong MT, Le QV, Manning CD (2020) ELECTRA: Pre-training Text Encoders as Discriminators Rather than Generators. *ICLR 2020 the Eighth International Conference on Learning Representations*.

Croissant Y, Millo G (2018) Panel Data Econometrics with R. *John Wiley & Sons*.

de Finetti B (1963) Foresight: Its Logical Laws, Its Subjective Sources. *Studies in Subjective Probability*.

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North* 1.

Doudchenko N, Imbens GW (2016) Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *National Bureau of Economic Research*.

Engle RF, Hendry DF, Richard JF (1983) Exogeneity. *Econometrica* 51(2):277.

Fader PS, Hardie BGS (2009) Probability Models for Customer-Base Analysis. *Journal of Interactive Marketing* 23(1):61–69.

Fader PS, Hardie BGS, Shang J (2010) Customer-Base Analysis in a Discrete-Time Non-

contractual Setting. *Marketing Science* 29(6):1086–1108.

Farouni R (2015) Bayesian Factor Analysis. rfarouni.github.io. Retrieved https://rfarouni.github.io/2015-04-26-fa/.

Ferman B, Pinto C (2019) Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity. *The Review of Economics and Statistics* 101(3):452–467.

Funk S (2006) Netflix Update: Try This at Home. sifter.org. Retrieved https://sifter.org/simon/journal/20061211.html.

Gelman A (2005) Analysis of Variance? Why It Is More Important than Ever. *The Annals of Statistics* 33(1):1–53.

Gobillon L, Magnac T (2016) Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls. *The Review of Economics and Statistics* 98(3):535–551.

Goldfarb A, Tucker C, Wang Y (2022) Conducting Research in Marketing with Quasi-Experiments. *Journal of Marketing* 86(3):1–20.

Hahn J, Shi R (2017) Synthetic Control and Inference. *Econometrics* 5(4):52.

Hainmueller J, Hopkins DJ, Yamamoto T (2014) Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis* 22(1):1–30.

Holland PW (1986) Statistics and Causal Inference. *Journal of the American Statistical Association* 81(396):945–960.

Hsiao C, Steve Ching H, Ki Wan S (2011) A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China. *Journal of Applied Econometrics* 27(5):705–740.

Imai K, Kim IS, Wang EH (2021) Matching Methods for Causal Inference with Time-Series Cross-Sectional Data. *American Journal of Political Science* 67(3).

Imbens GW, Rubin DB (2015) Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of Tricks for Efficient Text Classification. ACLWeb:427–431. Retrieved https://aclanthology.org/E17-2068/.

Kim S, Lee C, Gupta S (2020) Bayesian Synthetic Control Methods. *Journal of Marketing Research* 57(5):831–852.

Kirsch W (2018) An Elementary Proof of de Finetti's Theorem. arXiv.org. Retrieved https://arxiv.org/abs/1809.00882.

Koren Y, Bell R, Volinsky C (2009) Matrix Factorization Techniques for Recommender Systems. *Computer* 42(8):30–37.

Li F, Ding P, Fabrizia Mealli (2023) Bayesian Causal inference: a Critical Review. *Philosophical Transactions of the Royal Society A* 381(2247).

Li KT, Van den Bulte C (2022) Augmented Difference-in-Differences. *Marketing Science* 42(4).

Martinez I, Vives-i-Bastida J (2023) Bayesian and Frequentist Inference for Synthetic Controls. arXiv.org. Retrieved (December 2, 2023), https://arxiv.org/abs/2206.01779.

Netzer O, Lattin JM, Srinivasan V (2008) A Hidden Markov Model of Customer Relationship Dynamics. *Marketing Science* 27(2):185–204.

Padilla N, Ascarza E (2021) Overcoming the Cold Start Problem of Customer Relationship Management Using a Probabilistic Machine Learning Approach. *Journal of Marketing Research* 58(5):981–1006.

Pang X (2010) Modeling Heterogeneity and Serial Correlation in Binary Time-Series Cross-sectional Data: a Bayesian Multilevel Model with AR(p) Errors. *Political Analysis* 18(4):470–498.

Pang X (2014) Varying Responses to Common Shocks and Complex Cross-Sectional Dependence: Dynamic Multilevel Modeling with Multifactor Error Structures for Time-Series Cross-Sectional Data. *Political Analysis* 22(4):464–496.

Pang X, Liu L, Xu Y (2021) A Bayesian Alternative to Synthetic Control for Comparative Case Studies. *Political Analysis* 30(2):1–20.

Pinkney S (2021) An Improved and Extended Bayesian Synthetic Control. arXiv.org. Retrieved `https://arxiv.org/abs/2103.16244`.

Redding SJ, Sturm DM (2008) The Costs of Remoteness: Evidence from German Division and Reunification. *American Economic Review* 98(5):1766–1797.

Rojas C, Wang E (2020) Do Taxes on Soda and Sugary Drinks Work? Scanner Data Evidence from Berkeley and Washington State. *Economic Inquiry* 59(1).

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rubin DB (1976) Inference and Missing Data. *Biometrika* 63(3):581.

Sarwar B, Karypis G, Konstan J, Riedl JT (2000) Application of Dimensionality Reduction in Recommender System - a Case Study. conservancy.umn.edu. Retrieved `https://conservancy.umn.edu/handle/11299/215429`.

Schmittlein DC, Morrison DG, Colombo R (1987) Counting Your Customers: Who Are They and What Will They Do Next? *Management Science* 33(1):1–24.

Tirunillai S, Tellis GJ (2017) Does Offline TV Advertising Affect Online Chatter? Quasi-Experimental Analysis Using Synthetic Control. *Marketing Science* 36(6):862–878.

Tuomaala E (2019) The Bayesian Synthetic Control: Improved Counterfactual Estimation in the Social Sciences through Probabilistic Modeling. arXiv.org. Retrieved `https://arxiv.org/abs/1910.06106`.

Wilson SE, Butler DM (2007) A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications. *Political Analysis* 15(2):101–123.

Xu Y (2017) Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis* 25(1):57–76.

Zhang Y, Bradlow ET, Small DS (2015) Predicting Customer Value Using Clumpiness: From RFM to RFMC. *Marketing Science* 34(2):195–208.

A

Theoretical Results

A.1 Proof of Proposition 1

We aim to provide a proof that demonstrates the relationship between *latent ignorability* and *strict exogeneity*. Such a relationship was first proposed by Pang et al. (2021), although no formal proof has been provided in the recent literature. Before outlining the framework for such a proof, we define *strict exogeneity* (Engle et al. 1983) using the language applied in this thesis.

Definition 7 (Strict Exogeneity)**.** A variable X_i is said to be strictly exogenous with respect to the error term ε_t if and only if the expectation of the error term, conditional on the exogenous variables, is zero for all time periods such that

$$\mathbb{E}[\varepsilon_t | X_i] = 0 \quad \forall t,$$

where ε_t represents the error term at time t, and X_i denotes the matrix of exogenous variables for unit i. This condition implies that the exogenous variables are uncorrelated with the error term, which ensures they do not contain information about the error process across all time periods.

