Deep Causal Reasoning for Recommendations

YAOCHEN ZHU, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
JING YI, School of Computer Science, Wuhan University, Wuhan, China
JIAYI XIE, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
ZHENZHONG CHEN∗, School of Remote Sensing and Information Engineering, School of Computer Science, Wuhan University, Wuhan, China

Traditional recommender systems aim to estimate a user’s rating to an item based on observed ratings from the population. As with all observational studies, hidden confounders, which are factors that affect both item exposures and user ratings, lead to a systematic bias in the estimation. Consequently, causal inference has been introduced in recommendations to address the influence of unobserved confounders. Observing that confounders in recommendations are usually shared among items and are therefore multi-cause confounders, we model the recommendation as a multi-cause multi-outcome (MCMO) inference problem. Specifically, to remedy the confounding bias, we estimate user-specific latent variables that render the item exposures independent Bernoulli trials. The generative distribution is parameterized by a DNN with factorized logistic likelihood and the intractable posteriors are estimated by variational inference. Controlling these factors as substitute confounders, under mild assumptions, can eliminate the bias incurred by multi-cause confounders. Furthermore, we show that MCMO modeling may lead to high variance due to scarce observations associated with the high-dimensional treatment space. Therefore, we theoretically demonstrate that controlling user features as pre-treatment variables can substantially improve sample efficiency and alleviate overfitting. Empirical studies on both simulated and real-world datasets demonstrate that the proposed deep causal recommender shows more robustness to unobserved confounders than state-of-the-art causal recommenders. Codes and datasets are released at https://github.com/yaochenzhu/Deep-Deconf.

CCS Concepts: • Information systems → Collaborative filtering; • Theory of computation → Bayesian analysis.

Additional Key Words and Phrases: Recommender systems; causal inference; unobserved confounders; generative models; variational inference

1 INTRODUCTION

Estimating users’ preference based on their past behaviors is of great importance in personalized recommendations. Collaborative filtering, which infers user’s rating to items based on the observed ratings from the user population, has been widely applied in modern recommender systems [14, 58]. However, since a user’s rating to an item is generally not independent of the item’s exposure to the user, the collected rating data are unavoidably biased [4]: Consider movie recommendations as an example. Since the genre of a movie affects both the likelihood of

∗Corresponding Author: Zhenzhong Chen.

This research was supported in part by the National Natural Science Foundation of China (Grant No. 62036005) and Tencent. The numerical calculations in this article have been done on the supercomputing system with the Supercomputing Center, Wuhan University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2157-6904/2024/3-ART
https://doi.org/10.1145/3653985

ACM Trans. Intell. Syst. Technol.
its exposure and user’s rating to it, a spurious dependence is created between user ratings and item exposures, which makes movies in the minority genres systematically under-represented by the collected data (see Fig. (1) as an example). This is a confounding phenomenon, and the movie genre is one of the confounders. Ignoring the influence of such confounders could lead to systematic confounding bias that degrades the recommendation quality for traditional recommender systems [3, 35, 59, 61].

If we compare recommendation (or exposing an item to a user) as a treatment and user’s ratings toward the item as the outcome, since whether or not an item is exposed to a user is pre-determined after the collection of the historical user data, the user’s rating associated with one exposure status of the item must be unobserved to the system. Therefore, similar to estimating treatment effects in clinical trials, eliminating the confounding bias in recommendations demands us to answer a counterfactual problem, i.e., what a user’s rating would be if a previously unexposed item is made exposed (recommended) to the user. This falls under the scope of causal inference, which aims to unbiasedly estimate treatment effects for a unit from observed outcomes from the population. In this article, Rubin causal model (RCM) [38] is adopted as the causal inference framework, where the deconfounded recommendation task can be framed as estimating the unit-level "treatment effects" for a user from observed ratings from all users (i.e., the population) [40]. To address such a counterfactual inference problem, classical causal inference demands us to find, measure, and adjust the influence of all the confounders. However, since any attribute that is shared among items (such as the genre of movies) could serve as a potential confounder, they are infeasible to enumerate. Alas, whether or not we have indeed exhausted all the confounders is not testable [38].

Observing that the item pools of modern recommender systems are usually large, we assume that single-cause confounder, which is a factor that exclusively affects the exposure and rating of one item, is negligible. Therefore, the deconfounding problem can be simplified because multi-cause confounders, i.e., confounders shared among items, can be properly handled by controlling and estimating substitute confounders inferred from the exposure status of all items via latent factor models [46]. This has been explored by Wang et al. [48] by proposing the Deconfounded Recommender (Deconf-MF). However, they mainly focused on using shallow methods such as Poisson matrix factorization to infer the substitute confounders, where a closed-form approximate inference solution can be deduced. However, this may fail to capture the complex item co-exposure relationship caused by confounders. Even if the exposure model is correctly specified, the rating potential outcome prediction model of Deconf-MF degenerates into a single-cause case where the item co-recommendation effect (the interference effect of recommending several items simultaneously) is ignored. Utilizing deep neural networks (DNNs) to model
both the item exposure and user ratings, which is demonstrated to be superior in classical collaborative filtering (i.e., item co-visit prediction) tasks [24], remains under-explored due to the intractable posteriors. In addition, the multiple causes induce an exponentially large causal space, which makes the observed ratings associated with one specific exposure vector in the population scarce. This could lead to a large variance in the estimated causal effects of seldomly observed item exposures. The variance, for a DNN-based causal model, appears as its tendency to overfit when the number of users is limited.

To address the above challenges, we propose a deep deconfounded recommender system, i.e., Deep-Deconf, to address the confounding bias problem in recommendations. In a nutshell, Deep-Deconf frames the recommendation as a multi-cause multi-outcome (MCMO) inference problem, where the item exposures and user ratings are regarded as multi-cause treatments and potential outcomes, respectively. Under this modeling strategy, item co-exposures are used to estimate the substitute confounder to eliminate the confounding bias, and item co-recommendation effects can be properly considered to predict new ratings. Specifically, to eliminate the confounding bias, based on a no single-cause confounder assumption, we infer and control user-specific latent variables as substitute confounders, where the factual generative distribution of the observed user ratings is parameterized by a DNN with factorized logistic likelihood and the intractable posteriors are approximated by variational inference, respectively. Furthermore, we demonstrate that the MCMO modeling strategy may suffer from high variance, mainly due to the scarcity of observation of rating potential outcomes associated with the high-dimensional treatment spaces of multiple items. To reduce the variance, we demonstrate that introducing user features as pre-treatment variables, which are factors that are independent of causes but are informative to predicting rating potential outcomes, such as user ages, genders, locations, etc., can substantially improve the sample efficiency. Finally, we proposed the theory of "duality of multi-cause confounders" to explain a previously ignored phenomenon where the recommendation performance increases first and then decreases with the rise of confounding levels. We demonstrate that this is because multi-cause confounders contain useful collaborative information, but greedily exploiting them could bias the model and degenerate the recommendation quality. Extensive experiments conducted on multiple simulated and real-world datasets show that Deep-Deconf is more robust to unobserved confounders than state-of-the-art causal recommenders. The main contribution is summarized as follows:

- We present Deep-Deconf, a deep neural network-based causality-aware recommendation algorithm built upon Rubin’s causal framework. Through controlling a user-specific latent factor and informative user features as pre-treatment variables, Deep-Deconf leads to confounding-bias-robust recommendation with substantially low variance.
- We theoretically demonstrate that the global and local Jacobians of the Deep-Deconf network can be associated with the treatment effects for all and a subsection of the user population, which leads to the improved explainability of the proposed Deep-Deconf.
- We demonstrate the side effects of large variance associated with the multi-cause causal inference modeling paradigm, where we theoretically and empirically show that controlling user features as pre-treatment variables can substantially lower the variance.
- Experiments conducted on both simulated and real-world datasets demonstrate that Deep-Deconf leads to improved robustness toward confounding bias compared with the baseline methods. Moreover, we propose the theory of "duality of multi-cause confounders" to explain the non-monotone change of recommendation performance for deconfounded recommender systems when the influence of unobserved confounders increases.

ACM Trans. Intell. Syst. Technol.
2 RELATED WORK

2.1 Causality-based Recommendations

Recently, researchers began to realize the importance of fairness [23], diversity [50] and causality. Among them, causality-based recommendation can be seen as an intervention analogous to the treatment in clinical trials, where users’ feedback can be compared to the outcome of the treatment [40]. Since randomized experiments are clearly infeasible for recommender systems, confounders, which are factors that affect both the item exposure and user ratings, pervasively exist and lead to severe bias in the collected data [44]. Therefore, causal inference can be utilized to eliminate the confounding bias and uncover the true causal relationships between item exposure and user ratings [40]. Existing causality-based recommender systems can be classified into three main categories: propensity score reweighting (PSW)-based methods, substitute confounder-based methods, and adjustment-based methods. In addition, methods from each category can also use two paradigms as the fundamental causal framework: Rubin’s potential outcome causal framework (RCM) [16], and Pearl’s structural causal model-based framework (SCM) [34].

2.1.1 Inverse Propensity Weighting (IPW)-based Methods. RCM models the outcome (ratings) associated with different treatments as the (rating) potential outcomes, where only one potential outcome is observable. Therefore, to calculate the causal effects of the treatment, i.e., the difference between the potential outcome under treatment and under no treatment, randomized experiments should be conducted. However, random recommendations are clearly infeasible in modern online systems. Consequently, unobserved confounders pervasively exist and cause discrepancies between the two groups in the collected user data. PSW aims to reweight users in the treatment group by the chances that they receive the treatment (i.e., the propensity scores), such that they can be viewed as random samples from the population [1, 17, 39]. Linear regression [40] and variational auto-encoders (VAEs) [63] have been used to estimate the propensity scores from historical ratings and user features. Recently, Xiao and Wang propose a doubly robust IPW based on deep variational information bottleneck to reduce the bias and variance of the estimator [51]. In addition, Li et al. proposed to use a small number of unobserved ratings to facilitate the estimation of balancing weights [22]. Furthermore, He et al. propose to rescale the nominal propensity estimated by the benchmark model to account for the unobserved confounders [12]. However, one problem of PSW is that the estimated propensity scores can be extremely small when the causal space is high-dimensional, which leads to large sample weights that make the training dynamics unstable. Moreover, the unbiasedness of IPW relies heavily on a correctly specified propensity score estimation model, which is generally non-testable from experiments.

