Going Beyond Popularity and Positivity Bias

Correcting for Multifactorial Bias in Recommender Systems

Huang, J.; Oosterhuis, H.; Mansoury, M.; van Hoof, H.; de Rijke, M.

DOI
10.1145/3626772.3657749

Publication date
2024

Document Version
Final published version

Published in
SIGIR '24

Going Beyond Popularity and Positivity Bias: Correcting for Multifactorial Bias in Recommender Systems

Jin Huang
University of Amsterdam
Amsterdam, The Netherlands
j.huang2@uva.nl

Harrie Oosterhuis
Radboud University
Nijmegen, The Netherlands
harrie.oosterhuis@ru.nl

Masoud Mansoury*
Delft University of Technology
Delft, The Netherlands
m.mansoury@tudelft.nl

Herke van Hoof
University of Amsterdam
Amsterdam, The Netherlands
h.c.vanhoof@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Two typical forms of bias in user interaction data with recommender systems (RSs) are popularity bias and positivity bias, which manifest themselves as the over-representation of interactions with popular items or items that users prefer, respectively. Debiasing methods aim to mitigate the effect of selection bias on the evaluation and optimization of RSs. However, existing debiasing methods only consider single-factor forms of bias, *e.g.,* only the item (popularity) or only the rating value (positivity). This is in stark contrast with the real world where user selections are generally affected by multiple factors at once. In this work, we consider multifactorial selection bias in RSs. Our focus is on selection bias affected by both item and rating value factors, which is a generalization and combination of popularity and positivity bias. While the concept of multifactorial bias is intuitive, it brings a severe practical challenge as it requires substantially more data for accurate bias estimation. As a solution, we propose smoothing and alternating gradient descent techniques to reduce variance and improve the robustness of its optimization. Our experimental results reveal that, with our proposed techniques, multifactorial bias corrections are more effective and robust than single-factor counterparts on real-world and synthetic datasets.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Recommender Systems; Unbiased Learning; Propensity Estimation

ACM Reference Format:
Jin Huang, Harrie Oosterhuis, Masoud Mansoury, Herke van Hoof, and Maarten de Rijke. 2024. Going Beyond Popularity and Positivity Bias: Correcting for Multifactorial Bias in Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657749

*Work done while the author was with Elsevier Discovery Lab and University of Amsterdam.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657749

1 INTRODUCTION

Rating prediction is a fundamental recommender system (RS) task where the goal is to predict user ratings on items. The task facilitates personalized recommendations to improve user satisfaction [6, 42, 43]. Rating prediction methods that are learned from user ratings can be biased as user interactions with RSs are subject to severe selection bias [32, 37, 41, 48, 49]. The effects of such bias can produce systematic errors in user preference prediction [18, 48, 58] and result in problems of over-specialization [3], filter bubbles [35, 39], and unfairness [10]. Two influential types of bias present in user rating behavior are popularity bias [8, 41, 49] and positivity bias [41], which arise as users are more likely to rate popular items or items that they prefer, respectively.

Single-factor bias. Widely-used methods for mitigating the effect of selection bias in user ratings make use of inverse propensity scoring (IPS) [20] and integrate it into the learning process [18, 21, 48]. Given the propensity of a rating, *i.e.,* the probability of the corresponding user rating the specific item, IPS weights each rating inversely to their propensity, and, thereby, corrects for the over-representation resulting from selection bias. The predominant model of popularity bias in previous work assumes that the propensity values only depend on the corresponding item. For positivity bias, the propensity values are assumed to only depend on the corresponding rating value. These single-factor propensity models can provide unbiased estimations with IPS, given that their assumptions about the factors that determine the selection bias in user data are correct. However, real-world user decisions about rating items generally depend on more than one factor, a scenario that existing methods do not address in practice [14, 19, 41].

Multifactorial bias. We consider a *multifactorial* bias that is determined by two factors, *i.e.,* item and rating value. This can be seen as a generalization of popularity and positivity bias that combines the essential properties of both. As we expect multifactorial bias to better capture actual user behavior, we also expect that the resulting propensities will lead to a better performance of IPS-based debiasing methods. To estimate multifactorial bias, existing propensity estimation methods [48], based on naive Bayes or logistic regression, can simply be used by accommodating multiple factors. Surprisingly, there is a lack of studies comparing the performance of IPS-based debiasing methods using single-factor bias

estimation against those using such a multifactorial bias estimation. This raises questions about the practical utility of multifactorial bias estimation and correction. Moreover, our experimental results on real-world datasets indicate that existing multifactorial bias estimation methods lead to unstable performance when applied to IPS-based debiasing methods. This could potentially explain their limited adoption in practice.

The practical challenges associated with multifactorial bias arise as the consideration of multiple factors greatly increases problems of data sparsity [12, 42]. For comparison, single-factor popularity bias estimation is based on the observation frequency of ratings per item, *i.e.,* how many users have rated an item. Single-factor positivity bias estimation is based on the difference in frequency of rating values between naturally observed ratings and a (small) unbiased dataset, *i.e.,* how much more often or less often a rating value is observed in natural user interactions than when users rate randomly sampled items. Both single-factor estimation techniques already have to deal with severe sparsity, as most items are not very popular and often only very little unbiased data is available [7, 12, 42]. Multifactorial bias estimation exacerbates this sparsity problem as it has to consider the frequencies of combinations of items and rating values. As a result, before a multifactorial bias approach can be effective, one has to first overcome this severe data-efficiency problem.

Contributions and findings. In this work, we develop a propensity estimation method for multifactorial bias that is determined by item and rating value factors. The results of our proposed multifactorial bias propensity estimation are integrated with an IPS-based debiasing method to correct for multifactorial bias. To deal with the severe sparsity problem multifactorial bias poses, we propose the adoption of propensity smoothing technique and an alternating gradient descent approach for more robust and stable IPS-based optimization.

To evaluate our multifactorial method, we compare the IPS-based debiasing method using our multifactorial bias estimation against those using single-factor bias estimation on a selection of real-world datasets: the Yahoo!R3 [32], Coat [48], and KuaiRec [16] datasets. Our experimental results show the effectiveness of our multifactorial method over state-of-the-art single-factor counterparts. Furthermore, we perform an extensive simulation-based experimental analysis where the effect of each of the two factors is varied. The results show that single-factor methods are only effective when their corresponding factor dominates selection bias, but perform poorly when the other factor is also important. In contrast, our multifactorial approach has much more robust performance, as it is always effective, regardless of how much effect each factor has, and provides considerably better performance when both factors have a substantial effect. This indicates that, once its sparsity problem is dealt with, our multifactorial approach provides the safest choice when it is unclear what factors determine selection bias.

2 CONCEPTUALIZATION OF SELECTION BIAS

In this section, we provide an overview of existing conceptualizations of selection bias, popularity bias, and positivity bias in the context of RSs. Some concepts have seen varied definitions across publications, potentially leading to confusion in their usage.

Selection bias. Ovaisi et al. [37] conclude that selection bias occurs when a data sample is not representative of the underlying data distribution. Primary studies delineate selection bias into two principal categories [17, 44, 48, 55]: self-selection bias, where users choose to interact with certain items more often, and algorithmic bias, where items showing to users are highly dependent on the algorithm in an RS. In this paper, we adopt a definition of selection bias in line with Ovaisi et al. [37].

