Enhancing Recommender Systems with Causal Inference Methodologies

by

Huiqing Huang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Data Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In the current era of data deluge, recommender systems (RSs) are widely recognized as one of the most effective tools for information filtering. However, traditional RSs are founded on associational relationships among variables rather than causality, meaning they are unable to determine which factors actually affect user preference. In addition, the algorithm of conventional RS continues to recommend similar items to users, resulting in user aesthetic fatigue and ultimately the loss of customer sources. Moreover, the generation of recommendations could be biased by the confounding effect, leading to inaccurate results. To tackle this series of challenges, causal inference for recommender systems (CI for RSs) has emerged as a new area of study. In this paper, we present four different propensity score estimation methods, namely hierarchical Poisson factorization (HPF), logistic regression, non-negative matrix factorization (NMF), and neural networks (NNs), and five causal effect estimation methods, namely linear regression, inverse probability weighting (IPW), zero-inflated Poisson (ZIP) regression, zero-inflated Negative Binomial (ZINB) regression, and doubly robust (DR) estimation. Additionally, we propose a new algorithm for parameter estimation based on the concept of alternating gradient descent (AGD). Regarding the study's reliability and precision, it will be evaluated on two distinct categories of datasets. Our research demonstrates that the causal RS can correctly infer causality from user and item characteristics to the final rating with an accuracy of 96%. Moreover, according to the de-confounded and de-biased recommendations, ratings can be increased by an average of 1.6 points (out of 4) for the Yahoo! R3 dataset and 1.2 points (out of 2) for the Restaurant and Consumer data.

Acknowledgements

First, I would like to express my deep gratitude to my supervisor, Dr. Yeying Zhu, for her constant support and guidance throughout the duration of this research endeavour. Her profound expertise and perceptive perspectives were indispensable to the creation and completion of this work.

Dr. Liqun Diao and Dr. Alex Stringer have my sincere appreciation. Their willingness to engage with my research and provide constructive feedback during the reading of my thesis and participation in my defense contributed substantially to the improvement and precision of my research.

I am indebted to my family for their unending encouragement and support. Throughout the difficult phases of this research voyage, their confidence in my abilities and unwavering patience have been my inspiration.

Last but not least, I would like to express my gratitude to all those who have contributed to this research in one way or another, particularly Jinglan and Bob, who provided statistical and mathematical support, and their inspirations were essential to the completion of this study. I am sincerely appreciative of your support and encouragement throughout this research process.

Dedication

This is dedicated to my family for their unending love and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Recommender Systems

People are surrounded by immense amounts of information, whether essential or not, as a result of the proliferation of digital content, and the quantity is increasing exponentially every day, resulting in data overload. Furthermore, users must independently determine what information they need, which is a time-consuming and unreliable process that relies on human judgment. Tapestry [Goldberg et al., 1992], the first recommender system, was introduced in 1992 in an effort to save consumers' time and assist them in locating pertinent and high-quality content more efficiently. Tapestry's concept is to enable users to annotate documents and messages with various tags so that they can use these annotations to filter the information. This technique is also known as collaborative filtering because the tags of users (i.e., collaborators) can be used to assist others in filtering information.

In addition to filtering the data, RSs have been designed to make recommendations based on the users' interests and preferences (Schafer et al., 1999), thereby enhancing the users' experience by displaying what they believe the users will find enthralling. This feature is beneficial to e-commerce [Schafer et al., 1999] because it can increase users' engagement and retention, therefore catalyzing the behaviour flow (from viewing content to completing purchases). It can also aid users in discovering new products in which they may be interested but have not found on their own, thus increasing the companies' revenue and profitability. Since the introduction of the first commercial RS [Goldberg et al., 1992], a growing number of platforms, including Netflix [Gomez-Uribe and Hunt, 2016], YouTube [Davidson et al., 2010], and Amazon [Linden et al., 2003], have incorporated RSs into their commercial mechanisms.

Since the end of the 1990s, algorithms to improve RSs have been modified, and seven main types can be briefly categorized: collaborative filtering [Goldberg et al., 1992], content-based filtering [Van Meteren and Van Someren, 2000], hybrid filtering [Thorat et al., 2015], knowledge-based filtering [Burke, 2000], demographic-based filtering [Pazzani, 1999], association rule learning [García et al., 2009], and deep learning-based RSs [Zhang et al., 2019], with countless algorithms in total. All of these algorithms, however, have fatal defects, such as Bias, Low Diversity, and Low Interpretability.

Bias: There are various types of biases, including selection bias, exposure bias, popularity bias, conformity bias, and position bias [Xu et al., 2023b]. Selection bias, also known as missing not at random (MNAR) data, indicates that feedback cannot accurately reflect the items' quality due to user selections. Popularity bias occurs because RSs prefer to recommend high-rating products [Wei et al., 2021]. Exposure bias is caused by the fact that consumers are not exposed to products randomly, as they are typically shown alongside comparable products based on the classic RSs procedure. Consequently, diverse product categories are not recommended. Conformity bias refers to the tendency of individuals to adopt the behavior of others, even if it contradicts their own. Position bias is a result of people tending to choose the items in a prominent position when presented with a recommendation list. These biases will impair the outcomes of the RSs [Xu et al., 2023b], diminishing customer satisfaction and company revenue.

Low Diversity: RSs generate recommendations based on correlations or association-based algorithms; therefore, they will recommend comparable items based on what users have expressed an interest in. In conjunction with the influence of popularity bias, the scenario of displaying similar products will continue to deteriorate, ultimately resulting in customers losing interest and abandoning the platform.

Low interpretability: Since RSs do not examine the causal relationship between user activity and item preferences [Yao et al., 2021], we do not know if promoting the present item will enhance a user's experience. Instead, we only know that there are certain connections between the item and the user's happiness, and it is possible that another unidentified factor affects the user's experience [Sato, 2021]. If we can identify this cause, we can recommend what the user actually enjoys, thereby increasing the user's satisfaction. In order to investigate the causal relationship and make objective recommendations, causal inference is necessary.

1.2 Causal Inference

Causal inference (CI) is a statistical and machine-learning technique that aims to establish the causal relationship among variables, and it plays a crucial role in numerous fields, including statistics, epidemiology, economics, and social science [Imbens and Rubin, 2015]. There are two widely-used causal frameworks: structural causal models (SCMs) and Rubin's causal models (RCMs).

SCMs [Pearl, 2009b] can convert causality between variables into a causal graph, define structural functions, infer causality, and assess the causal effect of interventions or counterfactuals. Typically, a causal graph is represented by a directed acyclic graph (DAG), and Figure 1.1 illustrates three types of graphs. Suppose we have two variables X and Y; according to the Chain structure, X influences Y via a mediator Z. Variable Z in the Fork structure simultaneously influences variables X and Y; this is known as a confounder. Unobserved Z can result in a miscalculation of the correlation between X and Y due to the confounding effect, which can introduce bias into the standard RSs [Xu et al., 2023a]. Given the preceding example from **Low Interpretability**, RS would incorrectly assert that X and Y have a causal relationship, when in fact, Z is the true common cause of X and Y. In the last structure, the Collider structure, both X and Y affect the collider variable Z, resulting in a collision effect; hence, X and Y are now connected despite being marginally independent [Zhu et al., 2023].