Proof. To prove that latent ignorability extends strict exogeneity (Pang et al. 2021), we need to demonstrate that under latent ignorability, the conditional independence of the treatment assignment from the potential outcomes, given observed and latent covariates,

implies that there is no correlation between the treatment assignment and any error term in the model of potential outcomes.

Consider a linear model for the potential outcome such that

$$Y_i(w_i(a_i)) = X_i\beta + U_i\gamma + \varepsilon_i,$$

where ε_i represents the error term. The assumption of latent ignorability suggests that the treatment assignment $w_i(a_i)$ is independent of the potential outcomes Y_i, conditional on X_i and U_i

$$\mathbb{P}(w_i(a_i)|X_i, U_i) = \mathbb{P}(w_i(a_i)|X_i, Y_i(w_i(a_i)), U_i).$$

Given latent ignorability, it must then hold that

$$\mathbb{E}[\varepsilon_i|X_i, U_i, w_i(a_i)] = \mathbb{E}[\varepsilon_i|X_i, U_i] = 0,$$

which fulfills the condition of strict exogeneity for the error term ε_i relative to the covariates X_i and the latent variables U_i.

Therefore, by including latent variables U_i in our model, we are effectively adhering to the strict exogeneity assumption by controlling for all unobserved heterogeneity that could otherwise correlate with both the covariates X_i and the error term, as well as with the treatment assignment. This demonstrates that latent ignorability, through the inclusion of latent variables U_i, robustly extends the principle of strict exogeneity (Pang et al. 2021), ensuring the model against biases from unobserved heterogeneity. □

A.2 Proof of Proposition 2

We aim to provide a proof that demonstrates the relationship between *latent ignorability* and the *parallel trends* assumption, a connection implicitly mentioned by Pang et al. (2021). Before outlining the framework for such a proof, we again define *parallel trends* (Card and Krueger 1994) using the language applied in this thesis.

Assumption 9 (Parallel Trends)**.** For any unit i, the expected change in the observed outcomes $Y_{it}(0)$ over time, in the absence of treatment, is the same across units. If t and s represent two time periods, then

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0)|X_i] = \mathbb{E}[\varepsilon_{it} - \varepsilon_{is}],$$

where ε_{it} and ε_{is} are idiosyncratic errors for time periods t and s, respectively, and X_i represents observed covariates. This assumes that the change in observed outcomes over time is due to factors other than unobserved heterogeneity, which remains constant over time for each unit.

Proof. First, recall the expression for the potential outcome under no treatment for any time period t:

$$Y_{it}(0) = \beta X_{it} + u_i + \varepsilon_{it},$$

where β is a vector of coefficients, u_i is a unit-specific constant, and ε_{it} is the idiosyncratic error term. Given two time periods, t and s, we note the difference in potential outcomes as

$$Y_{it}(0) - Y_{is}(0) = (\beta X_{it} + u_i + \varepsilon_{it}) - (\beta X_{is} + u_i + \varepsilon_{is}).$$

Simplifying this expression, we observe that the unit-specific constant u_i cancels out:

$$Y_{it}(0) - Y_{is}(0) = \beta(X_{it} - X_{is}) + (\varepsilon_{it} - \varepsilon_{is}).$$

To derive the expected difference, we take the expectation of both sides and obtain

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0)] = \mathbb{E}[\beta(X_{it} - X_{is})] + \mathbb{E}[\varepsilon_{it} - \varepsilon_{is}].$$

Given that ε_{it} and ε_{is} are idiosyncratic error terms assumed to be i.i.d. with a mean of zero, the expectation of their difference is also zero:

$$\mathbb{E}[\varepsilon_{it} - \varepsilon_{is}] = 0.$$

In addition, since $\beta(X_{it} - X_{is})$ represents the fixed effects of covariates across time and does not depend on the unit-specific latent variable u_i, its expectation is a function of time only. Therefore, we can express the expected difference in potential outcomes as

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0)] = \mathbb{E}[\beta(X_{it} - X_{is})].$$

This expected difference, being solely a function of time (and covariates) and not of the unit-specific latent variable u_i, is what constitutes the parallel trends assumption. Hence, under the condition of latent ignorability and the assumption that U_i is a unit-specific constant, we demonstrate that the expected difference in potential outcomes under no treatment follows a parallel trend over time, hence satisfying the parallel trends assumption. This completes the proof that latent ignorability implies the parallel trends assumption when U_i is considered constant across time for each unit. □

A.3 Proof of Proposition 3

Proof. Denote the observed outcome for any unit i at time t by Y_{it}. Define the potential outcomes under treatment and control as $Y_{it}(1)$ and $Y_{it}(0)$, respectively. Let w_{it} be the treatment indicator, with $w_{it} = 1$ if unit i is treated at time t, and $w_{it} = 0$ otherwise.

The DiD estimator for the ATT is

$$\Delta_{\text{DiD}} = \left(\overline{Y}_{T,\text{post}} - \overline{Y}_{T,\text{pre}}\right) - \left(\overline{Y}_{C,\text{post}} - \overline{Y}_{C,\text{pre}}\right),$$

where $\overline{Y}_{T,\text{post}}$ and $\overline{Y}_{T,\text{pre}}$ denote the average outcomes for the treated units in the post-treatment and pre-treatment periods, respectively. Similarly, $\overline{Y}_{C,\text{post}}$ and $\overline{Y}_{C,\text{pre}}$ represent the corresponding averages for the control units.

The SCM constructs a synthetic control for the treated unit as a weighted average of control units:

$$\hat{Y}_{0t}(0) = \sum_{j=1}^{N} \hat{\beta}_j Y_{jt}(0),$$

where $\hat{\beta}_j$ are the weights assigned to control units j, determined by minimizing the pre-treatment prediction error, subject to $\hat{\beta}_j \geq 0$ for all j and $\sum_{j=1}^{N} \hat{\beta}_j = 1$.

Assuming SCM assigns equal weights to control units satisfying the parallel trends assumption with the treated unit, we have $\hat{\beta}_j = \frac{1}{N}$ for these units, and $\hat{\beta}_j = 0$ otherwise. The synthetic control outcomes in the pre-treatment and post-treatment periods are given by

$$\hat{Y}_{0,\text{pre}}(0) = \overline{Y}_{C,\text{pre}}, \quad \hat{Y}_{0,\text{post}}(0) = \overline{Y}_{C,\text{post}},$$

effectively equating to the average outcomes of control units that adhere to the parallel trends. The treatment effect on the treated, estimated by SCM in the post-treatment period, is

$$\delta_{0t} = Y_{0t} - \hat{Y}_{0t}(0) = \left(\overline{Y}_{T,\text{post}} - \overline{Y}_{C,\text{post}}\right),$$

which becomes equivalent to the DiD estimator Δ_{DiD} when the pre-treatment trends between the treated and control groups are parallel, i.e., $\overline{Y}_{T,\text{pre}} - \overline{Y}_{C,\text{pre}}$ is constant.