In some studies, the definition of selection bias slightly differs. Chen et al. [10] constrain selection bias exclusively to self-selection bias while delineating algorithmic bias as exposure bias. Exposure bias could be known as "previous model bias" when the previous recommendation policy controls what items to show [28], or "user-selection bias" in the scenario where the RS shows the items according to users' active search queries [54].

Popularity bias. Prior work mostly takes popularity to be a form of selection bias defined by Ovaisi et al. [37] or exposure bias defined by Chen et al. [10]. Popularity bias is often defined based on two primary reasons for occurrence: users are more likely to provide feedback on popular items [41, 49], and popular items are recommended more frequently than their popularity would warrant [2, 9, 31, 62]. Due to popularity bias, the observed logged data reveals a concentration of user interactions on popular items, shown as a long-tail distribution in the frequency of interactions across items. Therefore, there exists a widely shared consensus that popularity bias is closely associated with the long-tail phenomenon [2, 10, 49].

Positivity bias. Positivity bias is another form of selection bias, which is uniformly considered to refer to the scenario where users rate more often the items they like [40, 41]. In contrast, a rarely studied but relevant form of selection bias could occur when users rate more often the items they dislike.[1] These forms of bias can contribute to a scenario where observed user ratings are characterized by a skewed rating distribution compared to the true rating distribution [18, 41].

3 PRELIMINARIES

Before we define and address multifactorial bias specifically, we introduce our problem setting, provide a formal definition of selection bias, and summarize the IPS-based debiasing methods.

We follow the common RS setting where users from a set $\mathcal{U} = \{u_1, \ldots, u_N\}$ give ratings on items from a set $\mathcal{I} = \{i_1, \ldots, i_M\}$ [50]. User preferences are explicitly shown by these ratings, $y_{u,i} \in \mathcal{R} = \{1, 2, 3, 4, 5\}$ per user $u \in \mathcal{U}$ and item $i \in \mathcal{I}$. In practice, logged rating data $\mathcal{D}$ is often very sparse and subject to heavy selection bias as it is unrealistic for all users to provide ratings for all items. To indicate which ratings are available for optimization, we use an observation indicator matrix $\boldsymbol{O} \in \{0, 1\}^{|\mathcal{U}||\mathcal{I}|}$, where $o_{u,i} \in \boldsymbol{O}$ indicates whether the rating for user u on item i is recorded in the logged data ($o_{u,i} = 1$) or not ($o_{u,i} = 0$). One can expect $\boldsymbol{O}$ to be sparse and influenced by selection bias [18, 48, 49]. Next, we define several forms of selection bias and discuss their effects

[1] We avoid the use of the term negativity bias, as it commonly denotes a scenario where wrong impressions may sometimes outweigh good ones [4].

on rating prediction methods that learn from logged rating data: $\mathcal{D} = \{(u, i, y_{u,i}) \mid u \in \mathcal{U}, i \in \mathcal{I}, o_{u,i} = 1\}$.

3.1 Definition of selection bias

As discussed in Section 2, selection bias occurs if the process that decides whether a user rates an item is not a random selection. We formally define selection bias by using the propensities $p_{u,i}$, *i.e.,* the probabilities of a user rating an item: $p_{u,i} = P(o_{u,i} = 1 \mid u, i, y_{u,i})$.

Definition 3.1 (Selection bias). Logged rating data $\mathcal{D}$ is subject to *selection bias* if not every rating propensity has the same value:

$$\textit{Selection-bias}(\mathcal{D}) \iff \exists u, u' \in \mathcal{U}, \exists i, i' \in \mathcal{I},\ p_{u,i} \neq p_{u',i'}. \quad (1)$$

We further provide the following definitions of two influential forms of selection bias – positivity bias and popularity bias – to match our usage of the terms:

Definition 3.2 (Positivity bias). Logged rating data $\mathcal{D}$ is subject to *positivity bias* if propensities only depend on their rating values (Fig. 1a) and higher ratings correspond to higher propensities:

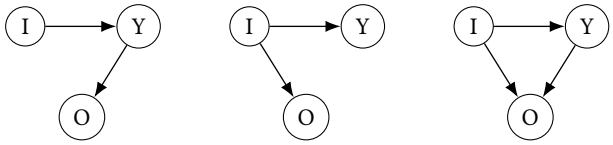
$$\begin{aligned}\textit{Positivity-bias}(\mathcal{D}) \iff \big(&\textit{Selection-bias}(\mathcal{D}) \wedge \\ &\forall u, u' \in \mathcal{U}, \forall i, i' \in \mathcal{I},\ \big(y_{u,i} > y_{u',i'} \longleftrightarrow p_{u,i} > p_{u',i'}\big)\big).\end{aligned} \quad (2)$$

Definition 3.3 (Popularity bias). Logged rating data $\mathcal{D}$ is subject to *popularity bias* if the propensities of ratings only depend on which item they correspond to (Fig. 1b):

$$\begin{aligned}\textit{Popularity-bias}(\mathcal{D}) \iff \big(&\textit{Selection-bias}(\mathcal{D}) \wedge \\ &\forall u, u' \in \mathcal{U}, \forall i, i' \in \mathcal{I},\ \big(i = i' \longrightarrow p_{u,i} = p_{u',i'}\big)\big).\end{aligned} \quad (3)$$

As discussed in Section 2, the definition of each form of bias exclusively focuses on the presence of its corresponding factor influences, excluding consideration of any other factors or biases. Importantly, our definitions only consider what variables the propensities of ratings depend on. Thereby, our usage of the terms is only concerned with the specific pattern the selection bias follows, and not with its resulting effects. In this regard, our approach contrasts with prior work that identifies types of selection bias by the highly-skewed rating distributions that they can produce [1, 10, 41, 49]. For example, a long-tailed rating distribution where a few items receive the most ratings (*e.g.,* Fig. 2b) is sometimes referred to as popularity bias or evidence thereof [1, 10, 49]. Similarly, a difference between rating value frequencies from natural user behavior and ratings on randomly sampled items (*e.g.,* Fig. 2a) is sometimes referred to as (evidence of) positivity bias [41].

However, these skewed distributions can occur for many reasons, and therefore, it is difficult to use their observation as evidence for a specific form of selection bias. For example, a long-tailed rating distribution could result from positivity bias per Definition 3.2: if there are only a few items with high rating values then these items will get the most ratings. Vice versa, the differences between rating distributions could result from popularity bias per Definition 3.3 in a case where the more popular items happen to have a higher rating on average (see Fig. 2c). To avoid this ambiguity and since our focus is on how selection bias should be modeled, we explicitly choose to base our definitions around the dependencies of propensities and will use the terms popularity bias and positivity bias accordingly.



(a) (b) (c)

Figure 1: The dependency between observance (O), items (I), and rating values (Y) for different bias assumptions: (a) Positivity bias: propensities only depend on rating values; (b) Popularity bias: propensities only depend on items; (c) Multifactorial bias: propensities depend on both factors.