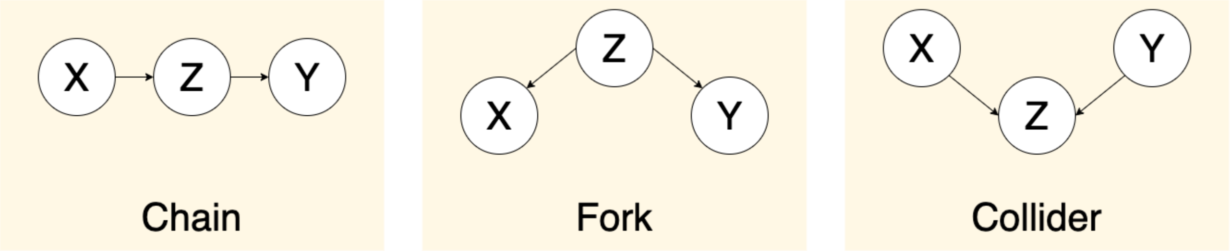


Figure 1.1: Three types of DAGs

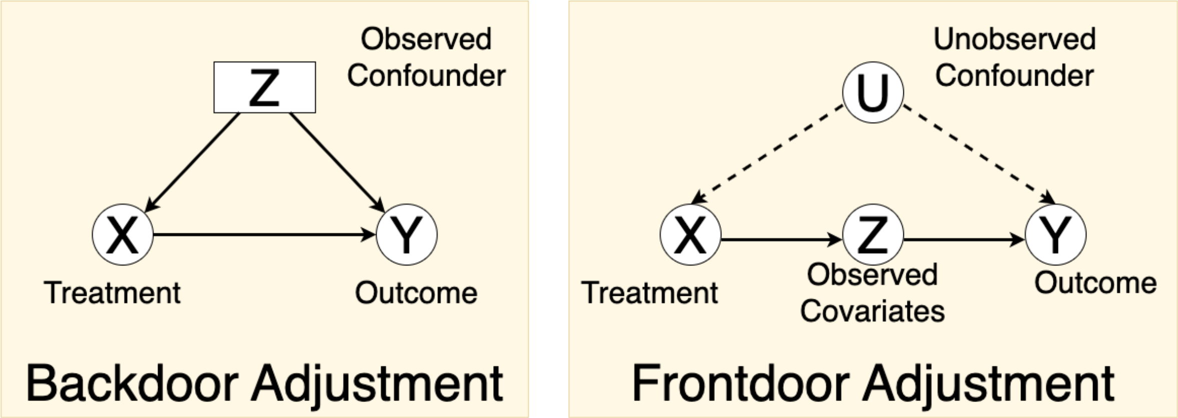
Once we have a DAG and its corresponding structural functions, we can estimate the causal effects of interventions or counterfactuals. In both SCMs and RCMs, intervention and counterfactuals are fundamental concepts. Intervention is how we will intervene in a treatment variable X with a do-calculus, do(X=x) [Pearl, 2009b]. In Figure 1.1-Fork, for instance, if we execute a do-operation on variable X, i.e., assign X the value x, the path from Z to X will be eliminated, allowing us to determine the causality between Z and Y. Counterfactual, as its name suggests, is the opposite of a factual; it is the result of assigning a different value to a treatment variable from what is observed in actuality.

RCMs or Potential Outcome Framework [Rubin, 1974] is a well-known CI framework that is logically comparable to SCMs [Pearl, 2009a], while it can estimate the causal effect without sketching a causal DAG. RCM requires the notion of potential outcomes, which can be denoted as Y_i^t, the value of outcome variable Y if individual i receives treatment T with value t. In the case of a binary treatment, we are interested in estimating the individual treatment effect (ITE) Y_i^1 - Y_i^0. However, measuring ITE is impractical due to the fact that only one treatment can be administered to a person, resulting in the observation of only one potential outcome. Therefore, researchers extended ITE into an average treatment effect (ATE), expressed as $E[Y_i^1 - Y_i^0]$, and then the causal effect could be estimated.

Yet, the prerequisite for correctly estimating causal impact is a randomized experiment, in which individuals are randomly assigned to various treatments such that there are no confounders. Unfortunately, this is not typical in the real world, and we must discover a way to de-confound. In the ideal situation in which we observe every confounding variable, we can intervene with do-operation by assigning the same value to each confounding variable for everyone, resulting in identical characteristics for all individuals. After the intervention, there is no difference between people, so we can randomly designate them to the control or treatment group. This method is referred to as Backdoor Adjustment [Pearl, 2009b] (Figure 1.2-Left).

Nevertheless, there are a tremendous number of variables in our lives, and most of the time we cannot observe all confounders [Xu et al., 2023a]. Here, Frontdoor Adjustment can be of assistance (Figure 1.2-Right). Frontdoor Adjustment states that if we can identify a collection of variables Z such that all paths from X to Y pass through Z and there is no unblocked backdoor path between Z and Y, we can infer the causality from X to Y and eliminate unobserved confounding variables [Pearl, 1995].

In an observational study where the treatment is not randomized, we can mimic a randomized experiment by estimating propensity scores using the confounding variables. Then, based on the propensity scores, the distribution of the observed covariates will be similar between treated and untreated objects [Austin, 2011]. Given observed covariates or confounders, the propensity score represents the probability that an individual will receive the treatment [Rosenbaum and Rubin, 1983]. "Evidence of residual bias in the propensity score is evidence of prospective bias in estimated treatment effects", Rosenbaum and Rubin [1983] thereby establishing the propensity score as an essential component of causal inference. In addition to combining the information of a vector of covariates into a scalar number, the propensity score estimation method is also a dimension reduction technique. After obtaining the propensity score e(X), the causal effect can be estimated using inverse probability weighting (IPW) [Hirano et al., 2003], which gives a weight $\frac{1}{e(\boldsymbol{X})}$



Left: All confounding variables have been accounted for; the backdoor between X and Y has been closed off, and the effect of X on Y can be estimated.

Right: Not all confounders can be observed; instead, observe certain covariates Z such that no backdoor paths between X and Z and every backdoor path between Z and Y are blocked by X, then the effect of X on Y can still be estimated.

Figure 1.2: Two types of adjustments

for a subject in the treatment group and $\frac{1}{1-e(\boldsymbol{X})}$ for a subject in the control group.

1.3 Causal Inference for Recommender Systems

1.3.1 Background

In CI for RSs, we aim to determine whether a user will like the item we recommend; in other words, if we present this product to a customer, will he or she like it? The question may appear congruent with the concept of conventional RS; however, let us consider the following example.

Assume Lisa is a movie enthusiast. In her rating history, she has given positive ratings to dozens of comedies but only two action films; these are the only two action movies she has watched so far, whereas she has watched a significant amount of comedies. One of the possible explanations is that it is difficult for her to have access to action films in her region. The traditional RSs will unquestionably introduce her to comedic films, but due to a lack of knowledge, they may overlook action films. Nonetheless, a causal RS can extract this insignificant characteristic from a large dataset. Hence, in contrast to a conventional RS, we attempt to show consumers an item that piques their attention, but only after we reveal it. Based on Lisa's past behaviour, we can assume that if we show her action movies, she will enjoy them, leading to a good score. If CI can demonstrate the aforementioned theory and expose Lisa to more action films, then, in comparison to a traditional RS, ours can accommodate user preferences more effectively.

In a causal RS, the treatment variable is whether we should expose an item to a user or not, and the outcome variable is the item's rating. Conducting a randomized trial in this context is prohibitively expensive and infeasible since we are unable to randomly display products to clients, as this would have a negative impact on the user experience and platform utility. Consequently, causal effects can only be estimated through the use of observational data. However, numerous biases will be produced due to the obscurity of the treatment assignment mechanism, as discussed in **Bias**. Therefore, we must de-bias or de-confound the observed data before determining the causal effect, and there are several techniques have been proposed for accomplishing unbiased learning for CI in RS.

1.3.2 Literature Review

CI for RSs has been investigated in order to evaluate biased data objectively. Typically, researchers use a two-phase learning strategy, wherein they first fit a model to estimate a propensity score and then use an unbiased estimator to estimate the causal effect [Li et al., 2022, Liang et al., 2016, Schnabel et al., 2016, Wang et al., 2020]. Schnabel et al. [2016] focuses on proposing an empirical risk minimization (ERM) to address selection bias, i.e., for a visited item, the user prefers to rate it if they like it; otherwise, they will not provide feedback, resulting in the data MNAR problem. This will not be the case in this study, as we presume that users will continue to provide ratings as long as they are exposed to the items. Sato et al. [2020] is similar to Schnabel et al. [2016], which also addresses selection bias and conducts ERM by the unbiased estimator, while using gradient descent to update the unbiased learning, as we will do in our study. However, Sato et al. [2020] used stochastic gradient descent (SGD), whereas we implemented a new algorithm based on the concept of alternating gradient descent (AGD).

Unlike Liang et al. [2016], Wang et al. [2020], Li et al. [2022] emphasized the absence of data. As previously discussed, users have limited access to items, resulting in missing data.

In addition, due to data biases such as popularity bias and exposure bias, individuals can only view a subset of items, leading to data MNAR. In order to overcome this issue, Li et al. [2022] introduces the imputation model. Moreover, instead of pre-selecting and training an algorithm, the authors choose to simultaneously update the propensity, imputation, and outcome models.

Liang et al. [2016] and Wang et al. [2020] are the closest to our research, and both studies fit hierarchical Poisson factorization (HPF) as an exposure model and concentrate on exposure bias, as explained in this article. In the meantime, Liang et al. [2016] attempts to adapt the exposure data to a popularity model. After obtaining the propensity score from the exposure model, the former uses non-negative matrix factorization (NMF) to estimate the causal effect, whereas the latter uses a linear regression model for the task. Combining the concepts of linear regression and NMF, we instead use the NMF as the exposure model. We will continue to use linear regression as the outcome model, but we will incorporate the proposed AGD algorithm to enhance computing efficiency. Nevertheless, rather than declaring the causal effect by indicating how much the rating will grow, both publications only demonstrate an improvement in rating prediction accuracy.

There are also a number of review articles on CI for RSs [Gao et al., 2022, Xu et al., 2023b, Zhu et al., 2023]; each of these articles comprehensively explained the concept of CI for RSs and introduced all potential types of bias and their associated remedies, which can be evaluated in the future.