2.1.2 Substitute Confounder-based Methods. Another class of RCM-based causal recommender systems aims to find, measure, and control all the unobserved confounders [62]. However, since exhausting confounders is clearly infeasible, recent work focuses on inferring and controlling substitute confounders as the surrogates [5]. One exemplar work is the Deconf-MF [48], where the item co-recommendations for a user are viewed as a bundled treatment, where user-specific latent factors are estimated to render the exposure of different items conditionally independent. Under mild assumptions, confounding bias can be eliminated by controlling such latent factors as substitutes. However, both the substitute confounder inference model and recommendation model in Deconf-MF are shallow matrix factorization-based models, which may have insufficient modeling ability for large-scale modern recommendation tasks. Utilizing deep neural networks (DNNs) to model both the item exposures and the rating potential outcomes associated with high dimensional treatment space, however, remains under-explored due to intractable posteriors and sparsity of observations. Faced with these challenges, we generalize the Deconf-MF to a DNN-based framework where the non-linear influence of confounders to the ratings can be captured and remedied. In addition, a pre-treatment variable controlling-based strategy is proposed to improve the data efficiency. Concurrent with our work, [29] proposed the DIRECT algorithm to
extend the Deconf-MF, which focuses on learning disentangled representations of the treatment with VAE for better explanation. Compared to DIRECT, Deep-Deconf mainly tackles network weights interpretability, variance reduction of MCMO modeling, and duality of multi-cause confounders, which are three new independent research questions worthy of in-depth investigations.

2.1.3 Adjustment-based Methods. Pearl’s SCM-based framework constructs a priori a direct acyclic graph (DAG) that depicts the causal relationships among variables of interest, where adjustment-based methods can be applied to eliminate the influence of known confounders. Compared to RCM, causal graphs can clearly illustrate the assumed causal relationships among user/item features, interaction histories, item exposures, and user ratings with nodes and links, which is more intuitive than the RCM-based methods. One exemplary graph adjustment-based strategy is the backdoor adjustment [34], in which the influence of pre-treatment factors on the observed treatment assignment (i.e., item exposure) can be eliminated such that items can be viewed as exposed in a random manner. Generally, the established SCM varies drastically among different methods, both in variables included for consideration and the assumed links among them [26, 49, 54, 55]. Moreover, there is also no consensus in approximate inference strategies to solve the adjustment formula [45, 56, 57, 60]. For example, [57] models the item popularity as the confounder and eliminates its influence to reduce the popularity bias, whereas [53] explicitly models the influence of users’ historical ratings on items’ exposure, where the influence is then removed via do-calculus.

2.2 Deep Learning for Causal Inference

Recent years have witnessed an upsurge of interest in combining DNNs with causal inference [28, 33, 41, 52]. Among them, the most relevant methods to Deep-Deconf are CausalVAE [27] and [36], which explored the deep latent-variable model for single cause inference tasks, e.g., twin weights and job training, based on RCM. These works mainly focused on modeling the joint distribution of treatment, hidden confounders, and potential outcomes via deep generation networks and approximating the intractable posteriors of hidden confounders through variational encoders. However, this strategy is not directly applicable to recommender systems because when multiple treatments exist, a sub-encoder and a sub-decoder may be required for each configuration of the treatments. However, since the number of treatments for recommendations is exponential to the number of items, extrapolating all the missing potential outcomes requires intractable numbers of inference and generation networks. The proposed Deep-Deconf circumvents this issue by utilizing a two-stage modeling strategy, where the exposure model first fits the joint distribution of substitute confounders and item exposures, and the outcome model then fits the distribution of rating potential outcomes conditional on the substitute confounders, item exposures, and user features. This leads to a parameter-efficient solution to the DNN-based multiple cause inference problems.

3 PROBLEM FORMULATION

Suppose a system with $U$ users and $I$ items. The observational data comprises the rating matrix $R \in \mathbb{R}^{U \times I}$ and the user features $X \in \mathbb{R}^{U \times S}$ where each row $r_u^T$, $x_u^T$ are the rating and feature vector for user $u$. Based on the Rubin causal model [16], the received treatment is represented by the exposure matrix $A$ where $a_{ui}$ denotes whether the item $i$ has been exposed to the user $u$ when rating $r_{ui}$ is provided. We denote the potential outcome random variable associated with an exposure $A_u$ for the user $u$ as $R_u(A_u)$. For each user, we only observe the value of $R_u(A_u = a_u)$, i.e., $r_u(a_u)$. The main quantity of interest is the ratings user $u$ would provide if $K$ extra items are recommended. Therefore, recommendation under causal reasoning is a counterfactual inference problem. To

---

1We use boldface capital symbols $R$ for matrices, boldface lower case symbols $r_u$ for vectors, non-boldface lowercase symbols $r_{ui}$ for scalars, non-boldface capital symbols $R_u$ for random variables ($R_u$ may be scalar, vector, or matrix based on context). Exceptions such as $U$ and $I$ can be easily identified from the context.
address this problem, we assume that stable unit treatment value assumption (SUTVA) holds where the ratings of user $u$ is independent of items’ exposure to user $v$, i.e., $R_u \perp A_v | A_u$. This excludes from consideration the interference among users (interested readers can refer to [30] for a discussion of spillover effects). The randomness of $R_u$ and $A_u$ sources from the fact that the user $u$ is an arbitrary user sampled from a large user population. Therefore, unless specified otherwise, the expectations are taken w.r.t. this population for the rest of the article. The purpose of Deep-Deconf is to estimate the expected causal effects of an exposure $A_u$ to a user $u$, $\mathbb{E}[R_{ui}(A_{ui} = a_{ui})]$, from the population observations so that unbiased recommendations can be made accordingly.

4 THE DEEP DECONFOUNDED RECOMMENDER

4.1 Debias via Factorized Variational Auto-encoder

4.1.1 Problem Analysis. Traditional recommender systems directly model the conditional distribution $p(R_{ui} | A_{ui})$ via matrix factorization or DNNs with observed ratings to calculate $\mathbb{E}[R_{ui}(A_{ui} = a_{ui})]$, relying on the assumption that $\mathbb{E}[R_{ui}(A_{ui} = a_{ui})] \approx \mathbb{E}[R_{ui}(A_{ui}) | A_u \approx a_u]$, i.e., users with similar historical interactions tend to have similar ratings. The approximation is unbiased only if there exist no variables that simultaneously affect $R_{ui}$ and $A_u$, i.e., no unobserved confounders [46]. However, unobserved confounders pervasively exists in collected ratings, which leads to systematic bias (e.g., amplified exposure bias due to similar bias in similar users) for naive methods [42].

4.1.2 Substitute Confounders. To eliminate confounding bias, classical causal inference techniques demand us to find, measure and control all confounders $C_{ui}$ and calculate $\mathbb{E}[R_{ui}(A_{ui})]$ from $\mathbb{E}_{C_{ui}}[\mathbb{E}[R_{ui}(A_{ui}) | C_{ui}, A_u]]$. However, it can guarantee unbiasedness provided that there are no uncontrolled confounders. This is known as the strong ignorability assumption, which is both infeasible and untestable from experiments [13]. Therefore, to circumvent exhausting and measuring all confounders $C_{ui}$, Deep-Deconf models the recommendation as a multi-cause inference problem. Instead of treating the exposure of an item $A_{ui}$ as an isolated cause, we consider the exposures of all the items to a user $A_u$ as a holistic treatment that could causally affect all the ratings. We first assume that the single-cause confounder, i.e., $S_u$, which influences only the exposure of one specific item and its ratings, does not exist. The validity of the assumption can be justified by the fact that the pool of candidate items for modern recommender systems is usually large, and therefore it is unlikely that a confounder influences only one of the items. Taking genre as an example, the genre of a movie affects both its exposure and rating, and it is a universal attribute that is shared among all movies. With the assumption of no-single cause confounders, we only need to control the multi-cause confounders. This is a more amenable objective, since controlling them is equivalent to controlling latent factors $Z_u \in \mathbb{R}^K$ that render the causes conditionally independent, i.e.,
p(Au | Zu) = \Pi_i p(Aui | Zu). A simple proof for the validity of the claim is that, if multi-cause confounders still exist after conditioning on such latent factors, the exposures cannot be conditionally independent, which renders a contradiction (see Fig. 2 for an illustration).

4.1.3 Inference Strategy. To find such a latent variable Zu, we first parameterize the generative distribution \( p(\theta | Zu) \) by a DNN with factorized logistic likelihood, i.e., \( p(aui | Zu) = \Pi_i \text{Bern}(aui | [\theta(Zu)]) \). Conditional on Zu = zu, the exposures au for user u can be viewed as generated from randomized Bernoulli trials. The factorized Bernoulli distribution satisfies the overlap assumption, i.e., \( p(\theta | Zu) > 0 \) provided that the exposure probability of the item i, i.e., \([\theta(Zu)]\), not equals to \{0, 1\}, which is crucial for the identifiability of the model [15].

We then calculate the intractable posterior q(\theta | Zu, Au) via the variational inference [20], where the prior for \( \theta \) is set to be the standard Normal \( N(0, I_K) \). It could be noted that the assignment model of Deep-Deconf resembles the Multi-VAE for recommendation with implicit feedback [24], because they both manage to reconstruct the binary exposure/click inputs. However, a characteristic that makes the assignment model fundamentally different is that it requires Au to factorize conditional on Zu. This renders the multinomial likelihood (which has been demonstrated to be more suitable for recommendations in [24]) invalid in our case. The assumption of no single-cause confounder and the utilization of user latent factors as surrogate multi-cause confounders lead us to the following equality,

\[
E[Ru(Au)] = E_{Zu}[E[Ru(Au) | Zu, Au]].
\]

(1)

If the exposure model of Deep-Deconf is correctly specified and accurately estimated, conditional on Zu = zu, the observed ratings for the users can be analyzed as were generated from a randomized experiment, and bias due to unobserved multi-cause confounders is eliminated.