3.2 Rating prediction from user ratings

Our goal is to optimize an RS model that best predicts the user ratings across all items. This is achieved by minimizing a loss function that compares the actual ratings $y_{u,i}$ and the predicted ratings $\hat{y}_{u,i}$:

$$\mathcal{L} = \frac{1}{|\mathcal{U}|\,|\mathcal{I}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \delta(\hat{y}_{u,i}, y_{u,i}), \quad (4)$$

where the comparison function δ can be an RS metric, *i.e.,* the commonly-used mean squared error (MSE): $\delta(\hat{y}_{u,i}, y_{u,i}) = (\hat{y}_{u,i} - y_{u,i})^2$.

The loss function in Eq. 4 represents our ideal goal but assumes that all ratings are available, something that is rarely the case in practice. A straightforward but naive estimate of the ideal goal is to average over the observed ratings in the logged data $\mathcal{D}$:

$$\mathcal{L}_{\text{Naive}} = \frac{1}{|\mathcal{D}|} \sum_{u,i \in \mathcal{D}} \delta(\hat{y}_{u,i}, y_{u,i}). \quad (5)$$

However, this naive estimate ignores the effect of selection bias and assumes that every rating is equally probable to be observed [48]. As a result, if logged data $\mathcal{D}$ is subject to selection bias, it is biased by rating propensities:

$$\mathbb{E}_o[\mathcal{L}_{\text{Naive}}] = \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} p_{u,i}\, \delta(\hat{y}_{u,i}, y_{u,i}) \not\propto \mathcal{L}. \quad (6)$$

3.3 IPS-based debiasing method

To mitigate the effect of selection bias, widely-used methods make use of inverse propensity scoring (IPS) [20] and integrate it into the learning process [18, 21, 48]. IPS weights each rating inversely to its propensity, $p_{u,i}$, and, thereby, corrects for the over- and under-representation resulting from selection bias:

$$\mathcal{L}_{\text{IPS}} = \frac{1}{|\mathcal{U}|\,|\mathcal{I}|} \sum_{u,i \in \mathcal{D}} \frac{\delta(\hat{y}_{u,i}, y_{u,i})}{p_{u,i}}. \quad (7)$$

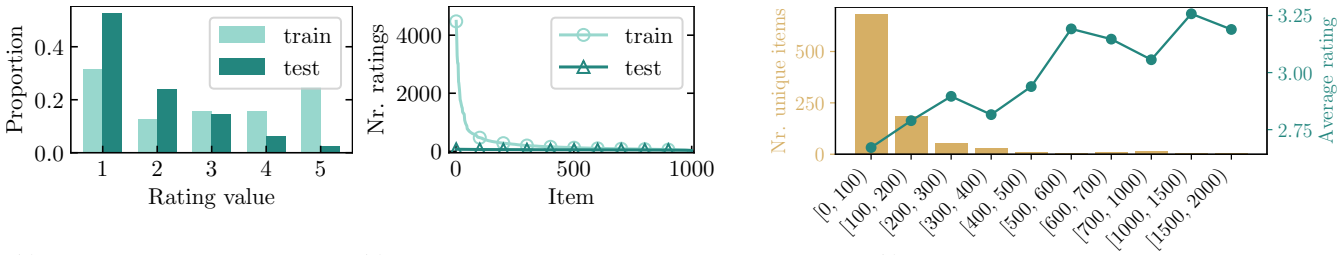
Thus, IPS gives more weight to observed ratings with small propensities and less weight to those with large propensities. As $\mathbb{E}_o[o_{u,i}] = p_{u,i}$, the IPS loss provides an unbiased estimate of $\mathcal{L}$:

$$\mathbb{E}_o[\mathcal{L}_{\text{IPS}}] = \frac{1}{|\mathcal{U}|\,|\mathcal{I}|} \sum_{u} \sum_{i} \frac{\mathbb{E}_o[o_{u,i}]}{p_{u,i}} \delta(\hat{y}_{u,i}, y_{u,i}) = \mathcal{L}. \quad (8)$$

Combined with a recommendation method, *e.g.,* matrix factorization (MF), IPS reduces the effect of bias in predicting user ratings.

3.4 Existing single-factor propensity estimation

IPS for rating estimation requires propensity estimation because propensities cannot be observed directly, since the exact way users



(a) Distribution of rating values. **(b) Distributions of interactions over items.** **(c) Item group by item popularity.**

Figure 2: Skewed distributions of (a) rating values or (b) item popularity in the logged training set (train) of the Yahoo!R3 dataset, and (c) the number and average ratings of items in a group that contains items with the number of interactions falling within a certain interval are counted from logged user ratings on the self-selected songs in the Yahoo!R3 dataset.

decide to rate items is not directly accessible. Methods exist for estimating propensities under our definitions of positivity bias (Definition 3.2) and popularity bias (Definition 3.3). Importantly, each existing method only corresponds to one of the definitions and thus assumes that propensities only depend on a single factor.

The predominant method of positivity bias estimation in previous work uses Bayes' rule [48]:

$$\hat{p}^{\text{pos}}_{u,i} = P(o = 1 \mid y = y_{u,i}) = \frac{P(y = y_{u,i} \mid o = 1)\, P(o = 1)}{P(y = y_{u,i})}. \quad (9)$$

The observation prior is estimated by the observation frequency: $P(o = 1) \approx |\mathcal{D}|/(|\mathcal{U}|\,|\mathcal{I}|)$, and the conditional rating-value probability estimate is the frequency of the rating in the observed data: $P(y = r \mid o = 1) \approx \sum_{u,i \in \mathcal{D}} \mathbb{1}[y_{u,i} = r]/|\mathcal{D}|$. Finally, to estimate the rating-value prior, a small sample of unbiased (missing completely at random (MCAR)) data $\mathcal{M}$ is used; such data could be obtained by having users rate randomly sampled items. The prior estimate is simply the rating-value frequency in $\mathcal{M}$: $P(y = r) = \sum_{u,i \in \mathcal{M}} \mathbb{1}[y_{u,i} = r]/|\mathcal{M}|$. Putting these components into Eq. 9, we see that positivity bias propensities are estimated as follows:

$$\hat{p}^{\text{pos}}_{u,i} = P(o = 1 \mid y = y_{u,i}) \approx \frac{|\mathcal{M}| \sum_{u',i' \in \mathcal{D}} \mathbb{1}[y_{u',i'} = y_{u,i}]}{|\mathcal{U}||\mathcal{I}| \sum_{u',i' \in \mathcal{M}} \mathbb{1}[y_{u',i'} = y_{u,i}]}. \quad (10)$$

The most widely-used model of popularity bias computes propensities on items based on item popularity [46, 57]:

$$\hat{p}^{\text{pop}}_{u,i} = P(o = 1 \mid i) \approx \frac{\sum_{u'} o_{u',i}}{\sum_{u'} \sum_{i'} o_{u',i'}}. \quad (11)$$

These estimated propensities may be small, especially for tail items, thus causing high variance in the IPS estimation. Propensity clipping is usually used as a variance reduction technique [51]; it clips propensity scores by a small value τ: $\bar{p}_{u,i} = \max(\hat{p}_{u,i}, \tau)$. Here, τ trades off the bias and variance of the IPS estimation with the clipped estimated propensities: If $\tau = 1$, it approaches the naive estimation, while if $\tau = 0$, it approaches the unbiased estimation.