1.3.3 About this paper

This study attempts to answer the question, “What is the rating difference if a user was exposed to an item v.s. if a user was not exposed to an item.” using music rating data from the Yahoo! R3 dataset [Marlin, 2008, Marlin and Zemel, 2009] and Restaurant & Consumer Data from the UC Irvine Machine Learning (UCI ML) Repository’s [Medelln and Serna, 2012]. Both are observational data; as previously mentioned, observational data lack of randomization in the mechanism for treatment assignment; in this study, we will focus on exposure bias. Similar to Li et al. [2022], we will also include MNAR in the study; however, we will treat missing data as zero and transform the missing data problem into a sparse data problem [Wang et al., 2020].

There will be two learning phases. To de-confound an exposure model, we will initially investigate four distinct algorithms: HPF, logistic regression, NMF, and neural networks (NNs). In order to analyze the causal effect, five additional outcome models will be fitted:

linear regression, IPW, zero-inflated Poisson (ZIP) regression, zero-inflated Negative Binomial (ZINB) regression, and doubly robust (DR) estimation. Even though DR estimation and NNs are common causal inference techniques, they have not yet been implemented for recommender systems. To estimate the parameters when estimating the causal effect, we will also provide a new algorithm based on the concept of AGD. Furthermore, different from the existing papers, in order to better illustrate the causal effect, we will not only present rating prediction accuracy, mean square(d) error (MSE), and standard error (SE) of the estimators, but also estimate by how much the average rating will increase as a result of the application of ATE, demonstrating the advantage of de-biased learning.

Chapter 2 will provide additional information about the datasets, while Chapter 3 will clarify the research assumptions and methodologies. In Chapter 4, we will describe the study process and interpret the findings. And finally, in Chapter 5, we will present a conclusion, limitation, and future direction.

Chapter 2

Data

2.1 About the Data

In this investigation, we will utilize two distinct datasets: music rating data (Dataset I) and restaurant rating data (Dataset II). Furthermore, to better apply an advanced deconfounded method, a NN model, we will expand Dataset II into Dataset III.

Dataset I [Marlin, 2008, Marlin and Zemel, 2009] is a large-scale dataset of song ratings acquired from the Yahoo! R3 website; it contains 1000 users, 15400 songs, and 365704 rating records. However, to cut down the computing cost, we will randomly reduce the number of users and the number of songs to 10% of the original. Following the reduction, Dataset I consists of 100 users, 1540 songs, and 2828 rating records.

Dataset II is a dataset obtained from a prototype recommender system downloaded from the UCI ML Repository. There are 138 users, 130 restaurants, and 1161 ratings on this site. Besides, there are 21 variables to characterize a user, such as personality and dress preferences, and 26 variables to describe each restaurant, such as open hours and location. In order to facilitate causality analysis and better match the user preferences with the restaurant attributes, after carefully reviewing the data, we will only preserve 7 user features and 8 restaurant features. There are 7 standard variables in both datasets: id, smoke (user: whether is a smoker; restaurant: whether there is a smoking area), alcohol (user: whether will drink alcohol or not; restaurant: whether will provide alcohol or not), cuisine type, parking (user: transportation type; restaurant: whether there is a parking lot), payment types, and price (user: budget level; restaurant: price level); and the restaurant contains an additional feature to describe whether it will provide other services or not. Except for id, all of the attributes listed above are categorical.

Dataset III is an expanded version of Dataset II; it comprises all features from the users and restaurant datasets, interactions (except for id) between them, rating scores, and exposure status, for a total of 59 features with 13800 rows. Then, we randomly divided Dataset III into a 7:3 training and testing dataset.

2.2 Data Issue

When reviewing data, we discover that there are two data issues. First, there exists missing data. When data are missing, we will lose some information about the item, resulting in biased and erroneous conclusions, which further erodes statistical power and affects the validity and applicability of the research. If we assert causation based on biased data, the causal RSs will be wrong, proposing unsatisfactory products and creating consumer attrition. Hence, missing data is a pressing issue that must be prioritized in CI for RSs.

Since Dataset I only contains user id, song id, and rating, it is a simple dataset; therefore, we are not concerned about data loss. Although the number of rated cells in a user-item rating matrix is less than 2%, we consider this to be another data problem, Sparse Data. However, data loss occurs in user and restaurant information in Dataset II: for user data, out of 138 users, there are 3 missing values for the smoke variable, 7 missing values for the transportation and the budget variables, and 5 missing values for the payment variable; for restaurant data, out of 130 restaurants, there are 45 missing values for the cuisine variable and 16 missing values again for the payment variable. Multiple imputations (MI) will be used to address the issue; specifics will be detailed in Section 4.1. Again, we will treat absent rating data in Dataset II as a problem involving sparse data.

After imputing the datasets, CI can be applied to determine the causal relationship. First, we will estimate the propensity score by fitting an exposure model, and then we will investigate the causal effect based on the de-confounded data. One of the traditional techniques for estimating propensity scores is logistic regression. Besides, because our exposure variable is binary, we can apply NMF by merely assuming a linear relationship between the exposure data and the user and item characteristics. However, when attempting to calculate the propensity scores, we must also consider the issue of Sparse Data.

Based on the concept of classic RSs, comparable things will continue to appear, making it impossible for a user to view all products. And since only a small fraction of items can be interacted with, few ratings will be recorded. To better distinguish whether a user has already listened to a song or visited a restaurant, we assume that un-visited items have a rating of 0 and visited items have a rating beginning with 0, resulting in a large number of 0s in the user-item rating matrix, which we refer to as Sparse Data [Gao et al., 2022].

A vast number of zeros will have a significant impact on the results, raising concerns about inefficient processing and inaccurate statistical analysis. Fortunately, HPF can alleviate this issue by extracting sparse factors and downplaying the significance of zero. In addition, we will employ NNs to investigate the undetected interaction between users and objects.

After estimating the propensity score, we can use it to examine the causal influence by fitting the outcome model to the data. IPW is a standard estimation approach for causal effects. In addition, we will experiment with linear regression based on the concept of Wang et al. [2020]. Recall that customers can only view a subset of products; since we believe un-visited items have a rating of 0, we suffer from the Sparse Data problem once more. We will apply ZIP regression and ZINB regression to overcome this hurdle.

Chapter 3

Methodologies

The followings are the assumptions and notations that will be applied throughout the study:

Assumption 1 (Stable Unit Treatment Value Assumption, SUTVA) *: There is no communication between users, indicating that they are independent of one another.*

Assumption 2 (Positivity) *: Each subject has a positive chance of being allocated to either the treatment group or the control group so that the treatment effect can be estimated.*

Assumption 3 *: If a user has never seen an item, we presume the rating will be 0; otherwise, the item may have a rating other than zero. For Datset 1, ratings can be* $\{0,1,2,3,4\}$*; for Dataset II and Dataset III, ratings can be* $\{0,1,2\}$*.*

Notation: Consider that there are n users and m items, where $\boldsymbol{a}$ is the exposure data matrix with dimension $n \times m$, and $\boldsymbol{Y}$ is the rating matrix (dimension $n \times m$ for Dataset I and Dataset II, and $(n \times m)$ x 1 for Dataset III). Then, for u=$\{1,...,n\}$ and i=$\{1,...,m\}$, a_{ui} represents whether user u is exposed to item i, and y_{ui} or $y_{ui}(a_{ui})$ represents the rating that item i receives from user u.

The following notations will be introduced for Dataset II and Dataset III only due to the fact that the restaurant dataset contains more information than the Yahoo! R3 dataset. Given that each user u possesses k features and each item i possesses l features:

$\boldsymbol{\theta}$ is a user information matrix extracted from Dataset II with dimension $n \times k$. $\boldsymbol{\mu}$ is an item information matrix extracted from Dataset II with dimension $m \times l$. $\boldsymbol{M}$ is a matrix containing all user information and restaurant information with dimension $(n \times m) \times ((k-1) \times (l-1) + k + l)$, which contains the rating, the user and restaurant information, the interactions between them except the id feature, and the exposure status, and it is extracted from Dataset III with the exception of the exposure column and rating column.

3.1 Propensity Scores Estimation Methods

In this section, we will describe the estimation methods for the propensity scores that will be used in this study, including logistic regression, non-negative matrix factorization, hierarchical Poisson factorization, and neural networks. In addition to defining the models and their coefficients, we will explain why these models will be applied.

3.1.1 Logistic Regression

Logistic regression is one of the classical methods to estimate propensity scores, and its formula can be written as follows:

$$\text{logit}\{e(\boldsymbol{X})\} = \boldsymbol{X}^T\boldsymbol{\beta}, \tag{3.1}$$

where for the Yahoo! R3 dataset, $\boldsymbol{X}$ represents the estimated propensity score with dimension $n \times m$ that is calculated by HPF, which will be explained later. For the restaurant dataset, $\boldsymbol{X}$ represents $\boldsymbol{M}$ (as explained in **Notation**) that contains all user and restaurant information. $\boldsymbol{\beta}$ is the coefficient matrix. After fitting a logistic regression and obtain $\hat{\boldsymbol{\beta}}$ by maximum likelihood estimation, we can estimate $\hat{e}(\boldsymbol{X})$ by:

$$\hat{e}(\boldsymbol{X}) = \text{expit}\{\boldsymbol{X}^T\hat{\boldsymbol{\beta}}\}.$$

3.1.2 Non-Negative Matrix Factorization

If we presume that there is a linear relationship between exposure and user and item characteristics, then the relationship can be expressed as:

$$\boldsymbol{a} = \boldsymbol{\theta} \times \boldsymbol{\beta} \times \boldsymbol{\mu}^T + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{3.2}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ correspond to a $n \times k$ user features matrix and a $m \times l$ item features matrix, respectively. With $\boldsymbol{a}$ is an exposure matrix with $n \times m$ dimensions, $\boldsymbol{\beta}$ is a coefficient matrix with $k \times l$ dimensions.