Hence, under the conditions that SCM assigns equal weights to control units satisfying the parallel trends assumption with the treated unit, SCM is mathematically equivalent to the DiD estimator, thus demonstrating SCM as a generalization of DiD under these specific conditions. □

A.4 Proof of Proposition 4

Proof. The functional form of the Bayesian causal MC model with independent multiple P's for unit i at time t and for outcome dimension p is

$$Y_{it}^{p} = w_{it}^{\top}\delta_{it}^{p} + X_{it}^{\top}\xi^{p} + Z_{it}^{\top}\zeta_{i}^{p} + A_{it}^{\top}\alpha_{t}^{p} + \Gamma_{i}^{\top} f_{t}^{p} + \varepsilon_{it}^{p}.$$

We begin by setting $Z_{it} = \varnothing$, implying that $Z_{it}^\top \zeta_i^p$ is removed from the model. This simplifies the Bayesian causal MC model to

$$Y_{it}^p = w_{it}^\top \delta_{it}^p + X_{it}^\top \xi^p + A_{it}^\top \alpha_t^p + \Gamma_i^\top f_t^p + \varepsilon_{it}^p.$$

Next, by setting $X_i = A_i$, we obtain

$$Y_{it}^p = w_{it}^\top \delta_{it}^p + X_i^\top \xi^p + X_i^\top \alpha_t^p + \Gamma_i^\top f_t^p + \varepsilon_{it}^p.$$

Assuming a single outcome dimension, we further simplify to

$$Y_{it} = w_{it}^\top \delta_{it} + X_i^\top \xi_t + X_i^\top \alpha_t + \Gamma_i^\top f_t + \varepsilon_{it}.$$

Given that in SCM, $\xi_t + \alpha_t$ can be seen as a single time-varying effect associated with the covariates X_i, we combine these terms. This yields the SCM model

$$Y_{it} = w_{it}^\top \delta_{it} + X_i^\top (\xi_t + \alpha_t) + \Gamma_i^\top f_t + \varepsilon_{it}.$$

By directly comparing the model to Abadie et al. (2010)'s underlying model of SCM, we notice that upon setting $\xi_t + \alpha_t$ to effectively represent the combined effect of covariates over time, the Bayesian causal MC model has been successfully reduced to match the SCM model. Therefore, by applying the specified conditions to the Bayesian causal MC model, we have demonstrated that it is a generalized form of SCM. □

A.5 Proof of Proposition 5

Proof. Similarly, the functional form of the Bayesian causal MC model with independent multiple P's for unit i at time t and for outcome dimension p is

$$Y_{it}^p = w_{it}^\top \delta_{it}^p + X_{it}^\top \xi^p + Z_{it}^\top \zeta_i^p + A_{it}^\top \alpha_t^p + \Gamma_i^\top f_t^p + \varepsilon_{it}^p.$$

We set $Z_{it} = A_{it} = \varnothing$, which removes these terms from the model. This simplification yields

$$Y_{it}^p = w_{it}^\top \delta_{it}^p + X_{it}^\top \xi^p + \Gamma_i^\top f_t^p + \varepsilon_{it}^p.$$

Given the constraint of considering only a single outcome dimension, we get

$$Y_{it} = w_{it}^\top \delta_{it} + X_{it}^\top \xi + \Gamma_i^\top f_t + \varepsilon_{it}.$$

This model aligns precisely with the functional form of Pinkney (2017)'s Bayesian SCM IFE model. Hence, this shows that our Bayesian causal MC model, under the specified conditions, is a generalized form of Pinkney (2017)'s Bayesian SCM IFE. □

B

Matrix Factorization

Matrix Factorization (MF)[1] is a technique widely used in recommender systems to reduce the dimensionality of complex data. According to Koren et al. (2009), the MF model characterizes both users and items by vectors of factors inferred from item rating patterns. By treating the data as a large user-item interaction matrix and further decomposing it into a set of latent factors, which are the product of two lower-dimensional matrices, the MF model allows for the capture of the underlying structure and hidden patterns in the data, ultimately used for prediction. In particular, such a model is best suited for collaborative filtering-based recommender systems.

In Appendix B, we first outline the prevalent strategies used in recommender systems. Following this, we review the generalized MF model as introduced by Koren et al. (2009). Subsequent sections, which adopt the empirical example used by Koren et al. (2009) to more effectively demonstrate the modeling steps, are dedicated to exploring the advanced features and algorithms behind the model. We then demonstrate its connection to SCM (Abadie et al. 2010). Appendix B initiates an open discussion on the interplay between these methodologies, ultimately framing our interpretation of the Bayesian causal MC model through the lens of MF, offering a distinct perspective from the conventional SCM-based approach.

[1] Please note that Appendix B represents an independent piece of writing produced during my previous research. While the emphasis differs from the rest of the thesis, we include this separate piece to provide additional understanding for readers. In particular, it aims to demonstrate how seemingly disparate methods (SCM, MF, and MC) can be integrated within a systematic framework.

B.1 Collaborative Filtering

As one of the two main strategies in recommender systems, the collaborative filtering approach, unlike the content filtering approach which creates a singular profile for each user or item to characterize its nature, aims to investigate the relationships between users and interdependencies among items to identify new user-item associations. Collaborative filtering can be easily adapted to various domains, addressing elusive data aspects that content filtering fails to capture, and providing more accurate predictions. Contrary to Koren et al.'s (2009) proposal, the *cold start* problem is not necessarily easier to address under content filtering. Instead, collaborative filtering can make recommendations based on similarities with other users or items in the data, even if there is implicit information about them.

There are two primary areas of collaborative filtering: neighborhood methods and LFMs. Neighborhood methods are more rudimentary as they focus solely on the relationships between items (i.e., an item-oriented approach) or between users (i.e., a user-oriented approach). The implementation of the MF model is based on the second area, LFMs. By characterizing both items and users by several latent factors inferred from rating patterns, these models provide clear dimensions for items and assess degrees of preference for users.

B.2 Generalized Model

The generalized MF model incorporates four components: user-item interactions, adding biases, implicit preferences, and user attributes.

B.2.1 User-Item Interactions

First, the MF model maps both users and items to a joint latent factor space of dimensionality f, where the user-item interactions are treated as inner products in f. Mathematically, each item i and each user u is associated with its corresponding vector, $q_i \in \mathbb{R}^f$ and $p_u \in \mathbb{R}^f$, respectively. For a given item i and a given user u, the elements of q_i and p_u measure the extent to which the item possesses those factors or the degree of interest that the user has in items. Intuitively, their interactions are expressed by $q_i^\top p_u$.

B.2.2 Adding Biases

Second, to account for the systematic differences among users and items, where some users tend to rate higher and some items are widely perceived as better, an adding bias

b_{ui} has been introduced by Koren et al. (2009), denoted as

$$b_{ui} = \mu + b_i + b_u,$$

where the intercept term μ stands for the overall average rating, and b_i and b_u represent the observed deviations of item i and user u, respectively. This added bias is independent of any user-item interactions, which explains why collaborative filtering is flexible in dealing with various data aspects.