With the corresponding estimated propensities, the IPS estimator can be used to mitigate the effect of popularity bias or positivity bias. However, existing single-factor forms of bias do not account for the fact that real-world user decisions toward rating items generally depend on more than one factor [14, 19, 41].

4 CORRECTION FOR MULTIFACTORIAL BIAS

In contrast with existing single-factor models of bias, we consider a multifactorial bias that is determined by two factors: the item and rating value. After defining our multifactorial bias, we introduce a stable propensity estimation method for it by adopting propensity smoothing technique. We use IPS-based optimization with our novel estimated propensities, resulting in an unbiased rating prediction method that corrects for multifactorial bias.

4.1 Definition of multifactorial bias

Multifactorial bias occurs if the process that decides whether a user provides a rating is not a random selection and is determined by multiple factors. In this paper, we consider a specific multifactorial bias that is determined by two factors: the item and rating value.

Definition 4.1 (Multifactorial bias). Logged rating data $\mathcal{D}$ is subject to *multifactorial bias* if the propensities of ratings depend on which item they correspond to and their rating values (Fig. 1c):

$$\textit{Multifactorial-bias}(\mathcal{D}) \iff \big(\textit{Selection-bias}(\mathcal{D}) \wedge \quad (12)$$
$$\forall u, u' \in \mathcal{U}, \forall i, i' \in \mathcal{I},\ (i = i' \wedge y_{u,i} = y_{u',i'}) \longrightarrow p_{u,i} = p_{u',i'}\big).$$

This definition encompasses any selection bias determined by both item and rating value factors and can naturally be extended to various types of multifactorial bias.

4.2 Propensity estimate for multifactorial bias

A novel method is required to estimate multifactorial propensities $p_{u,i} = P(o = 1 \mid y = y_{u,i}, i)$ that vary over different combinations of items and rating values. We propose to decompose the multifactorial propensity with Bayes' rule:

$$\hat{p}^{\text{mul}}_{u,i} = P(o = 1 \mid y = y_{u,i}, i) = \frac{P(y = y_{u,i}, i \mid o = 1) P(o = 1)}{P(y = y_{u,i}, i)}, \quad (13)$$

and use a maximum likelihood estimate for each component. Our observation prior estimate is the observation frequency: $P(o = 1) \approx |\mathcal{D}|/(|\mathcal{U}|\,|\mathcal{I}|)$. Our conditional joint rating-value and item probability estimate is their frequency in the observation data $\mathcal{D}$:

$$P(y = r, i \mid o = 1) \approx \sum\nolimits_{u,i' \in \mathcal{D}} \mathbb{1}[i' = i \wedge y_{u,i'} = r]/|\mathcal{D}|, \quad (14)$$

and our joint rating-value and item prior estimate is their joint frequency in the small unbiased (MCAR) data $\mathcal{M}$:

$$P(y = r, i) \approx \sum\nolimits_{u,i' \in \mathcal{M}} \mathbb{1}[i' = i \wedge y_{u,i'} = r]/|\mathcal{M}|. \quad (15)$$

While conceptually this propensity estimation is straightforward, it brings a severe practical challenge as it relies on the frequencies of combinations of items and rating values in the sparse observation data $\mathcal{D}$ and the even sparser unbiased data $\mathcal{M}$. As a result, estimates of the joint probabilities can be extremely small or even zero,

and, thereby, potentially result in invalid propensity estimates or extremely-high-variance IPS estimates.

To address these sparsity issues, we apply Laplace smoothing [30] to both the estimations of the joint conditional probability and joint prior. The conditional joint rating-value and item probability estimate is smoothed with parameter α_1:

$$P(y = r, i \mid o = 1) \approx \frac{\sum_{u,i' \in \mathcal{D}} \mathbb{1}[i' = i \wedge y_{u,i'} = r] + \alpha_1}{|\mathcal{D}| + \alpha_1 |\mathcal{I}|\,|\mathcal{R}|}. \tag{16}$$

The estimated joint rating-value and item prior is smoothed by α_2:

$$P(y = r, i) \approx \underbrace{\frac{\sum_{u,i' \in \mathcal{M}} \mathbb{1}[y_{u,i'} = r]}{|\mathcal{M}|}}_{\text{Estimate of } P(y=r).} \cdot \underbrace{\frac{\sum_{u,i' \in \mathcal{M}} \mathbb{1}[i' = i \wedge y_{u,i'} = r] + \alpha_2}{\sum_{u,i' \in \mathcal{M}} \mathbb{1}[y_{u,i'} = r] + \alpha_2 |\mathcal{I}|}}_{\text{Smoothed estimate of } P(i \mid y=r).}. \tag{17}$$

Instead of directly smoothing the joint prior $P(y = r, i)$, we decompose it into the product of the prior $P(y = r)$ and the conditional $P(i \mid y = r)$ and only smooth the latter. We found that this provided the most robust performance; most likely because item sparsity is much more extreme than rating-value sparsity.

4.3 A debiasing method for multifactorial bias

Using the results of our multifactorial bias propensity estimation, a rating prediction model can be optimized with IPS while accounting for multifactorial bias. Following Schnabel et al. [48], we choose inverse-propensity-scored matrix factorization (MF-IPS) as the debiased rating prediction method. With the propensity estimates $\hat{p}^{\text{mul}}_{u,i}$, we have our multifactorial method: MF-IPS^{Mul}. It minimizes the multifactorial IPS estimate of the MSE between the predicted ratings and the actual ratings with an added L_2-regularization term:

$$\mathcal{L}_{\text{MF-IPS}^{Mul}}(\Theta) = \frac{1}{|\mathcal{D}|} \sum_{u,i \in \mathcal{D}} \frac{\delta(\hat{y}_{u,i}, y_{u,i})}{\hat{p}^{\text{mul}}_{u,i}} + \lambda ||\Theta||_2^2, \tag{18}$$

where a predicted rating is computed by a standard MF: $\hat{y}_{u,i} = \boldsymbol{p}_u^\top \boldsymbol{q}_i + a_u + b_i + c$, which is the inner-product of embedding vectors $\boldsymbol{p}_u$ and $\boldsymbol{q}_i$ for user u and item i, together with user, item and global offsets a_u, b_i and c; and the parameter set $\Theta = \{\boldsymbol{p}_u, \boldsymbol{q}_i, a_u, b_i, c\}$ includes all parameters of MF.

In the optimization of our multifactorial method, we could follow common stochastic gradient descent and iteratively sample a batch of data and update parameter $\theta \in \Theta$ according to gradient of the loss function on each data batch using the Adam optimizer [22]:

$$\theta_t = \text{ADAM}(\theta_{t-1}, \nabla_{\theta_{t-1}} \mathcal{L}_{\text{MF-IPS}^{Mul}}). \tag{19}$$

However, we found this concurrent gradient descent method in IPS-based optimization to be unstable in experiments on real-world data (see Section 5). Many data batches contain widely varied propensity estimates, and due to the very low propensities under multifactorial bias, this appears to result in severe instability between updates.