The fundamental concept of matrix factorization (MF) is to infer user patterns and item patterns from the rating record, i.e., to decompose the rating matrix into a user matrix and an item matrix [Koren et al., 2009]. Therefore, given a user matrix and an item matrix, we should be able to derive a matrix that represents the user's preference for the item. From the preceding, we can deduce that $\boldsymbol{a}$ consists of a user matrix $\boldsymbol{\theta}$, an item matrix $\boldsymbol{\mu}$, an interaction matrix $\boldsymbol{\beta}$, and some error terms $\boldsymbol{\epsilon}$. Even though $\boldsymbol{a}$ is a binary matrix, if a regression is performed on it, $\hat{\boldsymbol{a}}$ should be a matrix with float numbers, and it can be considered as a preference matrix: the larger the number, the greater the likelihood that a user will like an item, and we should expose it. The linear regression problem can therefore be transformed into an MF problem:

$$\boldsymbol{a} \implies \boldsymbol{\theta} \times \boldsymbol{\beta} \times \boldsymbol{\mu}^{\boldsymbol{T}} \tag{3.3}$$

To obtain $\boldsymbol{\beta}$, we can perform a matrix transformation on it. Since $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ are full rank matrices, we can perform a left inverse transformation on $\boldsymbol{\theta}$ and right inverse transformation on $\boldsymbol{\mu}^T$, then we can estimate $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\theta}^{-1} \times \boldsymbol{a} \times (\boldsymbol{\mu}^T)^{-1},$$

where $\boldsymbol{\theta}^{-1}$ is a $k \times n$ matrix, $(\boldsymbol{\mu}^T)^{-1}$ is an $m \times l$ matrix, and $\hat{\boldsymbol{\beta}}$ is a $k \times l$ matrix.

Given $\hat{\boldsymbol{\beta}}$, $e(\boldsymbol{X})$, i.e., $\hat{\boldsymbol{a}}$, can be easily estimated:

$$\hat{e}(\boldsymbol{X}) = \hat{\boldsymbol{a}} = \hat{E}[\boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{\mu}] = \boldsymbol{\theta} \times \hat{\boldsymbol{\beta}} \times \boldsymbol{\mu}^T.$$

In this study, since $\boldsymbol{a}$, $\boldsymbol{\theta}$, and $\boldsymbol{\mu}$ are all non-negative matrices, $\hat{\boldsymbol{\beta}}$ and $\hat{e}(\boldsymbol{X})$ will also be non-negative; hence, we are employing non-negative matrix factorization (NMF) [Lee and Seung, 1999] for estimation.

3.1.3 Hierarchical Poisson Factorization

Hierarchical Poisson factorization (HPF) (or Poisson factorization (PF) in short) [Gopalan et al., 2015] is an extension of NMF which is a technique that models count data as a product of two matrices, while it can model more complex data structures, and its formula is as follows [Wang et al., 2020]:

$$a_{ui}|\boldsymbol{\pi_u}, \boldsymbol{\lambda_i} \sim Poi(\boldsymbol{\pi_u}\boldsymbol{\lambda_i}^T),\ \forall u, i, \tag{3.4}$$

where $\boldsymbol{\pi_u}$ represents the personal characteristic for user u and $\boldsymbol{\lambda_i}$ represents the item attributes for item i. Since Dataset I does not contain any information regarding users or songs, the features of them will be generated at random using a Gamma distribution, following $\boldsymbol{\pi_u} \overset{\text{iid}}{\sim} Gam(c_1, c_2)$ and $\boldsymbol{\lambda_i} \overset{\text{iid}}{\sim} Gam(c_3, c_4)$. To simplify, researchers [Wang et al., 2020, Liang et al., 2016] typically assign $\boldsymbol{\pi_u}$ and $\boldsymbol{\lambda_i}$ the same number of features, h; consequently, $\boldsymbol{\pi_u}$ and $\boldsymbol{\lambda_i}$ are both $h \times 1$ vectors, leading $\boldsymbol{\pi}$ will be a $n \times h$ user feature matrix, and $\boldsymbol{\lambda}$ will be a $m \times h$ song feature matrix.

Once we fit the HPF model and estimate the value of $\boldsymbol{\pi_u}, \boldsymbol{\lambda_i}$ for u=1,...,n and i=1,...,m, we can create a substitute confounder $\boldsymbol{\hat{a}}$ [Wang and Blei, 2019], i.e., $\hat{e}(\boldsymbol{X})$:

$$\hat{e}(\boldsymbol{X}) = \boldsymbol{\hat{a}} = \hat{E}[\boldsymbol{\pi_u}\boldsymbol{\lambda_i}^T|\boldsymbol{a}],$$

where $\boldsymbol{\hat{a}}$ is a $n \times m$ matrix.

3.1.4 Neural Networks

Other than using regression, researchers explore more innovative methods to adapt to more complex data structures [Westreich et al., 2010]. Considering the data complexity and computational efficiency in this study, we choose neural networks (NNs) as another estimation method (Figure 3.1) for the estimation of propensity scores. NN is a machine learning algorithm, and it aims to imitate how the human brain works [Clothiaux and Bachmann, 1994]. It consists of a large number of interconnected nodes, known as neurons, organized into layers. There are three types of layers: input layer, for receiving input; output layer, for returning results; and hidden layer, all layers that exist between the input and output layers.

The value of each unit in a layer except the input layer will consist of the units from the last layer. Therefore, after the input layer catches the input, we will apply activation functions to generate the input for the first neuron in the next layer, where each unit in the input layer will be assigned different weights, and we will repeat this step until we calculate all neurons for the next layer. Then we will follow the same procedure until we reach the output layer, which will return the final result for the model.

From Figure 3.1, we can see that there are three layers for our model with only one hidden layer. We will first pass all possible covariates as the input, which means the number of neurons in the input layer equals the number of covariates; then, we will use an activation function on them to create the neurons in the hidden layer. After that, we will apply another activation function to generate the unit in the output layer, which will be returned as a final result of the model, i.e, the predicted propensity score when given user and restaurant characteristics.

$$z_i = w_{0i} + w_{1i}x_1 + w_{2i}x_2 + ... + w_{ni}x_n, \quad \forall i \in \{1, ..., k\} \tag{3.5}$$

$$\hat{e}(\boldsymbol{X}) = w'_0 + w'_1 z_1 + w'_2 z_2 + ... + w'_k z_k \tag{3.6}$$

3.2 Causal Effect Estimation Methods

In this section, we will explain the models we use to estimate causal effects, such as linear regression, zero-inflated Poisson regression, zero-inflated Negative Binomial regression, inverse probability weighting, and doubly robust estimation; not only the detailed explanation of the principle but also the clarification of how we will adjust the models to suit the data better.

3.2.1 Linear Regression

In linear regression, ATE can be easily demonstrated as:

$$\boldsymbol{Y} = ATE \times \boldsymbol{a} + \gamma \times e(\boldsymbol{X}) + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{3.7}$$

where $\boldsymbol{Y}$ is the rating matrix; $\boldsymbol{a}$ is the exposure matrix; and γ is the scalar adjustment of $e(\boldsymbol{X})$, or, how much the propensity score will contribute to the ratings.