4.2 Rating Prediction via Deep Outcome Network

In this section, we introduce the deep rating potential outcome prediction network that predicts the rating potential outcomes Ru(Au). Since the outcome network no longer requires the ratings to be conditionally independent, more powerful multinomial likelihood can be put on ratings for more accurate predictions (see Fig. 3 for the structure of the deep outcome model). To gain more insights into the network, we first provide a theoretical analysis of the network weights for a special case where the network has a single linear layer with no activation, i.e.,

\[
r_u(a_u) = W^\tau Zu + Wa_u + \alpha_u + \epsilon_u.
\]

(2)
Fig. 4. The co-recommendation effects and large variance associated with MCMO modeling. In parallel universe 1, user $i$ had been recommended with item #3, which increased her expectation to item #2 and made her less satisfied with it. The large variance of MCMO is because the number of such counterfactuals is exponential to the number of items.

where $e_u$ is the residual vector for user $u$ and $\alpha$ is the constant term. For users with inferred substitute confounder equals $z_u$, we can calculate the expected causal effect for the exposure of item $j$ on the rating of item $i$ as

$$E[\Delta R_{ui} | z_u] = E[R_{ui}(a + e_j) - R_{ui}(a) | z_u] = [W^*e_j]_i = w_{ij}^*,$$

where $e_j$ is a one-hot vector with one at the $j$-th position. Eq. (3) shows that the network weights $w_{ij}^*$ can be interpreted as the conditional average treatment effect (CATE) of recommending item $i$ on the rating of item $j$. Moreover, since the R.H.S. of Eq. (3), $w_{ij}^*$ does not contain a $z_u$-related term, we have $E_{Z_u}[E[\Delta R_{ui} | Z_u]] = w_{ij}^*$, which leads to a further conclusion that $w_{ij}^*$ is also the average treatment effect (ATE) in the population. Note that allowing the causal effects of one item’s exposure on the rating of another item does not violate the SUTVA assumption, as the non-interference of users’ exposures to each other (i.e., $z_u \perp A_{ui} | Z_u$, $\forall u \neq v < U$) does not exclude from consideration the interference of different items’ exposures of a single user (i.e., $R_{ui} \perp A_{uj} | A_{ui}$, $\forall i \neq j < I$). On the contrary, it is beneficial to model such co-recommendation effects, since exposing one movie to a user may alter her expectation to movies with a similar genre and therefore causally influences her ratings towards these movies as well. An intuitive illustration can be referred to Fig. (4).

4.3 Variance Reduction via Pre-treatment Variables

4.3.1 Variance Analysis. Although confounding bias can be remedied by the controlling substitute confounders, the predicted ratings may suffer from a large variance due to the following two factors: First, multiple causes lead to a more severe data missing problem. Consider a system with $I$ items. The number of counterfactuals is $2^I$, but only one of the outcomes is observed. Therefore, the sample efficiency is exponentially reduced compared with the classical single cause problems. Furthermore, even if the observations are sufficient, the item exposures tend to depend heavily on the user preference. This makes the exposure and non-exposure group for different treatments highly imbalanced, which further increases the estimand variance [16].

4.3.2 Variance Comparison. In this section, we discuss the variance reduction technique of Deep-Deconf by introducing user features in the outcome model as pre-treatment variables, which are factors that remain uninfluenced by item exposures but are predictive to the ratings. To see how this works, we derive the estimand variance before and after introducing user features as pre-treatment variables. For simplification, the network weights $W^u$ are for now reduced to diagonal, i.e., $w^u = \text{diag}(W^u)$, and $z_u, x_u$ are one-dimensional. In this scenario, the co-recommendation effects vanish and we can treat the recommendation of each item $R_{ui}$ separately (Note that the randomness of $R_{ui}$ is still solely due to the sampling of $u$ from the population since the item $i$ is fixed).
To intuitively show the variance reduction, we put aside for now the powerful multinomial likelihood on ratings, and focus on the Gaussian likelihood and the simple Ordinary Least Square (OLS) optimizer. The outcome model before the introduction of user features can be specified as

\[ r_{ui}(a_{ui}) = w^T z_u + w^T_0 a_{ui} + \alpha_i + \epsilon_{ui}. \]

(4)

Conditional on \( Z_u = z_u \), with the overlap assumption, when the sample size \( U \) is large enough, we can meaningfully calculate the following statistics from the samples,

\[ \bar{r}_{ui}(t) = \frac{1}{U} \sum_{u: a_{ui} = t, Z_u = z_u} r_{ui} / U^1, \]

(5)

where \( U^1 = \sum_{u: Z_u = z_u} I(a_{ui} = t), \ t \in \{0, 1\} \) is the size of exposure and non-exposure group in the sub-population. We define CATE on the rating \( r_{ui} \) as \( \tau_{ui} = \mathbb{E}[r_{ui}(1) - r_{ui}(0) | z_u] \). Since unconfoundedness holds when conditional on \( Z_u \), the average rating difference between exposure and non-exposure use group \( \bar{r}_{ui} = r_{ui}(1) - r_{ui}(0) \) is asymptotically unbiased for \( \tau_{ui} \). Furthermore, if we define \( w^{ols}_i \) as the coefficient \( w^*_i \) learned from the \( U \) users drawn from the population, similar deductions as Eq. (2) can demonstrate that \( w^{ols}_i \approx \bar{w}_{ui} \). The variance of \( w^{ols}_i \), under the assumption of homoskedasticity, converges in probability to \( \frac{\sigma^2_{\tau_{ui}|Z_u,A_{ui}}}{U(p_i(1-p_i))} \), when \( U \) approaches infinity, where \( p_i \) is the limit of \( U^1 / U \) and \( \sigma^2_{\tau_{ui}|Z_u,A_{ui}} \) is the population conditional variance.

After that, we consider introducing user features \( x_u \) as additional pre-treatment variables. The new model becomes

\[ r_{ui}(a_{ui}) = w^T z_u + w^T_0 a_{ui} + w^T x_u + \alpha_i + \epsilon_{ui}. \]

(6)

After introducing an additional covariate \( x_u \) in the model, the same algebra shows that \( w^{ols}_i \) estimated by OLS is still asymptotically unbiased for \( \bar{w}_{ui} \). But the limiting variance of the estimand becomes \( \frac{\sigma^2_{\tau_{ui}|x_u,Z_u,A_{ui}}}{U(p_i(1-p_i))} \) (Proofs are in Appendix A.2). Therefore, as long as the user features are indicative to the variation of the ratings, \( \sigma^2_{\tau_{ui}|x_u,Z_u,A_{ui}} \) can reduce considerably compared to the marginal variance \( \sigma^2_{\tau_{ui}|Z_u,A_{ui}} \), which increases the precision of the estimand.

### 4.4 Theoretical Analysis in the Non-linear Case

The previous derivations of CATE interpretation of network weights and variance reduction are mainly based on a simple linear network. However, as the main contribution of this article is proposing a deep deconfounded model for recommendations, we generalize the analysis to the non-linear case. We begin by defining the DNN-based rating potential outcome prediction model as

\[ r_u(a_u) = f_{nn}(a_u, z_u, x_u), \]

(7)

where the function \( f_{nn} : \mathbb{R}^{l+K+F} \rightarrow \mathbb{R}^l \) is non-linear but differentiable almost anywhere. The generalization is achieved from two aspects. First, we consider the global property of \( f_{nn} \). Since the exposure \( a_u \) is a binary vector, the prior of substitute confounder \( z_u \) is set to \( \mathcal{N}(0, I_l) \), and the user features \( x_u \) are rescaled to zero mean and unit variance, we can form a global approximation of \( f_{nn} \) with its Taylor expansion at the original point \((0, 0, 0)\) and keep only the linear term,

\[ f^0_{nn}(a_u) = f^0_{nn}(a_u, z_u, x_u) = W^{a^T}_0 a_u + W^{z^T}_0 z_u + W^{x^T}_0 x_u + \alpha_0. \]

(8)

where the coefficient matrices \( W^{a^T}_0, W^{z^T}_0, W^{x^T}_0 \) are the Jacobians at the original point, and \( \alpha_0 \) is the expected user ratings where no items are recommended. The reason to justify the approximation of \( f^0_{nn} \) to \( f_{nn} \) is that generally, \( f_{nn} \) cannot be highly-nonlinear (\( f_{nn} \) is generally composed of 0-2 hidden layers); otherwise, the outcome model will overfit on the negative samples which are not truly negative and fail to generalize to the recommendation of new
where $W$ can be interpreted as both the CATE and the ATE of the recommendation of item $i$ on the rating of item $j$.

A more meaningful and accurate generalization strategy is to show the local property of the outcome prediction model $f_{nn}$. For user $\hat{u}$ with exposures $a_{\hat{u}}$, substitute confounders $z_{\hat{u}}$, and user features $x_{\hat{u}}$, we can linearize $f_{nn}$ at the point $(a_{\hat{u}}, z_{\hat{u}}, x_{\hat{u}})$ by Taylor expansion,

$$
\hat{r}_{\hat{u}}(a_{\hat{u}}) = f_{nn}(a_{\hat{u}}, z_{\hat{u}}, x_{\hat{u}}) = W_{a}\hat{u}(a_{\hat{u}} - a_{\hat{u}}) + W_{z}\hat{u}(z_{\hat{u}} - z_{\hat{u}}) + W_{x}\hat{u}(x_{\hat{u}} - x_{\hat{u}}) + \alpha_{\hat{u}},
$$

where $W_{a,z,x}$ are the network Jacobians at $(a_{\hat{u}}, z_{\hat{u}}, x_{\hat{u}})$, and $\alpha_{\hat{u}}$ is the expected ratings for user $\hat{u}$ where no items are recommended. With the local linearization of $f_{nn}$, similar analysis can be applied to $f_{nn}^{0}$. Note that the trade-off is that $w_{a,j}^{0}$ is no longer the ATE of recommending item $i$ on the rating of item $j$ for the entire population, but is only the CATE (which is also considered as the individual treatment effects in many existing literatures) for users who are similar with $\hat{u}$ (which is measured by $z_{\hat{u}}$). Therefore, this strategy establishes a corresponding relationship between local Jacobians of $f_{nn}$ and CATE for a sub-population (see Fig. 5 for details).