An existing alternative to the concurrent gradient descent is the alternating least squares (ALS) method [53]. ALS iteratively alternates between optimizing user and item embeddings via least squares to reduce optimization instability. The alternating updates mitigate the effect of noise and outlier interactions [53]. We build on the idea of alternating gradient descent from ALS and extend it to optimize generic loss functions using the Adam optimizer. Algorithm 1 shows the procedure of optimizing MF-IPS^{Mul} with our alternating gradient descent method. It begins with parameter initialization, then updates parameters over multiple epochs according to the loss on logged user ratings $\mathcal{D}$. The optimization continues until the stop condition is reached, *e.g.,* decreasing performance on the validation set or reaching a predefined number of epochs. Importantly, in each epoch, the item-related parameters $\boldsymbol{q}_i, b_i$ (line 6–8) and other parameters $\boldsymbol{p}_u, a_u, c$ (line 3–5) are updated independently and alternately. Thereby, our optimization alternately updates a subset of parameters while keeping the remaining parameters fixed in each epoch. Our experimental results on real-world data indicate this leads to increased stability and robustness (see Section 5).

Algorithm 1: Our optimization method for MF-IPS^{Mul} with our alternating gradient descent approach.

Input: Observed rating data: $\mathcal{D}$; estimated propensities: $\hat{p}$.
Output: MF-IPS^{Mul} parameters: $\boldsymbol{p}_u, \boldsymbol{q}_i, a_u, b_i, c$.

1 Initialize parameters $\boldsymbol{p}_u, \boldsymbol{q}_i, a_u, b_i, c$;
2 **while** *stop condition is not reached* **do**
  /* Epoch to update global & user embeddings and offsets. */
3 **for** *each batch of* $(u, i, y_{u,i})$ *in a random ordering of* $\mathcal{D}$ **do**
4 Update parameters $\boldsymbol{p}_u, a_u, c$ according to Eq. 19;
5 **end**
  /* Epoch to update item embeddings and offsets. */
6 **for** *each batch of* $(u, i, y_{u,i})$ *in a random ordering of* $\mathcal{D}$ **do**
7 Update parameters $\boldsymbol{q}_i, b_i$ according to Eq. 19;
8 **end**
9 **end**

This completes the description of our method to mitigate the effects of multifactorial bias. It optimizes a MF model for rating predictions using IPS with multifactorial bias propensity estimation that considers both item and rating value factors. In addition, we adopt propensity smoothing and alternating gradient descent to make our multifactorial method feasible and robust in practice.

5 EXPERIMENTS ON REAL-WORLD DATA

Our experimental analysis on real-world datasets aims to answer two research questions: (**RQ1**) Does our proposed multifactorial method better mitigate the effect of bias in logged rating data than existing single-factor debiasing methods? (**RQ2**) How do varying smoothing parameters and our alternating gradient descent approach affect our multifactorial method?

5.1 Experimental setup

Our experiments are based on two real-world datasets: Yahoo!R3 [32] and Coat [48], which are publicly available and widely used to evaluate debiasing methods.[2] Both have a training set consisting of biased ratings and a MCAR test set of user ratings on uniformly randomly selected items. We filter the users that do not appear in

[2]The KuaiRec dataset [16] contains biased user interactions (a sparse subset) and a set of fully observed user-item interactions (a dense subset). However, as highlighted by Lin et al. [27], its density (16.3% and 99.9% for the sparse and dense subsets, respectively) surpasses that of other datasets, diverging from our targeted bias and sparsity problem, as explained in Section 1. We extend our evaluation beyond the Yahoo!R3 and Coat datasets by conducting a *simulation* on KuaiRec in Section 6.

Table 1: Performance comparison for predicting ratings on the Yahoo!R3 and Coat datasets. Results are means of 10 independent runs with standard deviations in brackets. † indicates that our multifactorial method MF-IPS^{Mul} with alternating gradient descent significantly outperforms all other existing methods (paired-samples t-test ($p < 0.01$)).

Dataset	Method	MSE		MAE		RMSE		RMSE_U		RMSE_I	
Yahoo!R3	Avg	2.1321		1.2671		1.4602		1.4167		1.4153	
	MF	1.8296	(0.0318)	1.1305	(0.0173)	1.3526	(0.0117)	1.2593	(0.0159)	1.3325	(0.0130)
	VAE	1.4182	(0.0082)	0.9677	(0.0039)	1.1909	(0.0034)	1.1158	(0.0034)	1.1694	(0.0033)
	MF-IPS^{MF}	1.7877	(0.0297)	1.0621	(0.0024)	1.3370	(0.0111)	1.2140	(0.0050)	1.3067	(0.0109)
	MF-IPS^{Pop}	1.9432	(0.0048)	1.1425	(0.0058)	1.3940	(0.0017)	1.2783	(0.0046)	1.3711	(0.0008)
	MF-IPS^{Pos}	0.9891	(0.0013)	0.7928	(0.0079)	0.9945	(0.0006)	0.9267	(0.0048)	0.9774	(0.0015)
	MF-IPS^{Mul} (ours)	**0.9629**†	(0.0015)	**0.7700**†	(0.0120)	**0.9813**†	(0.0007)	**0.9071**†	(0.0075)	**0.9626**†	(0.0025)
Coat	Avg	1.6521		1.0904		1.2854		1.2521		1.2605	
	MF	1.2916	(0.0108)	0.9283	(0.0074)	1.1365	(0.0048)	1.0907	(0.0049)	1.1085	(0.0049)
	VAE	1.1393	(0.0048)	0.8583	(0.0038)	1.0674	(0.0023)	1.0282	(0.0027)	1.0424	(0.0021)
	MF-IPS^{MF}	1.1597	(0.0175)	0.8687	(0.0165)	1.0769	(0.0082)	1.0366	(0.0076)	1.0512	(0.0074)
	MF-IPS^{Pop}	1.2284	(0.0142)	0.9042	(0.0115)	1.1083	(0.0064)	1.0666	(0.0066)	1.0828	(0.0066)
	MF-IPS^{Pos}	1.1728	(0.0120)	0.8708	(0.0129)	1.0830	(0.0055)	1.0395	(0.0073)	1.0576	(0.0069)
	MF-IPS^{Mul} (ours)	**1.1020**†	(0.0007)	**0.8552**†	(0.0023)	**1.0498**†	(0.0003)	**1.0110**†	(0.0009)	**1.0275**†	(0.0006)

the test sets to make predictions more precise, resulting in 129,179 biased ratings and 54,000 unbiased ratings of 5,400 users to 1,000 items in the Yahoo!R3 dataset, and 6,960 biased ratings and 4,640 unbiased ratings of 290 users to 300 items in the Coat dataset, respectively. The biased ratings are partitioned into a training and validation set according to a ratio of 4:1. To estimate propensities, we set aside 5% and 20% of the original test sets as the small unbiased data $\mathcal{M}$ for the Yahoo!R3 and Coat datasets, respectively. This ensures at least two interactions per item for estimating the conditional joint rating-value and item distribution.

To evaluate our method, we adopt evaluation metrics widely used in previous work [48, 50, 55]: MSE, root mean square error (RMSE), and mean absolute error (MAE). We further report the average RMSE performance per user (RMSE_U) and item (RMSE_I) [33], *i.e.,* we calculate the RMSE score for each individual user/item separately and then average them.