To align with Dataset I, the model will be reformulated as the following [Wang et al., 2020]:

$$y_{ui}(a_{ui}) = \boldsymbol{\theta_u}^T \boldsymbol{\beta_i} \times a_{ui} + \gamma_u \times e(\boldsymbol{X})_{ui} + \epsilon_{ui},\ \epsilon_{ui} \sim N(0, \sigma^2), \tag{3.8}$$

where γ_u has the same definition as γ in (3.7), but it's a user-specific coefficient, so different users have a different adjustment value; $e(\boldsymbol{X})_{ui}$ is the propensity score for user u to item

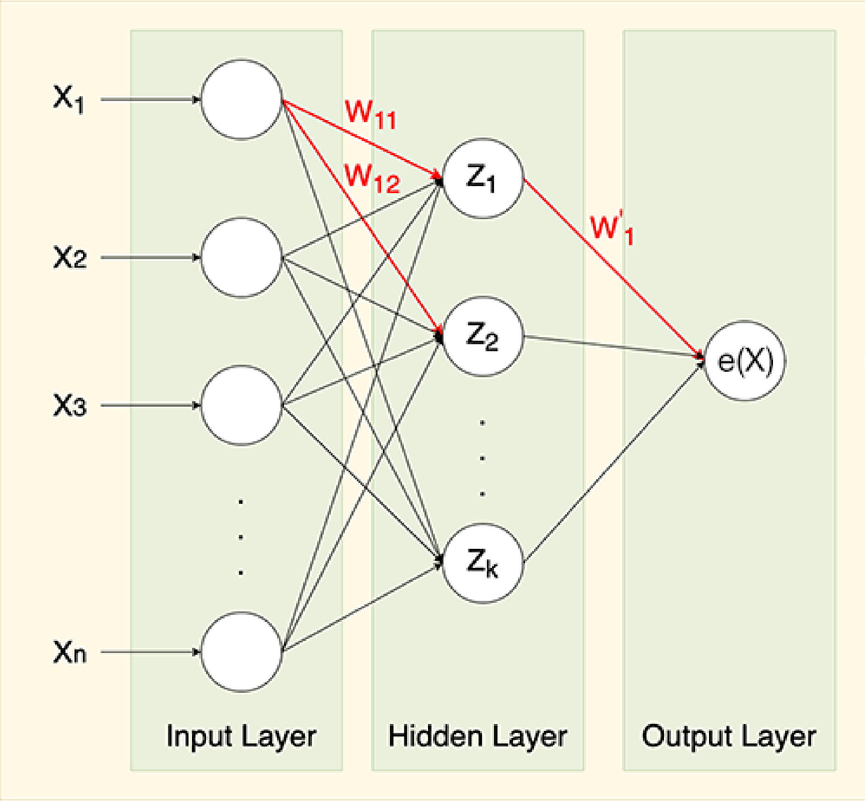


Figure 3.1: Neural Network for Propensity Score Estimation
$X_i,\ \forall i = \{1, ..., n\}$: represents the covariates as the inputs of the model
$Z_i,\ \forall i = \{1, ..., k\}$: represents the neurons in Hidden layer, $k \neq n$
$\hat{e}(\boldsymbol{X})$: represents the estimated propensity score
w_{1i}, or w^prime_i: represents the weight of the neurons

i; $\boldsymbol{\theta_u}^T\boldsymbol{\beta_i}$ is the product representing an expected ITE that is specific to every combination of every user u and every item i.

$\boldsymbol{\gamma_u}$, $\boldsymbol{\theta_u}$, and $\boldsymbol{\beta_i}$ are approximated via maximizing a posterior (MAP) estimation in investigation [Wang et al., 2020]. Alternatively, we propose a new algorithm based on the idea of alternating gradient descent (AGD) [Jain et al., 2013]. AGD is an iterative optimization procedure that can alternatively update several parameters simultaneously. Compared with standard gradient descent (SGD), AGD can provide better convergence properties, especially in high-dimensional problems, resulting in more compatible solutions than MAP as $\boldsymbol{\theta_u}$ and $\boldsymbol{\beta_i}$ can be multi-dimensional, and it can save more computational workload. Furthermore, instead of updating parameters by a given learning rate value, we will provide a list of learning rates, and the algorithm will decide the one that will optimize the accuracy the best during each iteration after the comparison.

We first convert $\boldsymbol{Y} = \boldsymbol{\theta}^T\boldsymbol{\beta} \times \boldsymbol{a} + \boldsymbol{\gamma} \times e(\boldsymbol{X}) + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ into an optimization problem:

$$\min_{\boldsymbol{\theta_u}, \boldsymbol{\beta_i}, \gamma_u} \sum_{u=1}^{n} \sum_{i=1}^{m} ||y_{ui}(a_{ui}) - \boldsymbol{\theta_u}^T\boldsymbol{\beta_i} \times a_{ui} - \gamma_u \times e(\boldsymbol{X})_{ui}||^2,$$

and we can derive the updated function for each parameter from it. Then at iteration t:

$$\begin{aligned}\boldsymbol{\theta_u}^{(t)} \leftarrow \boldsymbol{\theta_u}^{(t-1)} - lr \times 2 \times \sum_{i=1}^{m}(y_{ui}(a_{ui}) - \boldsymbol{\theta_u}^{(t-1)T}\boldsymbol{\beta_i}^{(t-1)} \times a_{ui} - \gamma_u^{(t-1)} \times \\ e(\boldsymbol{X})_{ui}) \cdot (-a_{ui} \times \boldsymbol{\beta_i}^{(t-1)}),\end{aligned} \tag{3.9}$$

$$\begin{aligned}\boldsymbol{\beta_i}^{(t)} \leftarrow \boldsymbol{\beta_i}^{(t-1)} - lr \times 2 \times \sum_{u=1}^{n}(y_{ui}(a_{ui}) - \boldsymbol{\theta_u}^{(t)T}\boldsymbol{\beta_i}^{(t-1)} \times a_{ui} - \gamma_u^{(t-1)} \times \\ e(\boldsymbol{X})_{ui}) \cdot (-a_{ui} \times \boldsymbol{\theta_u}^{(t)T}),\end{aligned} \tag{3.10}$$

$$\begin{aligned}\gamma_u^{(t)} \leftarrow \gamma_u^{(t-1)} - lr \times 2 \times \sum_{i=1}^{m}(y_{ui}(a_{ui}) - (-e(\boldsymbol{X})_{ui}) \times (\boldsymbol{\theta_u}^{(t)T}\boldsymbol{\beta_i}^{(t)}) \times a_{ui} - \\ \gamma_u^{(t-1)} \times e(\boldsymbol{X})_{ui}),\end{aligned} \tag{3.11}$$

where $e(\boldsymbol{X})_{ui}$ represents the propensity score for the user u and item i pair, and lr represents the learning rate; pseudocode is demonstrated in Algorithm 1.

3.2.2 Zero-Inflated Poisson Regression

As we informed previously, exposure data will mainly consist of 0, meaning the rating data will also include a high proportion of 0, which may impact the causal effect analysis. To better predict a user's rating when he or she is exposed to an item, researchers introduced zero-inflated Poisson (ZIP) regression [Lambert, 1992], which can manage the situation that we have more 0 than expected. ZIP contains two steps: Step 1 is to distinguish always 0 (i.e., an item that should not be exposed to a user) and not-always 0 (i.e., an item that should be exposed to a user) from data that are 0 by a logistic regression. Then in Step 2 we can consider data that are not-always 0 and not 0 as a new dataset, and we

Algorithm 1 Alternating Gradient Descent

```
  Input: None
  Output: User latent factors $\boldsymbol{\theta}_{1:U}$, item latent factors $\boldsymbol{\beta}_{1:I}$, and adjustment factors $\boldsymbol{\gamma}_{1:U}$
  Fit the outcome model to compute the rating
  Randomly initialize $\boldsymbol{\theta}_{1:U}, \boldsymbol{\beta}_{1:I}, \boldsymbol{\gamma}_{1:U}$
  $LEARNING\_RATE$ = [1e-2, 1e-3, 1e-4]
  while not converged do
    for $lr \in LEARNING\_RATE$ do
      for $u \leftarrow 1$ to $U$ do
        Update user factor $\boldsymbol{\theta}_{lr,u}^{(t)}$ (3.9)
      end for
    end for
    compare $\boldsymbol{\theta}_{lr}^{(t)}$ for each $lr$, choose the one reaches the best approximation as $\boldsymbol{\theta}^{(t)}$
    for $lr \in LEARNING\_RATE$ do
      for $i \leftarrow 1$ to $I$ do
        Update item factor $\boldsymbol{\beta}_{lr,i}^{(t)}$ (3.10)
      end for
    end for
    compare $\boldsymbol{\beta}_{lr}^{(t)}$ for each $lr$, choose the one reaches the best approximation as $\boldsymbol{\beta}^{(t)}$
    for $lr \in LEARNING\_RATE$ do
      for $u \leftarrow 1$ to $U$ do
        Update scale factor $\gamma_{lr,u}^{(t)}$ (3.11)
      end for
    end for
    compare $\gamma_{lr}^{(t)}$ for each $lr$, choose the one reaches the best approximation as $\boldsymbol{\gamma}^{(t)}$
  end while
  return $\boldsymbol{\theta}_{1:U}, \boldsymbol{\beta}_{1:I}, \boldsymbol{\gamma}_{1:U}$
```

assume it follows a Poisson distribution since the higher the rating, we should expect the smaller the count, and we can write it as [Hall, 2000]:

$$Y_{ui} = \begin{cases} 0, & \text{with probability } p_{ui} + (1 - p_{ui}) \times e^{-\lambda_{ui}} \\ k, & \text{with probability } (1 - p_{ui}) \times \frac{e^{-\lambda_{ui}} \lambda_{ui}^k}{k!} \end{cases},$$

where p_{ui} is the probability that user u gives item i a rating of 0 (i.e., will not show item i to user u); λ_{ui} is a parameter for the Poisson distribution; k is a score that user u will rate item i.