4.5 Potential Rating Prediction for Recommendations

Deep-Deconf is designed to unbiasedly predict rating potential outcomes associated with any exposure vector $a_{\hat{u}} \in \{0, 1\}^T$, i.e., $r_{\hat{u}}(a_{\hat{u}})$ with a low variance. However, which $a_{\hat{u}}$ should be selected for prediction is undetermined, as it demands the answer to the exact question we are trying to solve: which items should be exposed to the users. Wang et al. [48] proposed to use $a_{\hat{u}}^{obs+K}$, i.e., the exposure of the originally observed items and $K$ newly recommended items, as the exposure to calculate the ratings for recommendations. However, for Deep-Deconf, enumerating every $r_{\hat{u}}(a_{\hat{u}}^{obs+K})$ and taking the expectation is clearly infeasible. Therefore, we propose an approximate strategy for the prediction. We assume that the number of recommended items $K$ (e.g., Top 20) is small compared with the size of the item pool (e.g., 10,000). Based on this assumption, the observed exposure $a_{\hat{u}}^{obs}$ can be used as a surrogate for $a_{\hat{u}}^{obs+K}$ to calculate $r_{\hat{u}}(a_{\hat{u}}^{obs+K})$ for predictions.

5 EMPIRICAL STUDY

5.1 Datasets

5.1.1 Semi-simulated Datasets. Evaluating causal recommenders on real-world datasets faces great challenges, since usually we do not observe ratings for all the items of the users, and confounders make the model evaluations on a randomly split test set biased [11]. Therefore, most existing deep causal recommenders are evaluated by...
establishing simulated datasets from real-world datasets [63]. We create two semi real-world datasets, ML-causal and VG-causal, based on the real-world MovieLens-1m (ML-1m) and the Amazon-Videogames (Amazon-VG) datasets. In the simulation, we train two VAEs with factorized logistic likelihood and multinomial likelihood on the item exposures (binarized ratings) and user ratings to get the corresponding generative distributions from the user latent variables $z_{u,r}$. The decoders of the exposure and rating VAEs are denoted as $f_{\text{exp}}$, $f_{\text{rat}}$, respectively. We then simulate a $K$-dimensional confounder $c_u$ for each user from $N(0, I_K)$ and the user preference vector $\theta_u$ conditional on the confounders is specified as $\theta_u \sim Y_\theta \cdot c_u + (1 - Y_\theta) \cdot N(0, I_K)$; the constant $Y_\theta \in [0, 1]$ controls the strength of confounding. The user features are the noisy observation of her dimensional-reduced user preference vector $f_{\text{PCA}}(\theta_u + \epsilon_u)$, where $\epsilon_u \sim N(0, \lambda_u^{-1}I_K)$. The exposure vector $a_u$ for user $u$ is generated from

$$a_u = \arg\min_{\alpha_u \in (0,1)} \left( \sum_i |\alpha_{ui} - f_{\text{exp}}(c_u)_i| \right),$$

subject to $p_{\text{causal}}(A_{ui} = 1) = p_{\text{ori}}(A_{ui} = 1)$.

The constraint is to ensure that the global item exposure rate of the causal datasets (causal) is the same as that of the original datasets (ori). Moreover, we define the set $R = \{r_u \in \mathbb{R}^I \mid r_{ui} \in \text{range}(1, 5)\}$ as the set of possible user rating vectors for $I$ items. The rating of user $u$ is generated from

$$r_u = \arg\min_{r_u \in R} \left( \sum_i |y_{ui} - f_{\text{rat}}(\theta_u + Y_\theta \cdot c_u)_i| \right),$$

subject to $p_{\text{causal}}(R_{ui} = r) = p_{\text{ori}}(R_{ui} = r)$, $\forall r \in \text{range}(1, 5)$,

where $Y_\theta$ controls the strength of the basic confounding level (since zero confounding is non-existent). The constraint ensures the global rating distribution in the causal datasets is the same as the original datasets. The observed ratings $r_{u}^{\text{obs}}$ is calculated by masking the original ratings with exposures, $r_{u}^{\text{obs}} = r_u \cdot a_u$. The schematic illustration of the establishment can be referred to in Appendix B.3. The global item popularity distributions before and after the exposure under different levels of confounding effect are shown in Fig. (6). The statistics of the established ML-causal and VG-causal datasets are summarized in Table (5) for reference.

5.1.2 Real-world Datasets.} To demonstrate the practical applicability of the proposed Deep-Decof, we have included the experimental results on the Yahoo! R3 dataset [31] and the recently released KuaiRand

---

2https://grouplens.org/datasets/movielens/1m/

3https://jmcauley.ucsd.edu/data/amazon/
Table 1. Attributes of the established ML-Causal and VG-causal datasets. In the table, % density refers to the density of the rating matrix, #avg/std exposure refers to the corresponding statistics of the number of items that are exposed to the users.

| dataset      | #users | #items | %density | #avg ± std exposure |
|--------------|--------|--------|----------|---------------------|
| ML-causal    | 6,000  | 3,706  | 4.468%   | 165 ± 192           |
| VG-causal    | 7,253  | 4,338  | 0.406%   | 17 ± 16             |

dataset [9]. Specifically, the KuaiRand dataset is comparatively large-scale, which includes 1,413,574 biased user-item interaction data collected from daily user behaviors and 954,814 unbiased interactions collected based on randomized experiments on 23,533 users and 6,712 items. The statistics of the real-world datasets are summarized in Table 2 for reference.

Table 2. Attributes of the real-world Yahoo! R3 dataset and the KuaiRand dataset. #biased int. denotes the number of biased user-item interaction data collected from daily user behaviors, whereas #unbiased int. denotes the unbiased interactions collected based on randomized experiments.

| dataset      | #users | #items | %density | #biased int. | #unbiased int. |
|--------------|--------|--------|----------|--------------|----------------|
| Yahoo! R3    | 5,400  | 1,000  | 2.392%   | 129,179      | 54,000         |
| KuaiRand     | 23,533 | 6,712  | 0.895%   | 1,413,574    | 954,814        |

5.2 Implementation Details and Training Strategy
When establishing the causal datasets, the dimension $K$ of the user preference variables and the confounders is set to 100, 100, 100, respectively. The dimension of the user features is set to 10. The structure of the exposure and outcome models of Deep-Deconf is set to $\{I + \{F + K\} \rightarrow K \rightarrow I\}$, where $I$ is the number of items, $F$ is the dimension of user features, and $K$ is the latent dimension. The models are trained with Adam optimizer [19], with a learning rate of $1e^{-3}$ for 100 epochs. For the exposure model, 20% of the observed exposures of the validation users are held out for predictive check [37]. The selection of the outcome model generally follows the same procedure, where R@20 and N@20 on hold-out ratings are monitored as the metric. For all three datasets, we search $K \in \{50, 100, 150, 200\}$ and find that $K_e$ equals $K$ indeed achieves the best performance among all models with structures in the searching space.

5.3 Evaluation Strategy
We evaluate the model performance under the setting of strong generalization [25], where the observed item exposures and ratings for validation and test users are used only for inference purposes. When training the exposure model for substitute confounder estimation, we put aside 20% of the observed exposures for predictive checks [37], where the best model is selected by log-likelihood of the hold-out exposures. The outcome model is selected by how well it ranks the hold-out observed interactions for the validation users. The ranking quality is evaluated by R@K and N@K, where R@K is the Top-K recall. If we denote the item at ranking position $r$ by $i(r)$ and the set of hold-out items for the user by $I_u$, R@K is calculated as follows:

$$R@K(i) = \frac{\sum_{r=1}^{M} \mathbb{1}[i(r) \in I_u]}{\min(M, |I_u|)}.$$
Table 3. Model comparisons under different confounding levels. The best method is highlighted in boldface. For each method, the best-evaluated performance w.r.t. the confounding level is highlighted with blue and red for R@20 and N@20, respectively. Moreover, we have conducted the one-side t-test for Deep-Deconf and the method with the second-best performance and report the p-value.

| Methods       | $\lambda_p = 0.1$ | $\lambda_p = 0.3$ | $\lambda_p = 0.5$ | $\lambda_p = 0.7$ | $\lambda_p = 0.9$ |
|---------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|               | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 |
| WMF           | 0.318 | 0.306 | 0.325 | 0.312 | 0.321 | 0.310 | 0.322 | 0.314 | 0.313 | 0.301 |
| IPW-MF        | 0.325 | 0.319 | 0.330 | 0.324 | 0.328 | 0.319 | 0.327 | 0.321 | 0.320 | 0.314 |
| Deconf-MF     | 0.333 | 0.331 | 0.341 | 0.336 | 0.340 | 0.341 | 0.344 | 0.346 | 0.336 | 0.339 |
| BRD-MF        | 0.336 | 0.330 | 0.352 | 0.347 | 0.355 | 0.349 | 0.341 | 0.339 | 0.335 | 0.332 |
| Concat-VAE    | 0.369 | 0.354 | 0.385 | 0.376 | 0.405 | 0.396 | 0.385 | 0.378 | 0.388 | 0.379 |
| DRL-VAE       | 0.381 | 0.368 | 0.393 | 0.385 | 0.426 | 0.413 | 0.404 | 0.395 | 0.401 | 0.395 |
| VSR-VAE       | 0.377 | 0.372 | 0.398 | 0.390 | 0.422 | 0.414 | 0.401 | 0.393 | 0.400 | 0.392 |
| Deep-Deconf   | 0.386 | 0.379 | 0.407 | 0.399 | 0.431 | 0.420 | 0.410 | 0.404 | 0.411 | 0.401 |