We evaluate our multifactorial method MF-IPS^{Mul} by comparing it with the following baselines: (i) Avg, MF, and VAE [26] that ignore bias altogether. Avg simply predicts the average observed rating of each item: $\hat{y}_{u,i} = \frac{\sum_{u',i \in \mathcal{D}} y_{u',i}}{|\{(u',i,y_{u',i}) \in \mathcal{D}\}|}$. VAE has been proposed to apply variational autoencoders to collaborative filtering. We adopt Gaussian log-likelihood in VAE for rating predictions. (ii) MF-IPS^{MF}, a debiased method with propensity estimation using MF with logistic regression [17, 46, 48]. It has the potential to correct for multifactorial bias as it uses MF to model bias through learned multiple hidden factors. (iii) MF-IPS^{Pop} and MF-IPS^{Pos}, two debiased methods with single-factor popularity bias estimation and single-factor positivity bias estimation, respectively.

Additionally, to evaluate how our proposed alternating gradient descent approach affects rating prediction models, all MF-based models are optimized by two optimization methods: (i) *Concurrent* gradient descent: all parameters of methods are updated concurrently; (ii) *Alternating* gradient descent: the item-related parameters and other parameters are updated alternately.

Hyperparameters used in the MF-based methods are tuned per propensity estimation in the following range: the learning rate $\eta \in \{10^{-3}, 10^{-4}, 10^{-5}\}$, the L_2 regularization weights $\lambda \in \{10^{-7}, 10^{-6}, \ldots, 10^{-2}\}$, and the dimension of embeddings of users and items $d \in \{16, 32, 64, 128\}$. Hyperparameter tuning for VAE is conducted as follows: (i) aligning the learning rate, the regularization weights, and the dimension of the latent representation with the same range employed for MF-based methods; and (ii) adjusting the parameter that controls the strength of the Kullback-Leibler term within the range $\{0.05, 0.1, 0.2, 0.4, \ldots, 1.0\}$. For debiasing methods with multifactorial bias estimation, we also choose the smoothing parameters $\alpha_1, \alpha_2 \in \{1, 2, \ldots, 10\}$. Additionally, propensity clipping and normalization are used to reduce variance and improve the robustness of methods. Our experimental implementation is available at https://github.com/BetsyHJ/MultifactorialBias.

5.2 Overall performance

Table 1 displays our main experimental results on the Yahoo!R3 and Coat datasets. We make the following three observations. First, among all the methods, Avg has the worst performance; this is expected as it provides non-personalized predictions and ignores selection bias. Accordingly, MF does model individual user preferences and outperforms Avg. Importantly, VAE exhibits a considerable performance margin over MF due to its generative capabilities.

Second, the debiasing methods that consider the effect of bias improve the performance: MF-IPS $\succ$ MF (except for MF-IPS^{Pop} on Yahoo!R3).[3] A strong indication of the negative effect that selection bias has on rating prediction optimization.

Third, in debiasing methods, positivity bias estimation performs better than popularity bias estimation, but worse than multifactorial bias estimation: MF-IPS^{Mul} $\succ$ MF-IPS^{Pos} $\succ$ MF-IPS^{Pop}. This suggests that positivity bias has a stronger effect than popularity

[3]We write $A \succ B$ to indicate that method A outperforms method B.

Table 2: Performance comparison among MF-based methods when optimization is done with concurrent and alternating gradient descent on the Yahoo!R3 and Coat datasets. Results are means of 10 independent runs with standard deviations in brackets. † indicates that the method optimized by the alternating gradient descent method significantly outperforms the identical method optimized by the concurrent gradient descent method (paired-samples t-test ($p < 0.01$)).

Dataset	Method	Concurrent			Alternating		
		MSE	MAE	RMSE	MSE	MAE	RMSE
Yahoo!R3	MF	1.8296 (0.0318)	1.1305 (0.0173)	1.3526 (0.0117)	1.8335 (0.0236)	1.1688 (0.0077)	1.3540 (0.0088)
	MF-IPS^{MF}	1.7877 (0.0297)	1.0621 (0.0024)	1.3370 (0.0111)	1.7143† (0.0172)	1.0616 (0.0168)	1.3093† (0.0066)
	MF-IPS^{Pop}	1.9432 (0.0048)	1.1425 (0.0058)	1.3940 (0.0017)	1.9055† (0.0196)	1.1659 (0.0077)	1.3804† (0.0071)
	MF-IPS^{Pos}	0.9891 (0.0013)	0.7928 (0.0079)	0.9945 (0.0006)	0.9762† (0.0034)	0.7943 (0.0099)	0.9880† (0.0017)
	MF-IPS^{Mul} (ours)	0.9812 (0.0067)	0.7737 (0.0116)	0.9905 (0.0034)	**0.9629**† (0.0015)	**0.7700** (0.0120)	**0.9813**† (0.0007)
Coat	MF	1.2916 (0.0108)	0.9283 (0.0074)	1.1365 (0.0048)	1.2040† (0.0119)	0.9034† (0.0208)	1.0973† (0.0054)
	MF-IPS^{MF}	1.1597 (0.0175)	0.8687 (0.0165)	1.0769 (0.0082)	1.1641 (0.0154)	0.8730 (0.0287)	1.0789 (0.0072)
	MF-IPS^{Pop}	1.2284 (0.0142)	0.9042 (0.0115)	1.1083 (0.0064)	1.1923† (0.0049)	0.8787† (0.0124)	1.0919† (0.0022)
	MF-IPS^{Pos}	1.1728 (0.0120)	0.8708 (0.0129)	1.0830 (0.0055)	1.1717 (0.0065)	0.8672 (0.0106)	1.0825 (0.0030)
	MF-IPS^{Mul} (ours)	1.1397 (0.0295)	**0.8503** (0.0199)	1.0675 (0.0138)	**1.1020**† (0.0007)	0.8552 (0.0023)	**1.0498**† (0.0003)

bias in rating predictions. Despite the potential to capture multifactorial forms of bias, MF-IPS^{MF} does not always outperform MF-IPS^{Pos}, suggesting that it cannot adequately learn multifactorial bias. Nevertheless, multifactorial bias estimation provides the most robust and best overall performance; MF-IPS^{Mul} significantly outperforms all other methods on both datasets. By considering the effect of multiple factors on selection bias, the multifactorial method can better capture and correct for bias in real-world data.

Overall, the best-performing method is our multifactorial debiasing method with alternating gradient descent. Therefore, we answer **RQ1** in the affirmative: The proposed multifactorial method MF-IPS^{Mul} better mitigates the effect of bias in logged rating data than methods designed for single-factor biases.