3.2.3 Zero-Inflated Negative Binomial Regression

Zero-inflated Negative Binomial (ZINB) regression is slightly different from ZIP: it considers that each observation has a probability of occurring or not occurring. In our study, it means each item has a probability of exposing or not exposing to a user; and if we should expose the item, it will approximate a Negative Binomial (NB) distribution [Yusuf et al., 2017]:

$$Y_{ui} \begin{cases} = 0, & \text{with probability } p_{ui} \\ \sim NB(\lambda_{ui}, k), & \text{with probability } (1 - p_{ui}) \end{cases},$$

where p_{ui} is the probability that an item i will expose to user u; λ_{ui} is a parameter for the NB distribution; k is the sparse factor or the over-dispersion parameter for the NB distribution.

3.2.4 Inverse Probability Weighting

Inverse probability weighting (IPW) [Robins et al., 2000] is an approach to de-bias, more specifically, to adjust the selection bias in observational studies, and aims to assign different weights to different subjects. After calculating the propensity score $e(\boldsymbol{X})$, IPW will assign $\frac{1}{e(\boldsymbol{X})}$ as the weight for the treatment group (i.e., users who are exposed to the item) and $\frac{1}{1-e(\boldsymbol{X})}$ as the weight for the control group (i.e., users those are not exposed to the item) [Hernan and Robins, 2023, Chesnaye et al., 2021], then ATE can be evaluated by:

$$\widehat{ATE} = \frac{1}{n} \times \sum_{i=1}^{n} \Big[\frac{a_i \times Y_i}{\hat{e}(\boldsymbol{X}_i)} - \frac{(1 - a_i) \times Y_i}{1 - \hat{e}(\boldsymbol{X}_i)} \Big], \tag{3.12}$$

where Y_i is the observed outcome of treatment a_i, and $\hat{e}(\boldsymbol{X}_i)$ is the corresponding estimated propensity score.

3.2.5 Doubly Robust Estimation

In observational studies, treatments are not randomly assigned; to simulate a randomized trial so that we can infer causality, researchers typically estimate propensity scores and then investigate causal relationships based on these scores. This procedure is heavily reliant on the accuracy of the propensity score estimation method and the causal effect evaluation method; to correctly determine the causal effect, the underlying assumptions of these algorithms must be precise. Nonetheless, the doubly robust (DR) estimator [Robins et al., 1994] can correctly estimate causal effects if either of the two model assumptions holds, but not both, making it a flexible and reliable statistical method for estimating ATE. DR incorporates the models for the outcome variable and the exposure variable, and it can be written as follows [Funk et al., 2011]:

$$\begin{aligned}\widehat{ATE} = \frac{1}{n} \times \sum_{i=1}^{n} \Big[&\frac{a_i \times Y_i - (a_i - \hat{e}(\boldsymbol{X}_i)) \times \hat{m}_1(\boldsymbol{X}_i)}{\hat{e}(\boldsymbol{X}_i)} \\ &- \frac{(1-a_i) \times Y_i - (a_i - \hat{e}(\boldsymbol{X}_i)) \times \hat{m}_0(\boldsymbol{X}_i)}{1 - \hat{e}(\boldsymbol{X}_i)}\Big],\end{aligned} \tag{3.13}$$

where Y_i is the observed outcome for subject i of treatment a_i; $\boldsymbol{X}_i$ is the vector of covariates; $\hat{e}(.)$ is the estimator of propensity score; and $\hat{m}_t(.)$ is the estimator of the outcome model for treatment group $t = 1$ and control group $t = 0$, respectively.

However, this investigation will interpret the ATE based on the estimated exposure coefficient value rather than the above equation. A linear regression will be utilized as the outcome model, which can be expressed as follows:

$$\boldsymbol{Y} = \beta_0 + ATE \times \boldsymbol{a} + \boldsymbol{\beta} \times \boldsymbol{X} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}), \tag{3.14}$$

where β_0 is the intercept value; ATE is the coefficient of exposure matrix $\boldsymbol{a}$ indicating the average rating change when the exposure increases from 0 to 1; $\boldsymbol{\beta}$ is the coefficient vector representing how the covariates can affect the final rating; and $\boldsymbol{X}$ is a matrix containing all covariates information, which is $\boldsymbol{M}$ mentioned in **Notations**. The above model will be fitted with the propensity score-based weights. We will assign weight $\frac{1}{e(\boldsymbol{X})}$ to the treatment group and weight $\frac{1}{1-e(\boldsymbol{X})}$ to the control group, in which $e(\boldsymbol{X})$ can be estimated by NMF, logistic regression, and NNs; and again, we will conduct 99% quantile trim on weights to prevent outlier effect.

Chapter 4

Data Analysis

4.1 Data Processing

As discussed in Section 2, the restaurant dataset contains missing data, which can lead to biased estimates and decreased statistical power, as well as an inability to accurately depict the relationship among user characteristics, restaurant characteristics, and ratings, thereby preventing the inference of causality. Therefore, this obstacle must be eliminated prior to the implementation of CI.

Before addressing the missing data issue, it is necessary to identify the type of missing data. There are three different types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In this study, we simply assume that the probability of missingness is dependent on the observed variables and conclude that it is a MAR type [Rubin, 1976].

After identifying the type of missing data, the issue can be resolved in one of two ways: either by removing incomplete records or by replacing them with new values. Since the absent data for each feature could be contained in a separate row, deleting all rows simultaneously would result in a greater data loss than would otherwise be the case. In addition, there is a data gap of at least $\frac{1}{3}$ for the cuisine variable in the restaurant information. If the relevant rows are eliminated immediately, the data will be reduced to less than $\frac{2}{3}$ of its original size, jeopardizing the accuracy and dependability of this study. As a result, imputation will be used to address the issue of missing data so that as much data as feasible is preserved.

In this investigation, MI [Schafer, 1999] will be used to fill in the missing values. This will result in the production of multiple plausible imputations for the missing value, which will be analyzed and combined to produce estimates that account as precisely as possible for the uncertainty associated with the missing data. As Dataset II is the foundation for Dataset III, it is unnecessary to revisit the issue of missing data for Dataset III. Now that the datasets are complete, we can investigate the causality.

4.2 Method Application

A randomized trial is necessary for establishing causality because it allows for the random presentation of treatments to participants. However, this is expensive in terms of RSs in general, and we can only conduct observational studies. Therefore, we must eliminate any bias from the observed data so that each user has the same probability of receiving the treatment, also known as propensity scores.

Logistic regression, one of the conventional methods for estimating propensity scores, can simply restrict the output to values between 0 and 1. In addition, assuming there is a linear relationship between the exposure data and the user and item information, the coefficient can be reformulated as a combination of exposure data, user characteristics, and item characteristics. According to the structure of Dataset II, all explanatory variables are presented in metrics, indicating that the estimated coefficient is also a matrix that can be decomposed into multiple other matrices. Therefore, NMF will be utilized to calculate the propensity scores.

Since the exposure data is binary, it is straightforward to assume the exposure variable follows a Bernoulli distribution. However, when applied to sparse data, the zeros will have a substantial impact on the estimates. On the other hand, few rating data are available, and insufficient data will also result in a squandering of computer resources and a reduction in precision. Due to the aforementioned issues, the Bernoulli distribution will not be considered. However, this issue can be resolved by HPF's capabilities, which include the ability to detect sparse patterns and minimize the impact of zero.

Wang et al. [2020] uses HPF to construct a substitute confounder by estimating the user and item attributes, and then uses Backdoor Adjustment to compute the causal effect on the de-confounded data. This method is effective for evaluating the causal effects of the majority of methodologies; though, its application to IPW will result in complications. In IPW, a weight of $\frac{1}{e(\boldsymbol{X})}$ will be allocated to the treatment group, and $\frac{1}{1-e(\boldsymbol{X})}$ will be assigned to the control group; examining the estimated values of the substitute confounder reveals numbers greater than 1, resulting in a negative weight for the control group. To ensure that the propensity scores remain strictly within the range of 0 to 1, the true propensity scores must be calculated using substitute confounders that are generated by HPF. In order to expedite the process, logistic regression will be utilized once more.

In addition, a NN model, designated [Westreich et al., 2010] capable of identifying latent patterns in complex data, will be implemented. This model will aid in documenting the relationships between users and objects in order to recommend products that match the preferences of users. According to the definition of a NN model, all conceivable covariates must be included in its input. In addition, unlike other propensity score estimation methods that can pass user information, restaurant information, and user-restaurant rating matrix separately, a NN can only be furnished with a single dataset, so a new dataset containing all of the information is requisite. Consequently, Dataset III is constructed to accommodate this circumstance.