B.2.3 Implicit Preference

Third, to counter the *cold start* problem, an additional input may be supplied to help gather behavioral information, regardless of the user's willingness to provide explicit ratings. This input is known as Boolean implicit feedback or, more specifically, a set of items $N(u)$ indicating each user u's implicit preferences. Each item i is associated with a vector $x_i \in \mathbb{R}^f$, and for a user u who prefers items $i \in N(u)$, the preference is represented by the vector

$$|N(u)|^{-\frac{1}{2}} \sum_{i \in N(u)} x_i.$$

Here, the sum is normalized for better interpretability and standardization purposes.

B.2.4 User Attribute

Fourth, similar to implicit preference, another optional input is user attributes (e.g., demographic information). Let $A(u)$ denote a set of attributes that a user u may have. A distinct factor vector $y_a \in \mathbb{R}^f$ corresponds to each attribute, describing a user through the set of associated user attributes as

$$\sum_{a \in A(u)} y_a.$$

Model 5 (Basic Matrix Factorization Model)**.** After incorporating these four components, we propose the most basic form of the MF model (Koren et al., 2009) as

$$\hat{r}_{ui} = \underbrace{q_i^\top p_u}_{\text{User-item Interactions}} + \underbrace{\mu + b_i + b_u}_{\text{Adding Biases}} + \underbrace{|N(u)|^{-\frac{1}{2}} \sum_{i \in N(u)} x_i}_{\text{Implicit Preference}} + \underbrace{\sum_{a \in A(u)} y_a}_{\text{User Attribute}},$$

where $\hat{r}_{ui}$ is the estimated rating of item i by user u.

Two additional advanced features, temporal dynamics and varying confidence levels, can also be integrated into the equation above. We present them in Sections B.2.5 and B.2.6.

B.2.5 Temporal Dynamics

Temporal dynamics account for situations when customers' inclinations evolve and perceptions of product popularity change. Specifically, this results in the item bias b_i and user bias b_u becoming functions of time t, denoted by $b_i(t)$ and $b_u(t)$. Item-user interactions are also influenced by temporal dynamics. As users may change their preferences, the user vector p_u becomes a function of time t, denoted by $p_u(t)$. However, unlike human characteristics, the item vector q_i remains static. Implicit preferences are generally considered to evolve over time; however, the treatment of user attributes is somewhat controversial. Some user attributes may change (e.g., income level, age group, zip code), while others may not (e.g., gender). Thus, after accounting for temporal dynamics, the estimated rating of item i by user u is given by

$$\hat{r}_{ui} = q_i^\top p_u(t) + \mu + b_i(t) + b_u(t) + |N(u)|^{-\frac{1}{2}} \sum_{i \in N(u)} x_i(t) + \sum_{a \in A(u)} (y_a + y_a(t)) \,.$$

B.2.6 Varying Confidence Levels

Adding a *weight* coefficient, such as varying confidence levels denoted by c_{ui}, can make the estimate more realistic, as it helps prevent a small number of deliberately adversarial ratings from damaging the entire recommender system. Varying confidence levels allow us to quantify the likelihood of customers' implicit preferences. For example, it is sensible to assign a higher weight to a recurring event as an indicator that the customer is more likely to provide a positive rating, and vice versa. This approach is particularly relevant for non-subscription marketing research.

Model 6 (Complete Matrix Factorization Model with Temporal Dynamics and Varying Confidence Levels)**.** The complete MF model, extending temporal dynamics and varying confidence levels to Koren et al.'s (2009) basic MF model, is defined as

$$\hat{r}_{ui} = c_{ui} \left[q_i^\top p_u(t) + \mu + b_i(t) + b_u(t) + |N(u)|^{-\frac{1}{2}} \sum_{i \in N(u)} x_i(t) + \sum_{a \in A(u)} (y_a + y_a(t)) \right],$$

where $\hat{r}_{ui}$ is the estimated rating of item i by user u, $q_i^\top p_u(t)$ captures the temporal user-item interactions through latent factors, μ represents the global average rating, $b_i(t)$ and $b_u(t)$ are the time-dependent biases for item i and user u, respectively. The terms $|N(u)|^{-\frac{1}{2}} \sum_{iinN(u)} x_i(t)$ and $\sum_{ainA(u)} (y_a + y_a(t))$ account for the implicit preferences and user attributes, both static and time-evolving. The model assumes that the confidence level c_{ui} scales the impact of each part on the final rating estimate, enhancing the reliability of the rating data.

B.3 Algorithms

Model 6, closely related to SVD, faces challenges due to the sparse nature of the user-item interaction matrix, which often leads to overfitting. This is because SVD does not handle missing ratings well, which are prevalent in real-world datasets. To address these issues and improve estimation accuracy, we implement two machine learning algorithms: stochastic gradient descent (SGD) and alternating least squares (ALS), based on Koren et al.'s (2009) selection. These methods are designed to minimize the regularized squared error on the known ratings, offering a solution that balances fitting the model to the training data with maintaining the ability to generalize to unseen data.

B.3.1 Stochastic Gradient Descent

SGD optimization, suggested by Funk (2006), iteratively updates model parameters by looping through all ratings in the training set. For each rating, the prediction error is calculated and used to adjust the item and user parameters in the direction that reduces the error. This method is computationally efficient and allows for quick adjustments to the model parameters.

B.3.2 Alternating Least Squares

ALS, recommended by Bell and Koren (2007), alternates between fixing user parameters to solve for item parameters and vice versa, facilitating the solving of two independent least-squares problems. This method is particularly effective for datasets with implicit feedback and can leverage parallelization to enhance computational efficiency. Unlike SGD, ALS does not require the setting of a learning rate, making it easier to use in some scenarios.

For both SGD and ALS, the goal is to minimize the objective function:

$$\min_{q^*,p^*,b^*} \sum_{(u,i)\in\kappa} c_{ui}\left(r_{ui} - q_i^T p_u - \mu - b_u - b_i\right)^2 + \lambda\left(\|q_i\|^2 + \|p_u\|^2 + b_u^2 + b_i^2\right),$$

where κ is the set of known (u,i) pairs, c_{ui} adjusts for confidence levels, μ is the global average rating, λ is a regularization parameter to control overfitting, and γ is the learning rate for SGD. See Algorithm 3 for detailed steps. We note that the convergence criteria depend on the change in the objective function between iterations. A threshold can be set to determine convergence.