5.3 Smoothing and alternating gradient descent

To better understand the effect of propensity smoothing and alternating gradient descent, we perform the following additional analyses. Due to space limitations, some analyses are limited to the Yahoo!R3 dataset only. First, we look at how the performance of our multifactorial method changes when varying the smoothing parameters. Fig. 3 shows the MSE performance obtained for different smoothing parameters: α_1 and α_2 (see Eq. 16 and Eq. 17). We see that the highest performance is reached with $\alpha_1 = 10$ and $\alpha_2 = 2$, however, there is clearly a wide range of smoothing parameters that provide close to optimal performance. It appears that it is mainly important not to set the parameters too small, as the worst performance is reached with $\alpha_1 = 1$ and $\alpha_2 = 1$. The combined results of Fig. 3 and Table 1 reveal that the smoothing parameters do not need fine-tuning for the multifactorial method to outperform all other methods. Thus, we conclude that propensity smoothing is an effective and robust enhancement for multifactorial debiasing.

Second, we compare MF-based methods optimized by the concurrent method against those optimized by the alternating method, as shown in Table 2. These performance improvements are considerably enhanced with our alternating gradient descent method, which boosts the performance of MF-IPS^{Mul} on all datasets and all

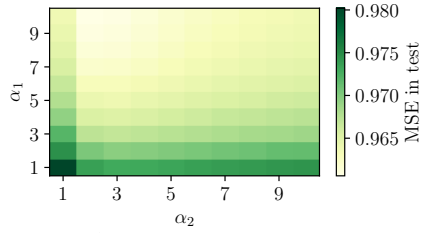


Figure 3: (Yahoo!R3) The effect of varying smoothing parameters α_1 and α_2 on MSE obtained by our multifactorial method.

metrics (with the exception of MAE on Coat). Performance gains are also seen for other methods but not as consistent as for MF-IPS^{Mul}. Due to the smaller multifactorial propensities, MF-IPS^{Mul} has more variance during optimization, and therefore, alternating gradient descent can provide a more consistent improvement here.

We further compare the learning curves of our multifactorial method when optimization is done with the concurrent and alternating gradient descent. Fig. 4 displays these in terms of the self-normalized IPS-weighted MSE performance [52] on the validation set and the MSE performance on the test set. Clearly, the alternating method exhibits more stable and faster learning than the concurrent method in the early stages of learning. While both converge around 500 epochs, the concurrent method converges to a slightly better MSE-IPS performance on the validation set compared to the alternating method. However, we see that this actually results in a slightly worse MSE performance on the test set, suggesting the concurrent method is more prone to overfitting. Therefore, it appears that alternating gradient descent is indeed less influenced by noise and outliers than the concurrent method, which we think is why it provides more stable and robust optimization.

Finally, we answer **RQ2**: propensity smoothing provides robust performance improvements to our multifactorial method and does not need fine-tuning; alternating gradient descent leads to less variance in learning curves and less overfitting than concurrent gradient descent. These advantages substantially increase the robustness, stability, and performance of our multifactorial method.

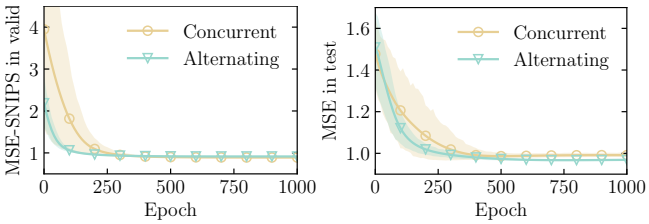


Figure 4: (Yahoo!R3) Learning curves tracking self-normalized IPS-weighted MSE on the validation set and MSE on the test set obtained by our multifactorial method. Results are means over 10 independent runs, shared areas show the 95% confident intervals calculated by using bootstrapping [13].

6 EFFECT OF BIASES ON USER RATINGS

We turn to our final research question: (**RQ3**) Can our multifactorial method MF-IPS^{Mul} robustly mitigate the effect of selection bias in scenarios where the effect of two factors on bias is varied?

6.1 Experimental setup for RQ3

Due to a lack of real-world datasets with different effects of each factor, we utilize a semi-synthetic setup. We simulate a short-video rating scenario by sampling user ratings on videos under different forms of selection bias. Our sampling source is the KuaiRec dataset [16] as it provides a fully observed user-item interaction matrix where 1,411 users rate almost all 3,327 items.

Since the dataset does not contain ratings but watch ratios on videos, we first convert these into 5-star user ratings. First, we sort the watch ratios in ascending order and then give the top 51.48% a rating of $y = 1$, the next 25.25% get $y = 2$, etc., such that the resulting ratings follow the rating distribution of the Yahoo!R3 dataset: $P(y = 1) = 0.5148$, $P(y = 2) = 0.2525$, $P(y = 3) = 0.1496$, $P(y = 4) = 0.0554$ and $P(y = 5) = 0.0277$.

The biased training set is constructed by sampling ratings with multifactorial selection bias. To simulate the joint effect of rating value and item factors, we first introduce two single-factor propensities: $\rho^{(\mathrm{R})}$ which is only dependent on the rating values, and $\rho^{(\mathrm{I})}$ which is only dependent on the items. Our simulated multifactorial propensity is then simply a linear interpolation between the two:

$$P(o = 1 \mid y = r, i) = \gamma \rho_r^{(\mathrm{R})} + (1 - \gamma)\rho_i^{(\mathrm{I})}, \tag{20}$$

where $\gamma \in [0, 1]$ controls the effect of each factor on the selection bias. Our simulation also covers single-factor scenarios: if $\gamma = 0.0$, the selection bias is *popularity bias*, only determined by the item factor; if $\gamma = 1.0$, it is *positivity bias*, only determined by the rating value factor. Importantly, when $\gamma \in (0, 1)$, the resulting selection bias is multifactorial as it is affected by both factors.

Our rating-value propensities are $\rho^{(\mathrm{R})} = [0.0123, 0.0102, 0.0213, 0.0568, 0.1795]$ corresponding to the ratings $[1, 2, 3, 4, 5]$. These values were chosen to match the positivity bias propensities estimated on the Yahoo!R3 datasets, and they lead to an expectation of ratings higher than 3 being over-represented. Item propensities are generated according to a power-law distribution following Bellogín et al. [5]: $\rho^{(\mathrm{I})} = (\eta - 1) \cdot (\mathrm{rank}(i)/k_{\min})^{-\eta}$, where $\mathrm{rank}(i) \in [1, |\mathcal{I}|]$ is the position of item i when sorted by their average ratings descendingly, and we set the power-law exponent $\eta = 1.4$ and the minimum value $k_{\min} = 20$. Hereby, more popular items have a higher rating on average as is often seen in real-world data (*e.g.,* Fig. 2c).

Some of our methods need a small unbiased MCAR set and we need an unbiased test set for evaluation. We sample unbiased data by uniform-randomly selecting 40 ratings from each user's ratings across all items. From this data, we set aside 20% for the small MCAR set and use the remaining 80% as the test set.

To answer **RQ3**, we compare the performance of our multifactorial method MF-IPS^{Mul} to that of MF with and without debiasing methods for single-factor bias correction: MF-IPS^{Pop} and MF-IPS^{Pos}. Additionally, we also consider debiasing with the ground truth propensities: MF-IPS^{GT}. This provides an unrealistic skyline that is only possible in a simulation setting where the true propensities are known. Due to space limitations, we only report MSE and MAE under optimization with alternating gradient descent.

6.2 Results for RQ3

Fig. 5 shows the performance of the different MF with various debiasing methods, under multifactorial selection bias, as γ varies the effects of the rating-value and item factors.