| Methods       | $\lambda_p = 0.1$ | $\lambda_p = 0.3$ | $\lambda_p = 0.5$ | $\lambda_p = 0.7$ | $\lambda_p = 0.9$ |
|---------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|               | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 |
| WMF           | 0.068 | 0.062 | 0.072 | 0.071 | 0.074 | 0.070 | 0.073 | 0.065 | 0.065 | 0.059 |
| IPW-MF        | 0.067 | 0.062 | 0.075 | 0.073 | 0.066 | 0.072 | 0.071 | 0.064 | 0.066 | 0.065 |
| Deconf-MF     | 0.076 | 0.069 | 0.084 | 0.077 | 0.082 | 0.073 | 0.080 | 0.079 | 0.078 | 0.073 |
| BRD-MF        | 0.075 | 0.076 | 0.099 | 0.082 | 0.091 | 0.077 | 0.085 | 0.080 | 0.074 | 0.079 |
| Concat-VAE    | 0.094 | 0.097 | 0.103 | 0.106 | 0.105 | 0.101 | 0.113 | 0.108 | 0.101 | 0.103 |
| DRL-VAE       | 0.099 | 0.100 | 0.104 | 0.111 | 0.113 | 0.109 | 0.114 | 0.117 | 0.110 | 0.106 |
| VSR-VAE       | 0.103 | 0.102 | 0.109 | 0.112 | 0.110 | 0.107 | 0.117 | 0.114 | 0.106 | 0.106 |
| Deep-Deconf   | 0.108 | 0.105 | 0.113 | 0.111 | 0.117 | 0.109 | 0.124 | 0.121 | 0.115 | 0.109 |

| Methods       | $\lambda_p = 0.1$ | $\lambda_p = 0.3$ | $\lambda_p = 0.5$ | $\lambda_p = 0.7$ | $\lambda_p = 0.9$ |
|---------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|               | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 | R@20 | N@20 |
| WMF           | 2.40E-4 | 2.23E-2 | 1.42E-3 | > 0.1 | 1.93E-3 | > 0.1 | 3.67E-4 | > 0.1 | 1.33E-3 | > 0.1 |
| IPW-MF        | 2.40E-4 | 2.23E-2 | 1.42E-3 | > 0.1 | 1.93E-3 | > 0.1 | 3.67E-4 | > 0.1 | 1.33E-3 | > 0.1 |

where $\mathbb{1}$ in the numerator is the indicator function, and the denominator is the minimum of $K$ and the number of hold-out items. N@K is the normalized DCG defined as follows:

$$DCG@K(u) = \frac{\sum_{r=1}^{K} \mathbb{1}[r \in E_u]}{\log(r+1)} - 1$$

The model is selected by N@20 on validation users where 20% of the observed ratings are hold-out for model evaluation. The R@20 and N@20 on test users with fully observed ratings for all items averaged over five different splits of the datasets are reported as the unbiased model performance.

5.4 Baselines

Our primary baseline is the deconfounded recommender (Deconf-MF) [48], which also models the recommendation as a multiple causal inference problem. In Deconf-MF, the substitute confounders and ratings are estimated by linear Poisson factorization. Another main baseline is DRL-VAE, where we adopt the key idea of [10], i.e., inferring latent confounders from user features via variational auto-encoder (VAE) [27], and we adapt it to the Multi-VAE backbone by controlling the inferred confounder as the same way as Deep-Deconf. Three other causality-based recommenders are based on propensity-score re-weighting, which eliminates the confounding bias by re-weighting the ratings by the inverse of propensity scores 

i.e., the chance of their exposures conditional on covariates (previous exposures and user features). The first method is the inverse propensity weighting matrix factorization (IPW-MF) [40], where the propensity scores are estimated by simple regression. Another method is BRD-MF [8], which rescales the nominal propensity score estimated by the IPS benchmark within a bound derived from the estimated strength of unobserved confounders. The third method is the variational sample re-weighting (VSR) [63], which considers the multiple causes as a bundled treatment and

---

4https://github.com/yixinwang/causal-recsys-public
estimates the propensity scores via latent variables instead of observations. For a fair comparison, the matrix factorization-based methods are augmented with user features in an SVD++ manner [21], and the VAE-based methods concatenate user features with exposures as the extra inputs, where improvement has been observed compared to their original forms. The non-causal baselines we include are weighted matrix factorization [14] (WMF, also augmented with user features) and concat-VAE, which is a variant of Deep-Deconf where the exposure model is removed to demonstrate the effectiveness of causal reasoning for recommendations. Bayesian parameter search [6] is used to find the optimal hyperparameters for MF-based recommenders and grid search is used for deep recommenders.

5.5 Experiments on the Semi-Simulated Datasets
In this section, we provide the experimental results on two semi-simulated datasets, i.e., the VG-causal and ML-causal datasets. Since in the simulation, the strength of confounding bias and noise level of user features are controllable, their influences on the model are thoroughly discussed.

5.5.1 Experimental Setups. In the simulation, we fix the basic confounding level $\gamma_b$ to 2.0 as with the empirical estimation of Wang et al. [48], and we vary the strength of user preference confounding effects by changing $\gamma_p \in \text{range}(0.1, 0.2, 0.9)$. The models are then evaluated under different confounding levels. In addition, we also conducted experiments with varied noise levels in user features to provide sensitivity analysis regarding variance reduction via controlling pre-treatment variables in outcome prediction models that are predictive of the ratings.

5.5.2 Comparisons with the State-of-the-Art. The comparison results are demonstrated in Table (3). From Table (3), we can find that the vanilla WMF performs the worst among all the methods that we draw comparisons with, especially when the ratings are heavily confounded. IPW-MF addresses the confounding bias by re-weighting the ratings by the propensity scores estimated through regression. While improvement has been observed over WMF in most cases, the unbiasedness of IPW-MF requires two strong assumptions, i.e., unconfoundedness and a correctly specified propensity model, which relies heavily on the expertise of the researchers. Therefore, it hinders IPW-MF’s further improvement. Deconf-MF weakens the unconfoundedness assumption of IPW-MF to the non-existence of single-cause confounders, and is the best matrix factorization-based baselines demonstrated in the middle part of three sub-tables of Table (3). However, since Deconf-MF models both the exposures and the ratings as linear Poisson matrix factorization, it fails to capture non-linear influences of unobserved confounders. Moreover, it treats a user’s ratings to different items separately in the outcome model, where co-recommendation effects cannot be considered to further improve recommendation performance. Since BRD-MF treats the interactions of a user separately, it is still based on the shallow matrix-factorization (MF) backbone. Therefore, it can still be outperformed by models with deep backbones. While DRL-VAE is based on the same Multi-VAE backbone as the proposed Deep-Deconf, the item co-exposures are not well-utilized to support the inference of latent confounders.

Consequently, Deconf-MF is outperformed by the Multi-VAE-based deep generative models, even if Concat-VAE is not causality-based and VSR is based on propensity score-reweighting, which requires a stronger unconfoundedness assumption for unbiasedness to hold. Combining the advantages of the non-linear collaborative modeling ability of Multi-VAE and the model-agnostic debiasing advantage of the substitute confounder-based causal inference for recommendations, Deep-Deconf achieves a systematic performance improvement compared with Deconf-MF while being more robust than CondVAE and VSR faced with unobserved confounders. Therefore, the experiments demonstrate the superiority of the Deep-Deconf to other causal recommenders.
5.6 Duality of Multi-Cause Confounders
From Table (3), we can find that the best results w.r.t. the confounding level, which are marked with colors blue and red for $R@20$ and $N@20$, respectively, appear in the middle of the Table. This shows an interesting phenomenon that the performance for all the methods improves first and then deteriorates with the increase of confounding level. This is against the naive intuition that the performance should reduce monotonically when the confounding level increases. The phenomenon can also be discovered from Fig. 1 of [48], but the authors provided no explanation to why it occurs.

5.6.1 Formulation of the Theory. We propose the "duality of multi-cause confounders" to explain such a phenomenon. When the confounding level is low, in the collected dataset, users tend to consume items at random, where the co-occurrence of different items contains little collaborative information. Therefore, the substitute confounder-based deconfounding models waste parameters to fit and control latent variables derived from uninformative random item exposures, which degenerates the model performances. However, in this case, the observed rating distribution is a more faithful representation of the population rating distribution, where the confounding bias in the dataset can be reduced. When the strength of confounding effects increases, although the co-occurrence of items demonstrates more regular patterns, the observed rating distribution deviates further from the true distribution of the user population. Thus, the influence of confounding bias outweighs the introduction of extra item collaborative information and reduces the performance.

In essence, the duality of multi-cause confounders results from their commonality among items, i.e., they also tend to be shared item attributes. Therefore, in contrast to their role in traditional single-cause inference problems, the confounders in recommendations, like yin and yang, exert their force in two opposite ways: On the one hand, since these confounders tend to be "shared" items attributes, they help explain why certain items tend to appear together and why certain items have never occurred at the same time. This introduces extra item collaborative information that is conducive to the recommendation of new items that are similar to the items that the user has interacted with, especially when the number of observed ratings is small and the datasets are very sparse. On the other hand, when the confounding effects are overly strong, the observed rating distribution diverges drastically from the true rating distribution, which leads to systematic bias due to the unbalanced representation of items in the dataset, where the recommendation performance can be severely degenerated. (See Appendix B.4 for an intuitive example).