Algorithm 3 Matrix Factorization Learning Algorithm

Objective: Minimize the regularized squared error in MF:

$$\min_{q^*,p^*,b^*} \sum_{(u,i)\in\kappa} c_{ui} \left(r_{ui} - q_i^T p_u - \mu - b_u - b_i\right)^2 + \lambda \left(\|q_i\|^2 + \|p_u\|^2 + b_u^2 + b_i^2\right)$$

▷ SGD for Parameter Updates

for each (u, i) in κ **do**
    Predict r_{ui} and compute error $e_{ui} = r_{ui} - (q_i^T p_u + \mu + b_u + b_i)$
    Update $q_i \leftarrow q_i + \gamma(e_{ui} p_u - \lambda q_i)$ ▷ Update item latent vector
    Update $p_u \leftarrow p_u + \gamma(e_{ui} q_i - \lambda p_u)$ ▷ Update user latent vector
end for

▷ ALS for Matrix Factorization

while convergence criteria not met **do** ▷ Iterate until convergence
    for each i **do**
        With p_u fixed, optimize q_i by minimizing the objective function
    end for
    for each u **do**
        With q_i fixed, optimize p_u by minimizing the objective function
    end for
end while

B.4 MF and SCM Interconnectedness

The interconnectedness between MF under collaborative filtering and the extended SCM, namely, Xu (2017)'s generalized SCM and Pinkney (2021)'s Bayesian SCM IFE, can broadly be seen in how they handle data and extract meaningful information from it.

MF is a broad class of latent variable models that includes factor analysis, encoder-decoder models in deep learning, and many others. It is a versatile technique applicable to a wide range of problems, from recommender systems to NLP and image recognition. Essentially, MF reduces the dimensionality of complex data by decomposing it into a set of latent factors, capturing the underlying structure in the data for future predictions.

Conversely, the extended SCM, designed for causal inference, utilizes a similar concept of latent factors to capture unobserved time-varying confounders that could affect the outcome variable. SCM is essentially a weighted combination of control units that closely match the characteristics of the treated unit prior to the intervention, akin to how MF uses latent factors to capture the underlying characteristics of users and items in a recommender system.

In the context of Bai (2009)'s IFE model, the least squares minimization is defined as

$$\sum_{i=1}^{N}(Y_{it} - X_{it}\beta - F_t\Lambda_i)^2,$$

essentially measuring the difference between the observed data and the data predicted by the model. This expression shares a similar structure with the regularized squared error minimization in MF such that

$$\min_{q^*,p^*,b^*} \sum_{(u,i)\in\kappa} c_{ui}\left(r_{ui} - q_i^T p_u - \mu - b_u - b_i\right)^2 + \lambda\left(\|q_i\|^2 + \|p_u\|^2 + b_u^2 + b_i^2\right),$$

aiming to minimize the difference between the observed data and the data reconstructed from the latent factors.

Hence, the extended SCM can be viewed as a specific manifestation of MF, tailored to the unique context of causal inference. They adopt the broad principles of MF – the use of latent factors to capture underlying structures in the data – and apply them within a context with specific assumptions and constraints. This essence of interconnectedness between MF and the extended SCM underlines that the former is a broad and versatile technique with wide applicability, while the latter represents a specific application of that technique. Stripping SCM back to a more primordial form is essentially a discovery of its roots in MF. Understanding SCM in the MF context allows us to gain a deeper insight into its principles and assumptions and potentially discover new ways to extend and apply it. Our proposed Bayesian causal MC model finds itself at the intersection of both MF and SCM, and in Section B.5, we discuss why the idea of MF could help us better counter the large N, small T, and multiple P problem.

B.5 A Hybrid Approach of MF and SCM

In the context of a marketing dataset with a large number of customers (large N), a short duration of time (small T), and multiple marketing outcomes (multiple P), the interconnectedness between MF under collaborative filtering and extended SCM can offer significant insights.

MF is adept at handling large-scale datasets by reducing the dimensionality of data through the decomposition into a set of latent factors. This is particularly beneficial for managing a large customer base (large N), enabling the capture of the underlying structure in customer data and unveiling hidden patterns for predictive purposes. However, MF generally presumes that all data points are i.i.d., an assumption that may not always be valid in panel data frameworks where observations are collected over time. Here, extended SCM becomes relevant.

SCM is equipped to manage time-varying confounders. Yet, traditional SCM might face challenges in large N and small T situations, often presupposing that the amount of pre-treatment periods is sufficient for estimating synthetic control weights – a notable small T issue.

Within the context of Bai (2009)'s IFE model, the large N and large T concept is leveraged to address scenarios featuring both a significant number of cross-sectional units and extensive time periods. This contrasts with our specific challenge of large N, small T, and multiple P. To address this, adaptation of SCM or MF techniques, or a hybrid approach, might be necessary. For example, MF could be employed to diminish the dimensionality of customer data (addressing large N and multiple P), followed by applying SCM or its variants to manage the time-varying data component (tackling small T). Furthermore, exploration into recent SCM and MF developments designed for such scenarios could be fruitful. For instance, some SCM variants have been introduced to accommodate a limited number of time periods, alongside MF methods tailored for time-series data analysis.

In conclusion, the synergy between MF under collaborative filtering and extended SCM can shed light on effective strategies for analyzing complex marketing datasets characterized by a vast customer base, brief time frames, and diverse outcomes. Acknowledging the strengths and limitations of each method, along with the potential for a hybrid approach, fosters the development of more advanced methodologies for prediction and inference. As the first to establish a connection between these methodologies, leveraging the advantages of both models has significantly contributed to the development of our proposed Bayesian causal MC model. In future research, we aim to strengthen its direct relationship with MF, creating a bridge between one of the SCM variants and the broader family of MF.

C

Model Replications

In Appendix C, we present only the essential replication codes for the four models discussed in this thesis. Please note that visualizations, output analysis, and the dataset have been omitted for clarity. Those interested in replicating these results are encouraged to visit my coding folders at https://rpubs.com/jiangzm, which is a collection of all previous `R` implementations documented in `Rmd` files. Some work is also documented in `ipynb` files, but these documents are not currently available online. If you are interested in accessing these documents, particularly those related to NLP models, please send an email to jiangzm@umich.edu, and I will provide the necessary code to you.

Please note that model replication is available exclusively for the German reunification study. Due to a non-disclosure agreement signed with the company that provided the CRM panel data, that particular dataset cannot be made public online. Those wishing to test and compare the following four models should refer to Hainmueller (2014) to access the German reunification data and run the codes below. If you encounter any difficulties, consider consulting the following resources that I wrote for replication guidance:

For Standard SCM replication, visit https://rpubs.com/jiangzm/1053384,

For Bayesian SCM IFE replication, visit https://rpubs.com/jiangzm/1054792, and

For Bayesian causal MC implementations, refer to https://rpubs.com/jiangzm/1069496 or https://rpubs.com/jiangzm/1095460.