After de-confounding the data and calculating the propensity scores, it is possible to estimate the causal effect. IPW is a conventional method for determining causal effects. Nonetheless, since the weight is related to the reciprocal of the propensity scores, if the value is too large or too small, one divided by the probability will result in an excessive number, which can have a disproportionate impact on the estimated treatment effect. To minimize the effect of these outlier observations, we will perform weight trimming by setting the 99% quantile as the upper limit. Linear regression will also be tested based on the study of Wang et al. [2020]. Wang et al. [2020] attempts to calculate the coefficients using the MAP estimation procedure prior to fitting the outcome model. In this study, however, we introduce a new algorithm inspired by AGD that will modify each variable one at a time using an alternating learning rate. This means that instead of providing a single value for the learning rate, a list of prospective values will be given, and the algorithm will determine which one optimizes accuracy the best before updating the variable accordingly. After the first variable has been modified, the second variable will be computed based on the amended one. Since this algorithm is implemented specifically for estimating the coefficients of linear regression, complete details are provided in subsubsection 3.2.1.

Given that our outcome data will also include a considerable number of zeros, the issue of sparse data resurfaces. As a solution, we will utilize ZIP and ZINB, which are appropriate for situations involving an excessive number of zeros. They can distinguish between always 0 and not-always 0; this means that in the current rating record, for those ratings that are zero, ZIP and ZINB can determine which user-item interaction has the potential to have a rating greater than zero (items that should be exposed to users, i.e., users will like the items) and which user-item interaction will always be zero (items that should not be exposed to users, i.e., users will not be interested in the items).

Due to the fact that the precision of the exposure model, also known as the estimation methods for propensity scores, will affect the precision of the outcome model, also known as the causal effect estimation methods, which means that incorrectly estimating the propensity score will damage the causal effect conclusion, DR estimation will be applied to mitigate bias and enhance the effectiveness of estimating treatment effects in observational studies.

Because of the disparity between the two datasets, HPF and logistic regression applied on HPF (Logit-HPF) will be tested on the music dataset, whereas NMF, logistic regression (Logit), and NN will be tested on the other. All of the causal effect estimation methods will be applied to both datasets, with the exception of DR, which will be tested only on the restaurant dataset due to the lack of additional information. Furthermore, recall that both HPF and Logit-HPF will yield estimated propensity scores for Dataset I, and the objective of executing a logistic regression, which is a uniquely parameterized propensity function, on the substitute confounder is to obtain values that are between 0 and 1. According to the definition of Assumption 3 in Imai and Van Dyk [2004], the results from HPF and Logit-HPF are logically equivalent; consequently, causal effect estimation methodologies can be applied to either propensity score result and conclude the same causality. To expedite the process, IPW will be only evaluated using the propensity scores generated by Logit-HPF, while the remaining algorithms will be evaluated using the substitute confounder generated by HPF.

4.3 Results

This section will present the results and evaluate each algorithm's performance. In the 'Dataset' column of the following tables, 'Music' corresponds to Dataset I. However, for 'Restaurant', if the exposure model is NMF, Dataset II is used because NMF decomposes the rating matrix into a user info matrix and an item info matrix, whereas Dataset II consists of both matrices; otherwise, Dataset III is used, which contains all information in a single matrix. All values shown in the table will be rounded to four digits.

The accuracy and MSE for each estimation method of the propensity score are displayed in Table 4.1. To compute the accuracy, we will first round the estimated propensity score to obtain the predicted exposure status, followed by the comparison of the predicted status with the actual status. Mean square(d) error (MSE) is a traditional method for evaluating the efficacy of a model. It is computed by averaging the squared differences between the predicted values and the actual values, which are the estimated propensity score and the true exposure; a lower MSE indicates that the predicted values are closer to the actual values, thereby demonstrating a more reliable model. For the 'Music' data, HPF can

Dataset	Method	Accuracy	MSE
Music	HPF	96.5182%	0.0371
	Logit-HPF	98.5221%	0.0146
Restaurant	NMF	93.5284%	0.0604
	Logit	93.7384%	0.0592
	NN	93.7386%	0.0587

Table 4.1: Accuracy and Mean Squared Error for Exposure Model

HPF stands for hierarchical Poisson factorization, Logit for logistic regression, Logit-HPF for implementing Logit after HPF, NMF for non-negative matrix factorization, and NN for neural network.

attain an accuracy of 96.5%. As previously discussed, HPF attempts to discover the latent attributes of users and items from the rating record in order to make future predictions and recommendations. However, the inner product of users and items features may be greater than 1, leading to a negative weight in IPW. To solve this issue, we will apply a logistic regression on the substitute confounder generated by HPF to obtain the propensity scores that are strictly between 0 and 1, denoted in the table as ‘Logit-HPF’, which can accomplish approximately 3% better accuracy and higher stability with lower MSE. For ‘Restaurant’, NMF achieves the lowest accuracy, while NN achieves the highest, because NMF can only conduct simple matrix calculations, whereas NN can discover hidden patterns and then predict the probability. Despite having the highest accuracy, NN also achieves the lowest MSE, indicating that it is closest to the actual value than other models, and it is the most accurate predictor of the outcome.

Accuracy and MSE are reported for each causal effect estimation algorithm in Table 4.2, and the corresponding average treatment effect and SE are described in Table 4.3. Similar to the evaluation of exposure models, the actual ratings are categorical with five levels, whereas the predicted ratings are continuous, then the accuracy will be determined by comparing the rounded predictions with the actual ones. In order to calculate the MSE, we will again just simply average the squared errors between the predictions and the true value. In Table 4.3, for ‘Music’, the predicted ratings generated by IPW based on the estimated propensity score produced by ‘Logit-HPF’ still provide the highest accuracy, up to 98.4%; however, the precision for each outcome model does not differ significantly, while others provide only half the MSE as compared to IPW’s. A possible explanation is that, HPF and logistic regression are two propensity score estimation approaches that are employed sequentially to process the estimated propensity score ’Logit-HPF’ that will be utilized to assess the causal effect. To attain a higher accuracy during the second propensity score calculation, the logistic regression may capture more patterns, even those that are

Dataset	Exposure Model	Outcome Model	Accuracy	MSE
Music	HPF	Linear	98.3766%	0.0472
		ZIP	98.3788%	0.0469
		ZINB	98.3788%	0.0469
	Logit-HPF	IPW	98.4827%	0.1626
Restaurant	NMF	DR	96.0981%	0.0422
		Non-DR	96.0981%	0.0363
	Logit	DR	96.1724%	0.0458
		Non-DR	96.0981%	0.0363
	NN	DR	96.0981%	0.0448
		Non-DR	96.0981%	0.0363

Table 4.2: Accuracy and Mean Squared Error for Outcome Model

DR stands for doubly robust estimation, and Non-DR refers to all other methods besides DR, such as linear regression, inverse probability weighting, zero-inflated Poisson regression, and zero-inflated Negative Binomial regression.

not necessary. Table 1 shows that this was indeed the case. Recall that a propensity score estimation is also a dimension reduction strategy because it may combine multiple features into a single scalar number, which will inevitably fit on the extraneous patterns, resulting in overfitting. Since we round the predictions first and then compare them with the actual ratings, this will not be a big problem when we try to determine the accuracy. However, the error continues to aggregate when we calculate MSE, which will directly compare the predictions and the actual ratings, resulting in a high MSE while high accuracy as well.

For ‘Restaurant’, all ‘Non-DR’ methods, i.e., linear regression, IPW, ZIP, and ZINB, provide comparable accuracy and MSE, regardless of the type of exposure models; this occurs because the predictions are close to one another, with absolute values no greater than 0.03, resulting in the same rounded values, and thus the same accuracy. ‘Logit-DR’, which refers to logistic regression as the exposure model and DR as the outcome model, achieves the highest accuracy with 96.1724% and the greatest MSE. ‘NMF-DR’ and ‘NN-DR’ yield the same accuracy as ‘Non-DR’ but with inferior generalization.