5.6.2 To Deconfound or Not to Deconfound. The duality of multi-cause confounders naturally leads to a further research question: In the battle of the benefits and drawbacks of multi-cause confounders, which one will prevail? The answer, which the authors believe, is that no matter the results, a deconfounded recommender should always be preferred over a non-causality-based one if unobserved confounders indeed pervasively exist to affect the ratings, because greedily exploiting the collaborative information associated with unobserved confounders come at a price: It also inherits the confounding bias in the historical ratings, which hinders the further improvement of the recommendation model. Consider the extreme case where the exposures are entirely due to previous recommendations. In this case, the new recommender system will perform no better than the previous one by greedily exploiting. Deconfounded recommenders remedy the bias by balancing the under- and over-represented items with either sample re-weighting or controlling substitute confounders. Through these mechanisms, the recommender system can provide a more bias-free estimate of user preferences. Furthermore, the above analysis naturally leads to another conclusion that a good strategy to reduce the confounding bias from the data collection perspective is to avoid greedy exploitation and to add random exploration to the recommendation policy that collects the user ratings, where the randomness negates the influence of unobserved confounders. This is interesting, as it provides another justification for exploration, i.e., a commonly adopted
Table 4. Performance of Deep-Deconf when user features have different noise levels (N.F. means no features)

|                | VG-causal |                | ML-causal |                |
|----------------|-----------|----------------|-----------|----------------|
|                | Low Conf. ($\lambda_b = 0.3$) | High Conf. ($\lambda_b = 0.7$) | Low Conf. ($\lambda_b = 0.3$) | High Conf. ($\lambda_b = 0.7$) |
| Noise          |    R@20   |    N@20   |    R@20   |    N@20   |
| 0.1            |  0.4071  |  0.3974  |  0.4098  |  0.4039  |
| 0.5            |  0.4034  |  0.3963  |  0.4036  |  0.3982  |
| 0.9            |  0.4010  |  0.3928  |  0.3993  |  0.3934  |
| N.F.           |  0.3982  |  0.3909  |  0.3948  |  0.3912  |

strategy in reinforcement learning-based recommender systems, from a causal perspective. A brief discussion of the relationship between exploration and deconfounding can be referred to in Appendix B.5.

5.7 Sensitivity Analysis to User Features

We have proved that introducing user features as pre-treatment variables can reduce the estimand variance for Deep-Deconf as long as they are informative to rating prediction. However, how the “informative level” of user features influences the estimand variance is unclear. Recall that in our simulation, user features are generated by setting $x_u = f_{PCA}(\theta_u + \epsilon_u)$, where $\theta_u$ is the user preference and $\epsilon_u \sim N(0, \lambda_u^{-1}I_k)$ is a random Gaussian noise. We control the informative level of the user features by setting $\lambda_u^{-1} \in \{0.1, 0.5, 0.9\}$ and evaluate Deep-Deconf. The performances are compared to a baseline where no user features are used. The results are summarized in Table (4). From Table (4), we can find that user features that are more informative to rating prediction (i.e., with less noise) indeed lead to a lower estimand variance, which is reflected by a higher performance when the ratings associated with one specific item exposure is extremely sparse. Moreover, the performance improvement is more significant when the simulated confounding level is high. When the user features are highly noisy ($\lambda_f = 0.9$), however, the model overfits the noise, and the performance degenerates to the baseline model where no user features are used.

5.8 Experiments on the Real-World Datasets

In this section, we introduce the experiments conducted on two real-world datasets, i.e., Yahoo! R3 and KuaiRand, where we show the effectiveness of the proposed Deep-Deconf on real-world scenarios, where unbiased ratings collected from randomized experiments are available for testing.

5.8.1 Experimental Setups. In our setting, all biased interactions are used as the training data, 50% of the unbiased data are used for validation, and the remaining 50% of the unbiased data are used for testing. As with other papers evaluated on these datasets, we select Recall@5 (R@5) and NDCG@5 (N@5) as the metrics similar to the experiments on the simulated datasets. For Yahoo! R3, we have 7-dimensional user features collected from a questionnaire with questions on users’ willingness to rate different songs that might influence their exposure, whereas the KuaiRand dataset includes 30 different user features, including the following number, fan number, etc.
Table 5. Experimental results on the real-world Yahoo! R3 and KuaiRand datasets, where the best results are highlighted in **bold**, and the second-best results are highlighted in **underline**.

| Dataset      | Yahoo! R3 | KuaiRand |
|--------------|-----------|----------|
| Method       | R@5      | N@5     | R@5  | N@5  |
| WMF          | 0.6826   | 0.5407  | 0.3220 | 0.3705 |
| IPW-MF       | 0.6901   | 0.5498  | 0.3224 | 0.3696 |
| Deconf-MF    | 0.7213   | 0.5775  | 0.3296 | 0.3750 |
| BRD-MF       | 0.7294   | 0.5783  | 0.3249 | 0.3746 |
| Concat-VAE   | 0.7629   | 0.6395  | 0.3405 | 0.3995 |
| DRL-VAE      | 0.7694   | 0.6418  | 0.3467 | 0.4129 |
| VSR-VAE      | 0.7712   | 0.6426  | 0.3458 | 0.4083 |
| Deep-Deconf  | **0.7795** | **0.6477** | **0.3583** | **0.4201** |

5.8.2 Comparisons with State-of-the-Arts. From the above table, we can find that generally, methods with a deep backbone outperform methods with a shallow MF-based backbone. In addition, deconfounding methods perform better than their vanilla counterpart. By inferring substitute confounders via an item exposure VAE with factorized Bernoulli likelihood and incorporating it with a deep Multi-VAE-based recommendation backbone, the proposed Deep-Deconf shows better recommendation performance and more robustness to exposure bias compared with existing methods in the real-world scenario. In addition, since user features from the KuaiRand dataset are richer and more predictive to ratings than the questionnaire-based user features in the Yahoo! R3 dataset, Deep-Deconf achieves better results on the KuaiRand dataset, which further demonstrates our theoretical analysis in Section 4.3 that controlling user features that are predictive to the ratings as pre-treatment variables in the outcome prediction model can reduce the variance in estimation.

6 CONCLUSIONS

In this article, we proposed an effective deep factor model-based causal inference algorithm, Deep-Deconf, for recommender systems. By controlling substitute confounders inferred through factorized logistic VAE that render the observed exposures randomized Bernoulli trials, Deep-Deconf alleviates the multi-cause confounding bias, leading to a more faithful estimation of user preferences. Moreover, we have proved that the variance of the estimated unbiased ratings can be substantially decreased by introducing user features as pre-treatment variables. We note that our algorithm can be plugged into any user-oriented auto-encoder-based recommender system by adding a decoder branch that constrains the user-latent variable to generate factorized exposures. Therefore, we speculate that these models can also benefit from the confounding reduction advantage of our method with a modest extra computational overhead.

Appendix

A THEORETICAL ANALYSIS OF DEEP-DECONF

A.1 Proof of Conditional Independence

In the exposure model of Deep-Deconf, we claim that a decoder that takes substitute confounder $z_a$ as input and reconstructs the exposure $a_u$ with factorized logistic likelihood can renders $a_u$ conditionally independent. Some may question that this is not possible since the weights that predict $a_{ui}$ for different $i$ are shared. This is not true. The fact is that the exposures cannot be marginally independent, as they are all governed by unobserved
confounders. However, conditional on $z_u$, they can be independent, and generation of $a_u$ from $z_u$ can also be implemented via a shared decoder. The reason is as follows. For user $u$ with latent confounder $z_u$, if we denote the input vector to the last layer of the exposure network as $f(z_u)$, the logit of the exposure to item $i$ is calculated as \( \logit p(a_{ui} \mid z_u) = \mathbf{W} f(z_u) \). Since $f(z_u)$ only depend on $z_u$, if the row vectors of the last layer weights $\mathbf{W}$ are independent, $z_u$ contains all information for $a_{ui}$ contained in $a_{ui}$. Therefore, $a_{ui}$ is conditionally independent of $a_{ui}$ given $z_u$, even if the weights used to infer different $a_{ui}$ are shared among all items.

A.2 Proof of Variance Reduction via User Features as Pre-treatment Variables

In this section, we derive the sampling variance of the coefficient of exposure indicator obtained by ordinary least square (OLS) estimator before and after the introduction of user features as pre-treatment variables. Recall that after we simplify the network weight $\mathbf{W}$ to a diagonal matrix $\mathbf{w} \cdot \mathbf{I}$, user features and substitute confounders to scalars, the single-layer outcome prediction network for the ratings before the introduction of user features as pre-treatment variables becomes

\[
\hat{r}_{ui}(a_{ui}) = \mathbf{w}^T \mathbf{z}_u + \hat{w}^T a_{ui} + \hat{a}_i + \epsilon_{ui}. \tag{14}
\]

The OLS estimators for the coefficients can be specified as

\[
(\hat{w}^T, \hat{w}_i^T, \hat{a}_i) = \arg\min_{w^T, w_i^T, a_i} \sum_{u=1}^{U} (r_{ui} - \mathbf{w}_i \mathbf{z}_u + \hat{w}_i^T a_{ui} + \hat{a}_i)^2. \tag{15}
\]

As we have demonstrated in the main paper, since conditional on $z_u$, the ratings $r_{ui}$ can be viewed as generated from randomized experiments, $\hat{w}_i^T$ is an unbiased estimator for the population conditional average exposure effects of item $i$ on rating $r_{ui}$. The SUTVA assumption ensures the non-interference of exposures of different users, and the homoskedasticity assumption ensures that the variance does not vary by the change of $a_{ui}$ and $z_u$ [16]. If we further assume $\mathbb{E}[\epsilon_{ui} \mid A_{ui}, Z_u] = 0$ (the above assumptions are known as the Gaussian-Markov assumption), according to the Gauss theorem [18], the sampling variance of $\hat{w}_i^T$ can be calculated as

\[
\hat{\sigma}_i^2 = \frac{\hat{\sigma}_{R_{ui}|A_{ui},Z_u}^2}{\sum_{u=1}^{U} (a_{ui} - \bar{a}_i)^2} = s^2 \cdot \left( \frac{1}{U_i^1} + \frac{1}{U_i^0} \right), \tag{16}
\]

where $s^2 = \hat{\sigma}_{R_{ui}|A_{ui},Z_u}^2$ is the OLS variance of $R_{ui}$, $\bar{a}_i = \sum_u a_{ui} / U$ is the average exposure number of item $i$ in the finite sample, $U_i^1$ is the exposure count of item $i$ and $U_i^0 = U - U_i^1$. The second equality can be derived with

\[
\hat{\sigma}_i^2 = \frac{\hat{\sigma}_{R_{ui}|A_{ui},Z_u}^2}{\sum_{u=1}^{U} (a_{ui} - \bar{a}_i)^2} = s^2 \cdot \left( \frac{1}{U_i^1} + \frac{1}{U_i^0} \right), \tag{16}
\]

where $s^2 = \hat{\sigma}_{R_{ui}|A_{ui},Z_u}^2$ is the OLS variance of $R_{ui}$, $\bar{a}_i = \sum_u a_{ui} / U$ is the average exposure number of item $i$ in the finite sample, $U_i^1$ is the exposure count of item $i$ and $U_i^0 = U - U_i^1$. The second equality can be derived with
The authors have noticed that the deconfounder algorithm [46], based on which we designed the Deep-Deconf, relies on the pinpoint requirement of the substitute confounders. Wang & Blei have also responded to the comments in [46], [47]. The questions of [7] were solved in [48] (see footnotes on Page 2), which stated that the disagreement between Wang & Blei and Ogburn et al. lies in the assumptions required for the model identifiability. Ogburn et al. believed that Deconfounder requires extra assumptions such as the inferred substitute confounder \( z \) does not pick up post-treatment variables, etc. However, Wang et al. responded that these assumptions do not satisfy the pinpoint requirement of \( z \) (see Theorem 7 in [48]). Similarly, the identifiability of Deep-Deconf also relies on the pinpoint requirement of the substitute confounders.