C.1 Standard SCM

```
library(Synth)

# Load German Reunification Dataset
d <- read.dta("repgermany.dta")

# Initial data preparation for predictors and dependent variable
dataprep_init <- dataprep(
  foo = d,
  predictors = c("gdp", "trade", "infrate"),
  dependent = "gdp",
  unit.variable = 1,
  time.variable = 3,
  special.predictors = list(
    list("industry", 1971:1980, c("mean")),
    list("schooling", c(1970, 1975), c("mean")),
    list("invest70", 1980, c("mean"))
  ),
  treatment.identifier = 7,
  controls.identifier = unique(d$index)[-7],
  time.predictors.prior = 1971:1980,
  time.optimize.ssr = 1981:1990,
  unit.names.variable = 2,
  time.plot = 1960:2003
)

# Synth initialization
synth_init <- synth(
  data.prep.obj = dataprep_init,
  Margin.ipop = .005,
  Sigf.ipop = 7,
  Bound.ipop = 6
)

# Main data preparation for predictors and dependent variable
dataprep_main <- dataprep(
  foo = d,
  predictors = c("gdp", "trade", "infrate"),
  dependent = "gdp",
  unit.variable = 1,
  time.variable = 3,
  special.predictors = list(
    list("industry", 1981:1990, c("mean")),
    list("schooling", c(1980, 1985), c("mean")),
    list("invest80", 1980, c("mean"))
  ),
  treatment.identifier = 7,
```

```r
  controls.identifier = unique(d$index)[-7],
  time.predictors.prior = 1981:1990,
  time.optimize.ssr = 1960:1989,
  unit.names.variable = 2,
  time.plot = 1960:2003
)

# Synth main calculation
synth_main <- synth(
  data.prep.obj = dataprep_main,
  custom.v = as.numeric(synth_init$solution.v)
)

# Synth table generation
synth_df <- synth.tab(
  dataprep.res = dataprep_main,
  synth.res = synth_main
)

# GDP data preparation
dataprep_gdp <- dataprep_main$Y0

# Extracting synthetic weights
synth_weight <- synth_main$solution.w

# Calculating synthetic GDP
synth_gdp <- dataprep_gdp %*% synth_weight
```

Listing C.1: Implementation of Standard SCM using R

C.2 Bayesian SCM with IFE

```r
library(rstan)

# Load German Reunification Dataset
d <- read.dta("/Users/apple/Desktop/repgermany.dta")

# Load only outcome for German Reunification
german_unification <- read.csv("german_unification.csv")

# Data preprocessing for average computations
df_avg <- d |> group_by(index, country) |>
  summarize_at(
    vars(gdp, infrate, trade, industry, schooling, invest70, invest80),
    mean, na.rm = TRUE
  )

# Investment data adjustment
```

```r
df_avg <- mutate(df_avg, invest80 = (1 / 100) * invest80)

# Calculate average investment for 1975
df_avg <- mutate(df_avg, invest75 = mean(c_across(c("invest70", "invest80")), na.rm = TRUE))

# Prepare data for Bayesian SCM IFE
df_J_P <- df_avg[, c(3:7, 10)]
rownames(df_J_P) <- 1:17
colnames(df_J_P) <- c("avg_gdp", "avg_infrate", "avg_trade", "avg_industry", "avg_schooling", "avg_invest75")

df_J_T <- german_unification[, 2:45]
colnames(df_J_T) <- 1960:2003
rownames(df_J_T) <- 1:17

# Stan model code (simplified for brevity)
stan_code <- "functions { ... }"  # Refer to Appendix of Pinkney (2019)

# Data list preparation for Stan
data_list <- list(
  T = length(unique(d$year)),
  J = length(unique(d$country)),
  L = 8,
  P = 6,
  X = data.matrix(t(df_J_P)),
  Y = data.matrix(df_J_T),
  trt_times = max(d$year) - 1990
)

# Control list for Stan
control_list <- list(max_treedepth = 14, adapt_delta = 0.95)

# Stan model fitting
fit <- stan(model_code = stan_code, data = data_list, init = "0.1",
            control = control_list, chains = 4, warmup = 250, iter = 500)

# Posterior analysis and output generation
# Example code for generating synthetic GDP outputs and their credibility intervals
```

Listing C.2: Implementation of Bayesian SCM with IFE using Stan and R

C.3 Bayesian Causal MC with Independent Multiple P's

```r
# Load external C++ code
```

```r
Rcpp::sourceCpp("bpCausal-main/src/blasso.cpp")

# Load essential R scripts (refer to Pang et al. 2021)
source("bpCausal-main/R/blasso_default.R")
source("bpCausal-main/R/blasso_core.R")

# Define the main function for Bayesian causal MC model
BCMC <- function(data, index, Yname_vector, Dname, Xname, Zname, Aname,
                 re, ar1, r, niter = 15000, burn = 5000,
                 xlasso = 1, zlasso = 1, alasso = 1, flasso = 1,
                 a1 = 0.001, a2 = 0.001, b1 = 0.001, b2 = 0.001,
                 c1 = 0.001, c2 = 0.001, p1 = 0.001, p2 = 0.001) {

  # Applying Bayesian Causal inference on each outcome variable
  out <- lapply(Yname_vector, function(Yname_single) {
    bpCausal(data = data,
             index = index,
             Yname = Yname_single,
             Dname = Dname,
             Xname = Xname,
             Zname = Zname,
             Aname = Aname,
             re = re,
             ar1 = ar1,
             r = r,
             niter = niter,
             burn = burn,
             xlasso = xlasso,
             zlasso = zlasso,
             alasso = alasso,
             flasso = flasso,
             a1 = a1, a2 = a2,
             b1 = b1, b2 = b2,
             c1 = c1, c2 = c2,
             p1 = p1, p2 = p2)
  })

  return(out)
}

# Running BCMC for multiple outcomes
OUT <-
  BCMC(data = hypo_synth_trt, index = c("id", "T"),
       Yname_vector = c("F_purch", "M_purch", "R_purch", "F_redem", "M_
    redem", "R_redem"),
       Dname = "D", Xname = c(), Zname = c(), Aname = c(),
       re = "both", ar1 = TRUE, r = 8)

```

```r
# Function to estimate counterfactual outcomes
counterfactual_est <- function(x) {
    # Iteration count
    niter <- dim(x$sigma2)[2]

    # Counterfactual outcomes
    yct_i <- x$yct
    yct_i <- matrix(c(yct_i[, (1):niter]), nrow(yct_i), niter)

    # Original outcomes and identifiers
    yo_t <- x$yo_t
    id_tr <- x$raw.id.tr
    time_tr <- x$time.tr

    # Mean counterfactual estimates
    m_yct_mean <- apply(yct_i, 1, mean)

    # Credibility intervals
    m_yct_ci_l <- apply(yct_i, 1, quantile, 0.025)
    m_yct_ci_u <- apply(yct_i, 1, quantile, 0.975)

    # Compile results
    result_x <- data.frame(
        id = id_tr,
        T = as.integer(time_tr),
        original_outcome = yo_t,
        counterfactual_estimate = m_yct_mean,
        ci_lower = m_yct_ci_l,
        ci_upper = m_yct_ci_u
    )

    return(result_x)
}

# Apply counterfactual estimation across all outputs
counterfactual_results <- lapply(OUT, counterfactual_est)
```

Listing C.3: Implementation of Bayesian Causal MC for Independent Multiple P's using R and Rcpp