Table 4.3 presents the estimated average treatment effect (represented by ‘ATE’) and SE for causal effect estimation methods. The values in the ‘ATE’ represent the average change in rating if we assign another treatment value. For example, the 1.6141 in the first line of the table stands for if we decide to expose an item to a user (treatment value changes from 0 to 1), on average, we should anticipate a 1.6141 increase in rating, which

Dataset	Exposure Model	Outcome Model	ATE	SE	Pr(¿| t |)
Music	HPF	Linear	1.6141	0.0050	≈ 0
	Logit-HPF	IPW	1.6006	0.0263	≈ 0
Restaurant	NMF	Linear	1.1963	0.0072	≈ 0
		IPW	1.1891	0.0099	≈ 0
		DR	1.1951	0.0092	≈ 0
	Logit	Linear	1.1975	0.0072	≈ 0
		IPW	1.1973	0.0096	≈ 0
		DR	1.1976	0.0095	≈ 0
	NN	Linear	1.1965	0.0072	≈ 0
		IPW	1.1960	0.0097	≈ 0
		DR	1.1985	0.0098	≈ 0

Table 4.3: Average Treatment Effect and Standard Error

means we predict the user will rate 2 (out of 4, a categorical variable) as we will only show unseen items and we assume them with rating 0 before recommending. Overall, the causal RS could have a positive impact on the 'Music' data, increasing the average rating by 1.6 (out of 4); additionally, Pr($> |t|$) is close to zero, indicating the significance of the exposure variable. For 'Restaurant', the table illustrates that if an item is exposed to a user, the average rating should grow by approximately 1.2 (out of 2), with DR showing the most significant effect most of the time and linear regression having the second highest ATE and the lowest SE. In contrast, IPW exhibits the smallest causal effect and largest SE across all categories of the exposure model. Despite this, they all emphasize the importance of exposure variables. Notice that results for zero-inflated models are excluded from the table as a consequence of the dataset's complete separation. This study aims to investigate the causal effect of exposure, i.e., the change in rating if an item is exposed to a user; thus, exposure is a crucial explanatory variable. On the assumption that an unseen object can only have a rating of 0 and a visited item can have a rating between 0 to 2, there will be a complete separation, as exposure 0 only corresponds to rating 0. Therefore, a zero-inflated model cannot determine a rating other than zero for the control group when applied, leading it to be unable to distinguish between always 0 and not always 0, resulting in a convergence issue. Consequently, all SE and p-values for models with zero inflation are omitted from the table.

Chapter 5

Conclusion, Limitation, and Future Direction

5.1 Conclusion

In this paper, we investigated different algorithms to remove the confounding effect of the CI for RSs. We first used logistic regression and NMF, these two classical methods, to estimate the propensity scores. Meanwhile, we built a NN to better capture the patterns between users and items. For the sake of Sparse Data, we also implemented HPF to de-emphasize the effects of excessive zeros in the data. According to our data, if a user's rating is zero, they either do not like the item or will like it if exposed to it; however, if the rating is not zero, the user is intrigued by the item. HPF benefits from non-zero ratings to discover latent factors to allow user-item pairs to be more similar, as opposed to allowing the pair to be less similar based on the zero rating, allowing it to struggle less with zero data and increase computing efficiency.

After we estimate the propensity scores, we can proceed with the estimation of the causal effect. In addition to the traditional approach of IPW, based on the idea of Wang et al. [2020], we also attempted linear regression. But, instead of using MAP estimation as they stated in their study, we proposed a new update algorithm based on the concept of AGD, which can update variables alternatively with an unfixed learning rate. Since Sparse Data also exists in our outcome model, to retrieve the sparse factor and increase the accuracy of statistical analysis, we applied ZIP And ZINB, which are more suitable to use when there are over-abundant zeros. Our research shows that, after de-biasing the data, the causal RS results lead to an average of 1.6 points (out of 4) higher in the user-item rating for the Yahoo! R3 dataset, and 1.2 points (out of 2) higher for the restaurant data.

5.2 Limitation and Future Direction

The present study has limitations that must be acknowledged cautiously. First, despite using two datasets and zero-inflated models, which are designed to handle data with an excess of zero counts, to maintain validity and reliability, our models are still affected by over-dispersion issues, indicating excessively high accuracy, as a result of the presence of too many zeros (i.e., missing rating value) in our data. The second limitation derives from the assumptions made regarding the missing value; we consider the missing data to be MAR. Nonetheless, the sensitive character of certain information may indicate that the missing category is MNAR. People who smoke, for instance, prefer not to answer the smoking question because they are aware that smoking is an unhealthy habit. Therefore, MI may not be able to resolve this issue effectively. Finally, the causal effect concludes that, for the restaurant data, if we recommend an item to a user, it may receive an average rating score of 1.2 points higher (out of 2) than if it were not recommended. However, this prediction remains unvalidated due to the absence of actual feedback to corroborate the predicted outcome, and the practical implications of these results should be interpreted with caution until additional empirical evidence can be gathered to support these assertions.

In the future, in order to handle sparse data, we could reapply this de-confounding framework to a dataset with more rating data and then employ zero-inflated models once more. Concerning the missing data type, we could provide a sensitivity analysis and additional techniques, including pattern-mixture models and informative missingness methods. Lastly, the unvalidated prediction could be addressed by launching an updated causal RS with a built-in feedback collection mechanism to collect actual user responses, thereby providing empirical evidence to support the model's predictive capabilities and enhancing the results' credibility and practical utility. In addition, although we tested nine algorithms in this study, the field of machine learning is constantly evolving, and more advanced and unbiased methods are continuously arising and can be evaluated based on survey papers from Gao et al. [2022], Xu et al. [2023b], Zhu et al. [2023].

References

Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

Robin Burke. Knowledge-based recommender systems. *Encyclopedia of Library and Information Systems*, 69(Supplement 32):175–186, 05 2000.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

Nicholas C Chesnaye, Vianda S Stel, Giovanni Tripepi, Friedo W Dekker, Edouard L Fu, Carmine Zoccali, and Kitty J Jager. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, 15(1):14–20, 08 2021. ISSN 2048-8505.

Eugene E. Clothiaux and Charles M. Bachmann. Neural Networks and Their Applications. In Bruce C. Hewitson and Robert G. Crane, editors, *Neural Nets: Applications in Geography*, pages 11–52. Springer Netherlands, Dordrecht, 1994. ISBN 978-94-011-1122-5.

James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Fourth ACM Conference on Recommender Systems*, pages 293–296, 09 2010.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 03 2011. ISSN 0002-9262.

Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal inference in recommender systems: A survey and future directions. *arXiv preprint arXiv:2208.12397*, 2022.

Enrique García, Cristóbal Romero, Sebastian Ventura, and Carlos Castro. An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Model. User-Adapt. Interact.*, 19, 07 2008.

Enrique García, Cristóbal Romero, Sebastian Ventura, and Carlos Castro. An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Model. User-Adapt. Interact.*, 19:99–132, 2009.

David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 12 1992. ISSN 0001-0782.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19, 01 2016. ISSN 2158-656X.

Prem Gopalan, Jake M. Hofman, and David M. Blei. Scalable recommendation with poisson factorization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, page 326–335, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Daniel B Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, 2000.

M.A. Hernan and J.M Robins. *Causal Inference: What If*. Taylor & Francis, 2023. ISBN 9781420076165.

Miguel A Hernán and James M Robins. Causal inference, 2010.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99 (467):854–866, 2004.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. ISBN 0521885884.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pages 665–674, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Haoxuan Li, Chunyuan Zheng, Xiao-Hua Zhou, and Peng Wu. Stabilized doubly robust learning for recommendation on data missing not at random, 2022.

Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI*, 2016.

Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert systems with applications*, 41(4):2065–2073, 2014.

Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

Benjamin M Marlin. *Missing Data Problems in Machine Learning*. Ottawa : Library and Archives Canada = Bibliotheque et Archives Canada, [2010], 2008. ISBN 9780494578988.

Benjamin M Marlin and Richard S Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 5–12, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584355.

Rafael Medelln and Juan Serna. Restaurant consumer data, 2012.

Michael J Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, 13:393–408, 1999.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 12 1995. ISSN 0006-3444.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 01 2009a.

Judea Pearl. *Causality.* Cambridge University Press, Cambridge, UK, 2009b.

Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40 (3):56–58, 03 1997. ISSN 0001-0782.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444.

Masahiro Sato. Online evaluation methods for the causal effect of recommendations. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 96–101, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582.

Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. Unbiased learning for the causal effect of recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 378–387, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832.

J Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 158–166, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131763.

Joseph L Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15, 1999.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, number 10, pages 1670–1679. JMLR.org, 2016.

Poonam B Thorat, Rajeshwari M Goudar, and Sunita Barve. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 110(4):31–36, 01 2015.

Robin Van Meteren and Maarten Van Someren. Using content-based filtering for recommendation. In *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 workshop*, volume 30, pages 47–56. Barcelona, 2000.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 426–431, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832.

Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1791–1800, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325.

Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, 08 2010.

Shuyuan Xu, Juntao Tan, Shelby Heinecke, Jia Li, and Yongfeng Zhang. Deconfounded causal collaborative filtering. *arXiv preprint arXiv:2110.07122*, 2021.

Shuyuan Xu, Yingqiang Ge, Yunqi Li, Zuohui Fu, Xu Chen, and Yongfeng Zhang. Causal collaborative filtering. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 8 2023a.

Shuyuan Xu, Jianchao Ji, Yunqi Li, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. Causal inference for recommendation: Foundations, methods and applications. *arXiv:2301.04016*, 1 2023b.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15 (5):1–46, 10 2021. ISSN 1556-4681.

OB Yusuf, T Bello, O Gureje, et al. Zero inflated poisson and zero inflated negative binomial models with application to number of falls in the elderly. *Biostatistics and Biometrics Open Access Journal*, 1(4):69–75, 05 2017.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 02 2019. ISSN 0360-0300.

Yaochen Zhu, Jing Ma, and Jundong Li. Causal inference in recommender systems: A survey of strategies for bias mitigation, explanation, and generalization. *arXiv:2301.00910*, 2023.