Fig. 8. Item popularity distribution of the established ML-causal and VG-causal datasets.

simple algebra based on the fact that \( a_{ui} \in \{0, 1\} \) and therefore \( a_{ui}^2 = a_{ui} \). The \( \hat{\sigma}_{R_{ui} | A_{ui}, z_u}^2 \) can be calculated as the common variance across the two potential outcome distributions,

\[
\hat{\sigma}_{R_{ui} | A_{ui}, z_u}^2 = \frac{1}{U - 2} \left( \sum_{u, a_{ui} = 1} (r_{ui}(1) - \bar{r}_{ui}(1))^2 + \sum_{u, a_{ui} = 0} (r_{ui}(0) - \bar{r}_{ui}(0))^2 \right).
\] (17)

After we multiply and divide \( U \) on the R.H.S. of Eq. (16), we have \( \hat{\nu}_i \to \frac{\sigma_{\hat{\nu}_i | A_{ui}, z_u}}{U (\bar{p}_i (1 - \bar{p}_i))} \) as \( U \) approaches infinity, where \( \bar{p}_i = \lim_{U \to \infty} U_i^1 / U \) is the population probability of the exposure of item \( i \); this finishes our deduction of estimand variance before introducing the user features as pre-treatment variables. Post the introduction of the user features \( x_u \), the outcome prediction network becomes

\[
r_{ui}(a_{ui}) = w^2 z_u + w^2 a_{ui} + w^2 x_u + \alpha_i + \epsilon_{ui}.
\] (18)

Supposing again the variance does not vary by the change of the treatment indicator \( a_{ui} \), the surrogate confounder \( z_u \), and the user feature \( x_u \) (i.e., homoscedasticity), with the SUTVA assumption and the assumption of \( \mathbb{E}[\epsilon_{ui} | A_{ui}, z_u, x_u] = 0 \), the limiting sampling variance of the OLS estimator for Eq. (18), i.e., \( \hat{\nu}_i^{new} \), given the general case of \( \hat{\nu}_i \), can be directly calculated as

\[
\frac{\sigma_{\hat{\nu}_i | A_{ui}, z_u}}{U (\bar{p}_i (1 - \bar{p}_i))}.
\]

By comparing the form of \( \hat{\nu}_i^{new} \) and \( \hat{\nu}_i \) we can find that the only difference lies in the OLS variance term in the numerator: if the features of a user \( x_u \) are indicative to the prediction of her ratings \( r_{ui} \), \( \sigma_{\hat{\nu}_i | A_{ui}, z_u, x_u}^2 \) reduces considerably compared with \( \sigma_{\hat{\nu}_i | A_{ui}, z_u}^2 \), which leads to a substantial decrease of estimand variance. This is especially favorable in our multiple causal case where the low sample efficiency due to large causal space hides a precise inference.

A.3 Discussion of Model Identifiability

The authors have noticed that the deconfounder algorithm [46], based on which we designed the Deep-Deconf, has raised some discussions among researchers, including [7], [32]. Wang & Blei have also responded to the comments in [46], [47]. The questions of [7] were solved in [48] (see footnotes on Page 2), which stated that the focus of Deconf-MF is on estimating the expected potential outcome (ratings) if \( K \) extra items are exposed (i.e., top-K recommendations), so it has different assumptions with counterexamples that rely on do-operators. The disagreement between Wang & Blei and Ogburn et al. lies in the assumptions required for the model identifiability. Ogburn et al. believed that Deconfounder requires extra assumptions such as the inferred substitute confounder \( z \) does not pick up post-treatment variables, etc. However, Wang et al. responded that these assumptions do not satisfy the pinpoint requirement of \( z \) (see Theorem 7 in [48]). Similarly, the identifiability of Deep-Deconf also relies on the pinpoint requirement of the substitute confounders.
B EXPERIMENTS ON REAL-WORLD DATASETS

B.1 Schematic Illustration of Dataset Establishment

A schematic illustration for the establishment of the real-world causal datasets is shown in Fig. (7).

B.2 Item Popularity Distributions

The popularity of an item is defined as the number of users who have rated that item [43]. Since the item popularity distribution on the ML-causal and VG-causal datasets only depends on the exposure, we visualize the distribution with an arbitrary confounding effect. The results are illustrated in Fig. (8). From Fig. (8) we can discover that the item popularity distribution of both the ML-causal and VG-causal datasets exhibits both long-tail and right-skewed characteristics, which faithfully reflects the item popularity distributions in the real-world scenario.

B.3 Rating Distribution Pre- and Post- Exposure

The global rating distribution of a recommendation dataset is defined as
\[ P(R = r) \propto \sum_{u, i} 1(r_{ui} = r), \forall r \in \{1, 2, 3, 4, 5\}, \]
where \( r = 0 \) denotes the rating is unobserved is excluded from consideration. If no confounder exists, the exposure matrix \( A_{\text{rand}} \) is a random matrix where the elements in \( A_{\text{rand}} \) are independent Bernoulli variables. The individual rating distribution for a user \( \hat{R}_{ui} = R_{ui} \times A_{ui} \times A_{rand} \), the element-wise independence of \( A_{\text{rand}} \) ensures that the global and individual rating distributions of the observed ratings \( R_{\text{obs}}_{ui} \) is an unbiased estimator to those of the population ratings \( R_{p} \). Unobserved confounders, however, create a spurious dependence of \( R_{\text{obs}}_{conf} \) on \( A_{\text{conf}} \), and therefore lead to systematic bias of the rating distributions in \( R_{\text{obs}}_{conf} \) after the exposure. Although in practice, the population rating matrix \( R_{p} \) is unobtainable, in our simulation, we have the users’ ratings for all items (although only exposed ratings are visible to the algorithm for training). Therefore, in this paper, we can visualize the confounders’ effect on both global and individual rating distributions after exposure under various levels of confounding effects.

We set \( \lambda_{b} = 0 \) and \( \lambda_{d} = 0 \) where no confounding effect exists and then fix \( \lambda_{b} \) to 2, and vary \( \lambda_{d} \) from 0.1 to 0.9. Comparisons of global rating distribution are shown in Fig. (6). From Fig. (6), we can find that an obvious observation of the confounding effect is that positively rated items are more likely to be exposed than their negatively rated counterparts, and the stronger the confounding effects are, the more unbalanced the exposure of highly-rated and lowly-rated items. This is quite interesting, since all we have done to simulate the confounders is to (a) train two VAEs to model the exposure and rating distributions of the ML-1m and Amazon-VG datasets and (b) make the latent variables that generate the exposure and ratings for a user correlated by taking a weighted sum of the confounder and user variable. This phenomenon also has a real-world explanation: Users tend to rate items they like and ignore items they dislike, which leads to systematic bias due to the gross under-representation.
of items with negative ratings. Furthermore, we visualize the difference of the individual rating distributions before and after exposure under various levels of confounding effects. The averaged KL-divergence between the true and observed individual rating distribution is illustrated in Fig. (9) for reference.

B.4 More on Duality of Multi-cause Confounders

In the experiments, we have discovered that the recommendation quality improves first and then degenerates with the increase of confounding level. We conclude that “multi-cause confounders in recommender systems like yin and yang, exert their forces in two opposite ways.” In this section, we provide a simple but intuitive example to further support the claim. Suppose that in our toy system, a subset of users is “blue lovers” who rate all blue items five and all red items one —similarly, a subset of “red lovers” rate just the opposite way. Moreover, blue lovers tend to be recommended with blue items, and red lovers with red items. Since color affects both the exposure and the rating of an item, item color is a confounder in the system. In addition, it is a multi-cause confounder because color is an attribute that are shared among all the items.

The observed ratings for the two blue lovers and one red lover under low and high confounding levels are illustrated in Fig. (10). The left part of Fig. (10) shows observations under no confounding effects. In such a case, the exposure probability of an item is independent of its color (1/3 for both red items and blue items), and the observed global and individual rating distributions \( p(r = 5) = p(r = 1) = 1/2 \) exactly matches the population rating distributions. However, since the item co-occurrences are random, no item collaborative information can be utilized to recommend new items with similar rating patterns. In contrast, the right part of Fig. (10) shows rating observations with a high confounding level, where the exposure probability of an item clearly depends on its color and the user’s preference. Under this circumstance, new items with similar user rating patterns with the items have already interacted by the users can be readily recommended based on item collaborative information (e.g., item #3 to user #1, and item #1 to user #2 based on their similar rating pattern to item #2 for both users). However, the observed rating distribution (all five) deviates drastically from the population distribution, which introduces a systematic bias between the true and observed distribution that could degrade the model performance.