C.4 Bayesian Causal MC with Concurrent Multiple P's

```python
import argparse
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```python
from jax import jit, random, numpy as jnp
import numpyro as npr
from numpyro import infer, distributions as dist

# Function to create beta parameter
@jit
def make_beta(beta_off: jnp.ndarray,
              lambd: jnp.ndarray,
              eta: jnp.ndarray,
              tau: jnp.ndarray):

    cache = jnp.tan(0.5 * jnp.pi * lambd) * jnp.tan(0.5 * jnp.pi * eta)
    tau_ = jnp.tan(0.5 * jnp.pi * tau)
    out = jnp.diag(cache) @ (beta_off * tau_)

    return out

# Function to create Sigma matrix
@jit
def make_Sigma(Sigma_diag):
    L = Sigma_diag.shape[1]
    Sigma = Sigma_diag[:,:,jnp.newaxis] * jnp.eye(L)

    return Sigma

# Function to create Phi matrix
@jit
def make_Phi(Sigma: jnp.ndarray,
             F: jnp.ndarray,
             beta: jnp.ndarray):
    # Calculate Phi matrix: KxLxL @ TxL @ LxJ = KxTxJ
    Phi = F[jnp.newaxis, :, :] @ Sigma @ beta[jnp.newaxis, :, :]

    return Phi

# Function to create counterfactual Y matrix
def make_Y(Y_1_pre: jnp.ndarray,
           Y_1_post: jnp.ndarray,
           Y_0: jnp.ndarray):
    # Concatenate Y_1_pre and Y_1_post to form a complete Y matrix
    Y = jnp.concatenate([Y_1_pre, Y_1_post]) # TxK
    Y_reshaped = Y.reshape(1, *Y.shape) # 1xTxK
    Y = jnp.concatenate([Y_reshaped, Y_0], axis=0) # JxTxK
    Y = jnp.transpose(Y, axes=(1, 0, 2)) # TxJxK
    return Y

# Main model function
def model_sur_scm(Y_0: jnp.ndarray, Y_1_pre: jnp.ndarray, L: int):
```

```
    # Model initialization
    J = Y_0.shape[0] + 1
    T = Y_0.shape[1]
    K = Y_0.shape[2]
    T_post = T - Y_1_pre.shape[0]

    # Define parameters and priors
    eta = npr.sample("eta", dist.Uniform())

    # Sample lambda, tau, beta_off, Sigma_diag, F, kappa, delta
    # For each sampling, use appropriate numpyro distribution
    with npr.plate("L",L):
        lambd = npr.sample("lambda", dist.Uniform())

    with npr.plate("J", J):
        tau = npr.sample("tau", dist.Uniform())

    with npr.plate("L", L, dim=-2), npr.plate("J", J, dim=-1):
        beta_off = npr.sample("beta_off", dist.Normal())

    with npr.plate("K", K, dim=-2), npr.plate("L", L, dim=-1):
        Sigma_diag = npr.sample("Sigma_diag", dist.Normal())

    with npr.plate("T", T, dim=-2), npr.plate("L", L, dim=-1):
        F = npr.sample("f", dist.Normal())

    with npr.plate("K", K, dim=-2), npr.plate("J", J, dim=-1):
        kappa = npr.sample("kappa", dist.Normal())

    with npr.plate("K", K, dim=-2), npr.plate("T", T, dim=-1):
        delta = npr.sample("delta", dist.Normal(scale=2))

    with npr.plate("K", K):
        sig_err = npr.sample('L_sigma', dist.HalfCauchy())

    # Data augmentation for post-treatment period
    Y_1_post = npr.sample("Y_1_post", dist.Normal().mask(False))

    # Calculate transformed variables: beta, Sigma, Phi
    beta = make_beta(beta_off, lambd, eta, tau)
    Sigma = make_Sigma(Sigma_diag)
    Phi = make_Phi(Sigma, F, beta) # Phi: KxTxJ
    Y = make_Y(Y_1_pre, Y_1_post, Y_0) # Y: TxJxK

    # Calculate mu and reshape for likelihood
    # Sample Y from Normal distribution with mu_reshaped and sig_err
    mu = Phi + delta[:, :, jnp.newaxis] + kappa[:, jnp.newaxis, :]
    mu_reshaped = mu.transpose([1, 2, 0]).reshape(T * J, K)
```

```
101     Y_reshaped = Y.reshape(T * J, K)
102
103     with npr.plate("T*J", T * J, dim=-2), npr.plate("K", K, dim=-1):
104         npr.sample("Y", dist.Normal(loc=mu_reshaped, scale=sig_err), obs=
        Y_reshaped)
105
106 # Function to get data
107 def get_data(path: str = None):
108     # Load and process German reunification data
109     # Include normalization and whitening steps
110     # Return processed data and whitening parameters
111
112 # Main function
113 def main(args):
114     # Initializations
115     rng_key = random.PRNGKey(args.seed)
116     rng_key, rng_key_mcmc, rng_key_predict = random.split(rng_key, 3)
117     x_values,Y_0, Y_1_obs, Y_1_pre, whitening1, whitening2, whitening3 =
        get_data()
118
119     T = Y_0.shape[1]
120     L = args.num_latent
121     J = Y_0.shape[0] + 1
122
123     # Inference
124     nuts_kernel = infer.NUTS(model_sur_scm, max_tree_depth=8,
        target_accept_prob=0.8)
125     mcmc = infer.MCMC(nuts_kernel, num_warmup=args.iter, num_samples=args.
        iter, num_chains=1)
126     mcmc.run(rng_key_mcmc,Y_0, Y_1_pre, L)
127
128     # Print
129     mcmc.print_summary()
130     posterior_samples = mcmc.get_samples()
131
132     # Posterior Predictive Distribution
133     ppd = infer.Predictive(model_sur_scm, posterior_samples, num_samples=
        args.iter, parallel=True) # Y is TxJxK
134     Y_counterfactual = ppd(rng_key_predict,Y_0, Y_1_pre, L)["Y"]
135     K=3
136     Y_1_counterfactual = jnp.array(Y_counterfactual).reshape([args.iter,T,J
        ,K])[:,:,0,:]
137     Y_1_counterfactual = Y_1_counterfactual.reshape([args.iter,T,K])
138     Y1_1_counterfactual = Y_1_counterfactual[:,:,0]
139     Y2_1_counterfactual = Y_1_counterfactual[:,:,1]
140     Y3_1_counterfactual = Y_1_counterfactual[:, :, 2]
141     y1p_mu = Y1_1_counterfactual.mean(axis=0)
142     y2p_mu = Y2_1_counterfactual.mean(axis=0)
```

```
    y3p_mu = Y3_1_counterfactual.mean(axis=0)

if __name__ == "__main__":
    parser = argparse.ArgumentParser(description="parse args")
    parser.add_argument("-n", "--num-latent", default=30, type=int)
    parser.add_argument("-seed", default=20240423, type=int)
    parser.add_argument("-iter", default=3000, type=int)

    args = parser.parse_args()
    main(args)
```

Listing C.4: Implementation of Bayesian Causal MC for Concurrent Multiple P's using JAX and NumPyro