B.5 Exploration and Deconfounding

Although we establish the Deep-Deconf with Rubin’s causal model in the main paper, we temporarily switch to Pearl’s causal structural models to demonstrate how exploration eliminates confounding bias from the data collection perspective for better illustrative effects. Specifically, we provide three causal graphs as Fig. (11). Fig. (11) - (a) is the hypothetical causal graph that traditional non-causality-based recommender systems assume to generate the collected rating data, where the exposure of item \( V \) is independent of user \( U \). Therefore, it can also be viewed as assuming the data are generated from randomized experiments from Rubin’s causal perspective. Fig.
(a) Assumption

(b) Exploitation

(c) Exploration

**Fig. 11.** $U$: user, $V$: item, $R$: rating, $X$: interaction history, $E$: a stochastic exploration strategy. (a) The hypothetical causal graph that traditional non-causality-based recommender systems assume to generate the collected rating data, where the exposure of item $V$ is independent of user $U$. (b) The true causal graph that generates the data collected by an exploitation-based strategy, where the exposure of item $V$ is dependent on the user’s interaction history $X$. (c) The true causal graph that generates the data collected by an exploration-based strategy, where the exposure of an item is based solely on the random exploration strategy $E$.

(11) - (b) shows a simple case of the true data generation process where the recommender systems used to collect the data greedily exploit the historical user ratings and user’s preference to make recommendations. Note that if we use a naive model that assumes the data generation process of (a) to fit on rating data actually generated according to (b), the influence of unobserved confounders is mistakenly captured as the user preference, which leads to confounding bias in these models. Fig. (11) - (c) shows the causal graph where the recommender system uses a random exploration strategy (i.e., randomly picking up an item to show the users) to collect data. In (c), we can find that since the item exposure is due solely to the random exploration, the exposure bias ($U \rightarrow V$) and confounding bias ($U \leftarrow X \rightarrow V$) is negated by the randomness. Therefore, models with the assumption of (a) can still be unbiased when trained on data collected by (c). Most recommender systems are a balance between exploitation in (b) and exploration in (c). Through comparisons among the three causal graphs, we have analyzed the exploration and exploitation, i.e., two commonly used strategies in reinforcement learning-based recommender systems [2], from a causal perspective. Based on this, we have provided a new justification for exploration strategies other than reducing uncertainty in estimations of user preference.

**REFERENCES**

[1] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In Proceedings of the 12th ACM Conference on Recommender Systems. 104–112.

[2] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. 2012. A contextual-bandit algorithm for mobile context-aware recommender system. In Proceedings of International Conference on Neural Information Processing. Springer, 324–331.

[3] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2021. Bundle recommendation and generation with graph neural networks. IEEE Transactions on Knowledge and Data Engineering (2021).

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

[5] Jiawei Chen, Yan Feng, Martin Ester, Sheng Zhou, Chuan Chen, and Can Wang. 2018. Modeling users’ exposure with social knowledge influence and consumption influence for recommendation. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 953–962.

[6] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In Proceedings of the 13th ACM Conference on Recommender Systems. 101–109.

[7] Alexander D’Amour. 2019. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. arXiv preprint arXiv:1902.10286 (2019).
[8] Sihao Ding, Peng Wu, Fuli Feng, Yitong Wang, Xiangnan He, Yong Liao, and Yongdong Zhang. 2022. Addressing unmeasured confounder for recommendation with sensitivity analysis. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 305–315.

[9] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenzhang Lei, Peng Jiang, and Xiangnan He. 2022. KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 3953–3957.

[10] Junruo Gao, Mengyue Yang, Yuyang Liu, and Jun Li. 2021. Deconfounding representation learning based on user interactions in recommendation systems. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 588–599.

[11] Wenshuo Guo, Serena Wang, Peng Ding, Yixin Wang, and Michael I. Jordan. 2021. Multi-source causal inference using control variates. arXiv preprints (2021). arXiv:2103.16689 [cs.LG]

[12] Xiangnan He, Yang Zhang, Fuli Feng, Chonggang Song, Lingling Yi, Guohui Ling, and Yongdong Zhang. 2023. Addressing confounding feature issue for causal recommendation. ACM Transactions on Information Systems 41, 3 (2023), 1–23.

[13] Paul W. Holland. 1986. Statistics and causal inference. J. Amer. Statist. Assoc. 81, 396 (1986), 945–960.

[14] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In Proceedings of the Eighth IEEE International Conference on Data Mining. 263–272.

[15] Kosuke Imai and David A. Van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. J. Amer. Statist. Assoc. 99, 467 (2004), 854–866.

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

[17] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 781–789.

[18] Richard Arnold Johnson, Dean W. Wichern, et al. 2002. Multivariate Linear Regression Models. In Applied Multivariate Statistical Analysis. Chapter 7, 360–417.

[19] Dawen Liang, Minshu Zhan, and Daniel PW Ellis. 2015. Content-aware collaborative music recommendation using pre-trained neural networks.. In Proceedings of the Annual Conference of the International Society for Music Information Retrieval. 295–301.

[20] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In Proceedings of the 2018 World Wide Web Conference. 689–698.

[21] Dawen Liang, Minhui Zhan, and Daniel PW Ellis. 2015. Content-aware collaborative music recommendation using pre-trained neural networks.. In Proceedings of the Annual Conference of the International Society for Music Information Retrieval. 295–301.

[22] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2021. Mitigating confounding bias in recommendation via information bottleneck. In Proceedings of the 15th ACM Conference on Recommender Systems. 351–360.

[23] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. Proceedings of the 31st ACM International Conference on Neural Information Processing Systems (2017).

[24] Yunan Luo, Jian Peng, and Jianzhu Ma. 2020. When causal inference meets deep learning. Nature Machine Intelligence 2, 8 (2020), 426–427.

[25] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. 2021. Multi-cause effect estimation with disentangled confounder representation. In Proceedings of the Thirty-first International Joint Conference on Artificial Intelligence.

[26] Jing Ma, Mingjing Wan, Longqi Yang, Jundong Li, Brent Hecht, and Jaime Teevan. 2022. Learning causal effects on hypergraphs. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 1202–1212.

[27] Benjamin M Marlin and Richard S Zemel. 2009. Collaborative prediction and ranking with non-random missing data. In Proceedings of the Third ACM Conference on Recommender systems. 3–12.

[28] Elizabeth L Ogburn, Ilya Shpitser, and Eric J Tchetgen Tchetgen. 2019. Comment on “blessings of multiple causes”. J. Amer. Statist. Assoc. 114, 528 (2019), 1611–1615.

[29] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. 2020. Deep structural causal models for tractable counterfactual inference. Proceedings of the International Conference on Neural Information Processing System 33 (2020).

[30] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. Causal inference in statistics: A primer. John Wiley & Sons.

[31] Bruno Pradel, Nicolas Usunier, and Patrick Gallinari. 2012. Ranking with non-random missing ratings: influence of popularity and positivity on evaluation metrics. In Proceedings of the Sixth ACM Conference on Recommender Systems. 147–154.
[36] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked causal variational autoencoder for inferring paired spillover effects. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 1679–1682.

[37] Donald B Rubin. 1984. Bayesianly justifiable and relevant frequency calculations for the applies statistician. The Annals of Statistics (1984), 1151–1172.

[38] Donald B Rubin. 1990. Formal model of statistical inference for causal effects. Journal of Statistical Planning and Inference 25, 3 (1990), 279–292.

[39] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In Proceedings of the International Conference on Machine Learning. 1670–1679.

[40] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. Proceedings of IEEE (2021).

[41] Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 713–722.

[42] Harald Steck. 2011. Item popularity and recommendation accuracy. In Proceedings of the fifth ACM conference on Recommender systems. 125–132.

[43] Harald Steck. 2013. Evaluation of recommendations: Rating-prediction and ranking. In Proceedings of the 7th ACM conference on Recommender systems. 213–220.

[44] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2020. Causal inference for recommender systems. In arXiv preprint arXiv:2003.04948.

[45] Tianxin Wei, Fulí Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 1791–1800.

[46] Yixin Wang and David M Blei. 2019. The Blessings of Multiple Causes. J. Amer. Statist. Assoc. 114, 528 (2019), 1574–1596.

[47] Yixin Wang and David M Blei. 2020. Towards clarifying the theory of the deconfounder. arXiv preprint arXiv:2003.04948 (2020).

[48] Yixin Wang, Dawn Liang, Laurent Charlin, and David M Blei. 2020. Causal inference for recommender systems. In Proceedings of the Fourteenth ACM Conference on Recommender Systems. 426–431.

[49] Minghao Zhao, Qilin Deng, Kai Wang, Runze Wu, Jianrong Tao, Changjie Fan, Liang Chen, and Peng Cui. 2021. Bilateral filtering graph convolutional network for multi-relational social recommendation in the power-law networks. ACM Transactions on Information Systems 40, 2 (2021), 1–24.

[50] Xiaolin Zheng, Menghan Wang, Renjun Xu, Jianmeng Li, and Yan Wang. 2020. Modeling dynamic missingness of implicit feedback for sequential recommendation. IEEE Transactions on Knowledge and Data Engineering 34, 1 (2020), 405–418.

[51] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for sequential recommendation. In Proceedings of the Web Conference 2021. 2980–2991.

[52] Shuyuan Xu, Juntao Tan, Shelby Heincke, Vena Jia Li, and Yongfeng Zhang. 2023. Deconfounded causal collaborative filtering. ACM Transactions on Intelligent Systems and Technology 11, 4 (2023), 1–25.

[53] Mengyue Yang, Quanyu Dai, Zhenhua Dong, Xu Chen, Xiuqiang He, and Jun Wang. 2021. Top-n recommendation with counterfactual user preference simulation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2342–2351.

[54] Chenxiu Xu, Hao Huang, and Shinjae Yoo. 2019. Scalable causal graph learning through a deep neural network. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 1853–1862.

[55] Shuyuan Xu, Juntao Tan, Shelby Heincke, Vena Jia Li, and Yongfeng Zhang. 2023. Deconfounded causal collaborative filtering. ACM Transactions on Recommender Systems 1, 4 (2023), 1–25.

[56] Jingsen Zhang, Xu Chen, and Wayne Xin Zhao. 2021. Causally attentive collaborative filtering. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 3622–3626.

[57] Yang Zhang, Fulí Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention explanation, and generalization. arXiv preprint arXiv:2301.00910 (2023).
[63] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. 2020. Counterfactual prediction for bundle treatment. *Proceedings of Advances in Neural Information Processing Systems 33* (2020).