C.5 Yelp's Fusion API

```
import pandas as pd
import requests
import time as t

# Load CRM panel data
JA_Cov = pd.read_csv("CRMpaneldata.csv")

# Define a function to search Yelp based on project location and full
    address
def search_yelp(proj_loc, full_address, api_key):
    # Yelp API endpoint for business search
    endpoint = "https://api.yelp.com/v3/businesses/search"
    headers = {
        "Authorization": f"Bearer {api_key}",
    }
    params = {
        "term": proj_loc,
        "location": full_address,
        "limit": 1
    }

    # Make the API request
    response = requests.get(endpoint, headers=headers, params=params)
    if response.status_code == 200:
        # Return JSON response if successful
        return response.json()
    else:
        # Error handling
        print(f"API request failed for row {index}.")
        return None

```

```
# Place your Yelp API keys here
api_key_1 = " ... "
api_key_2 = " ... "

# Initialize a list to store Yelp data
yelp_data_list = []

# Time delay between requests to avoid hitting rate limit
sleep_time = 0.5

# Iterate through the CRM panel data
for index, row in JA_Cov.iterrows():
    proj_loc = row['proj_loc']
    full_address = row['full_address']

    # Alternate between two API keys to balance the quota usage
    current_api_key = api_key_1 if index % 2 == 0 else api_key_2

    # Call the Yelp search function
    result = search_yelp(proj_loc, full_address, current_api_key)

    # Process the result and append business info to the list
    if result and 'businesses' in result:
        for business in result['businesses']:
            yelp_data_list.append({
                'yelp_tag': ', '.join([cat['title'] for cat in business.get('categories', [])]),
                'rating': business.get('rating'),
                'review_num': business.get('review_count'),
                'price_level': business.get('price'),
                'transactions': ', '.join(business.get('transactions', [])),
                'yelp_url': business.get('url')
            })
    else:
        # Append None values if no business info is found
        yelp_data_list.append({
            'yelp_tag': None,
            'rating': None,
            'review_num': None,
            'price_level': None,
            'transactions': None,
            'yelp_url': None
        })

    # Wait for a specified time before making the next request
    t.sleep(sleep_time)

```

```
# Convert the list of Yelp data into a DataFrame
yelp_data = pd.DataFrame(yelp_data_list)

# Concatenate the original data with the retrieved Yelp data
JA_C_yelp = pd.concat([JA_Cov, yelp_data], axis=1)
```

Listing C.5: Integrating Yelp's Fusion API for Covariate Data Enrichment in Python

C.6 FastText Tag Embeddings

```
from gensim.models.fasttext import FastText
import pandas as pd
import numpy as np
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt
import time as t

# Load the FastText model using pre-trained data
model = FastText.load_fasttext_format("cc.en.300.bin")

# Data Loading from CSV
X = pd.read_csv("CRMpaneldata.csv")

# Numerical data imputation
numerical_X = X.select_dtypes(include=['float64', 'int64']).copy()
numerical_X['price_level'].replace({0: np.nan}, inplace=True)

imputer = IterativeImputer(max_iter=10, random_state=0)
numerical_X_imputed = imputer.fit_transform(numerical_X)

numerical_X_imputed_df = pd.DataFrame(numerical_X_imputed, columns=
    numerical_X.columns, index=numerical_X.index)
X[numerical_X.columns] = numerical_X_imputed_df

# Preparing tags for embedding extraction
X['yelp_tag_list'] = X['yelp_tag'].apply(lambda x: str(x).split(','))

def get_embedding_from_tags(tags, model):
    embeddings = [model.wv[tag.strip()] for tag in tags if tag.strip() in
    model.wv]
    return np.mean(embeddings, axis=0) if embeddings else np.zeros(model.
    vector_size)

```

```
X['average_yelp_tag_embedding'] = X['yelp_tag_list'].apply(lambda tags:
    get_embedding_from_tags(tags, model))

# PCA for dimensionality reduction
embedding_matrix = np.vstack(X['average_yelp_tag_embedding'].apply(lambda x
    : np.array(x)))
pca = PCA(n_components=9)
reduced_embeddings = pca.fit_transform(embedding_matrix)

# Compute cosine similarity for original and reduced embeddings
similarity_matrix_original = cosine_similarity(embedding_matrix)
similarity_matrix_reduced = cosine_similarity(reduced_embeddings)

# t-SNE for dimensionality reduction
tsne = TSNE(n_components=2, random_state=42)
reduced_embeddings_tsne = tsne.fit_transform(embedding_matrix)

# Generate a combined panel data frame with original data and extracted
    features
X['pca_feature_1'] = reduced_embeddings[:, 0]
X['pca_feature_2'] = reduced_embeddings[:, 1]
X['tsne_feature_1'] = reduced_embeddings_tsne[:, 0]
X['tsne_feature_2'] = reduced_embeddings_tsne[:, 1]
```

Listing C.6: Extracting and Analyzing FastText Embeddings from Yelp Tags in Python

C.7 BERT and ELECTRA

```
import pandas as pd
import numpy as np
from transformers import ElectraTokenizer, ElectraModel
import torch
from tqdm import tqdm

# Load data from CSV file
X = pd.read_csv("CRMpaneldata.csv")

# Initialize ELECTRA Small-Discriminator Model and Tokenizer
tokenizer = ElectraTokenizer.from_pretrained('google/electra-small-
    discriminator')
model = ElectraModel.from_pretrained('google/electra-small-discriminator')

# Function to generate embeddings in batches for given texts
def batch_bert_embedding(texts):
    # Replace NaN texts with empty strings
    texts = ['' if pd.isna(text) else text for text in texts]

    # Tokenize the batch of texts
```

```
    inputs = tokenizer(texts, padding=True, truncation=True, max_length=64, return_tensors="pt")

    # Generate embeddings without updating gradients
    with torch.no_grad():
        outputs = model(**inputs)

    # Extract the embeddings of the first token ([CLS] token) as sentence embeddings
    embeddings = outputs.last_hidden_state[:, 0, :].numpy()

    return embeddings

# Define batch size for processing
batch_size = 32

# List of UTM columns to process
utm_columns = ['utm_campaign', 'utm_medium', 'utm_content', 'utm_source']

# Process each UTM column to generate embeddings
for col in tqdm(utm_columns, desc='Processing UTM columns'):

    # Initialize an array to hold all embeddings for the current column
    all_embeddings = np.empty((0, 256))

    # Calculate the total number of batches needed
    total_batches = int(np.ceil(len(X) / batch_size))

    # Process each batch
    for i in tqdm(range(0, len(X), batch_size), desc=f'Processing {col}', total=total_batches):
        # Select the current batch of data
        batch = X[col][i:i + batch_size]

        # Generate embeddings for the batch
        batch_embeddings = batch_bert_embedding(batch)

        # Stack the embeddings to accumulate them
        all_embeddings = np.vstack([all_embeddings, batch_embeddings])

    # Convert the embeddings into a DataFrame
    bert_df = pd.DataFrame(all_embeddings, columns=[f"{col}_electra_{i}" for i in range(all_embeddings.shape[1])])

    # Concatenate the new DataFrame of embeddings with the original data
    X = pd.concat([X, bert_df], axis=1)
```

Listing C.7: Customer/Project-Level Covariate Data using BERT/Electra in Python