We first consider when γ equals 0, and the simulated selection bias reduces to popularity bias. Here, we see that MF-IPS^{Pos} performs worst and that MF-IPS^{Mul} and MF-IPS^{Pop} have performance comparable and similar to MF. This shows that assuming selection bias is dependent on only the rating value factor can substantially hurt performance when it is actually only dependent on the item factor. However, it appears that assuming dependency on both factors does not hurt performance at all, in this scenario.

Next, we consider when γ equals 1, and the simulated selection bias reduces to positivity bias. Here, we observe that MF and MF-IPS^{Pop} perform worse than all other methods by a large margin; and that MF-IPS^{Pos} has the best performance, while our multifactorial method MF-IPS^{Mul} performs slightly worse. This strongly suggests that assuming selection bias is dependent only on the item factor is detrimental to performance when it is in fact dependent only on the rating value factor. In contrast, the multifactorial model also made an incorrect assumption: a dependency on both factors, but this only resulted in a relatively small performance decrease.

Finally, we turn our attention to all other cases: where $\gamma \in (0, 1)$ and the selection bias is multifactorial bias. We see that as γ gets closer to 0 or 1, the performance of the corresponding single-factor debiasing method increases. In contrast, the performance of our multifactorial approach (MF-IPS^{Mul}) is much more stable for all values of γ, and its MSE value closely approximates that of the ground-truth method MF-IPS^{GT}. When $\gamma < 0.7$, MF-IPS^{Mul} has a substantially lower MSE than MF-IPS^{Pos}, and when $\gamma > 0.1$ the MSE of MF-IPS^{Mul} is substantially lower than MF-IPS^{Pop}. There is an exception when $\gamma > 0.8$, when MF-IPS^{Mul} is outperformed by MF-IPS^{Pos} and MF-IPS^{GT} by a small but noticeable margin.

Similar observations can be made in terms of MAE performance, however, the MAE results have more variance making the trends less clearly apparent. The increased variance is likely because all methods optimize the MSE in their loss, and thus, they do not necessarily fully minimize the MAE in the process.

Overall, our results show that the performance of single-factor debiasing methods varies greatly depending on how much selection bias is affected by their corresponding factor. Conversely, the

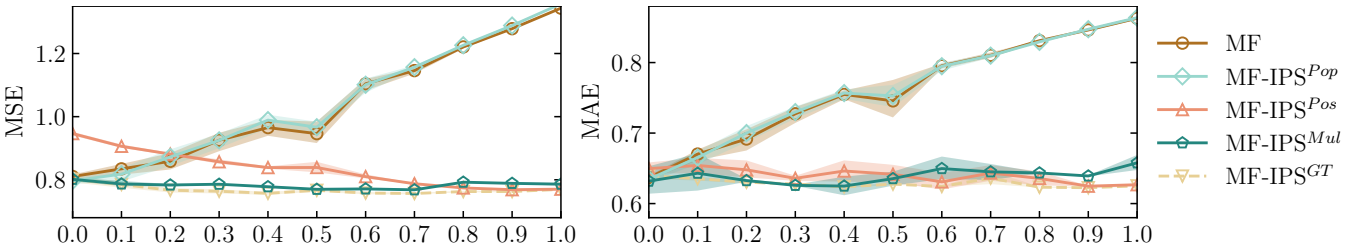


Figure 5: Performance in our simulated setting with different dependencies of bias on item and rating value factors through varying γ (x-axis, Eq. 20). Results are means over 10 independent runs; shared areas show 95% bootstrap confident intervals [13].

performance of our multifactorial method is hardly affected by how much selection bias depends on each factor, with only showing a minor decrease when selection bias is very close to positivity bias. Therefore, we answer **RQ3** in the affirmative: we conclude that our multifactorial method has the most robust performance and is the safest choice if selection bias could depend on multiple factors.

7 RELATED WORK

Selection bias is pervasive in user interactions with RSs and can be observed in both explicit feedback (*e.g.,* user ratings) [17, 48] and implicit feedback (*e.g.,* user clicks) [46, 57]. As discussed in Section 2, selection bias can arise for a variety of reasons, resulting in different forms of selection bias, such as the well-known popularity bias [2, 41, 49, 62] and positivity bias [40, 41]. Other forms of bias include incentive bias, manifested when users are incentivized to provide ratings for benefits and rewards [38], and conformity bias, manifested when users tend to rate items similarly to others in a group [23, 24]. Such forms of bias are determined by one factor, referred to as single-factor bias.

In reality, selection bias in user interactions can be characterized as a combination of multiple biases or a complex bias that is determined by more than one factor [56, 60]. Previous work suggests that selection bias is also affected by the additional factor of time [17]. Many contextual factors such as position, modality or surrounding items can result in selection bias in user rating behavior simultaneously [47, 56, 63]. Additionally, correlations between selection and both popularity and positivity were observed in multiple real-world datasets [18, 41]. Building on this, our focus in this paper is on a multifactorial bias determined by item and rating value factors, which can be seen as a generalization of popularity and positivity.

Debiased recommendation methods aim to mitigate the negative effects of bias and involve both bias estimation and correction [10, 48, 62]. A prevalent family of debiasing methods is based on inverse propensity scoring (IPS) [20, 21, 48]. IPS weights observations inversely to their observation probability; in theory, its estimation is unbiased but can suffer from high variance [48]. Propensity clipping [11, 46] and doubly-robust estimation [36, 45, 55] are two common ways to reduce variance for IPS. An alternative research direction involves two-tower methods, which jointly model user-item interactions and estimate bias present in the interactions [63]. Due to a lack of explicit signals of the bias effect on interactions, two-tower methods encounter challenges in distinguishing between user preference modeling and bias estimation [15]. In light of this, for our proposed multifactorial method we have chosen to build on the IPS-based debiasing method.

Bias or propensity estimation aims to estimate the probability of a user interacting with an item [25, 29, 34, 48, 61]. It is a key component in IPS weighting and significantly affects the performance of IPS in mitigating bias. A prevalent method for propensity estimation uses naive Bayes with maximum likelihood, which is commonly used to estimate popularity bias and positivity bias [8, 48, 59]. An alternative for propensity estimation is based on optimizing machine learning models. E.g., logistic regression and MF models can be trained to predict propensities that can best generate an observation matrix [17, 46, 48]. While the idea of estimating propensities through optimization is conceptually appealing, our experiments show that these estimates are often unstable and do not always provide propensities that work well with IPS.

8 CONCLUSION

We have considered a multifactorial selection bias that is determined by two factors: the item and rating value. We introduced a propensity estimation method for multifactorial bias and integrated it into the prevalent IPS-based debiasing approach. Furthermore, we proposed the adoption of propensity smoothing and a novel alternating gradient descent method to deal with the sparsity problem that arises in multifactorial bias estimation.

Our experimental results on two real-world datasets show the effectiveness of our multifactorial method over state-of-the-art single-factor counterparts. Moreover, through a simulation analysis, we found that the performance of our multifactorial method is stable as the effect of different factors is widely varied, in stark contrast with existing single-factor methods. Thereby, our multifactorial approach appears to be both substantially more robust and significantly effective than previous single-factor debiasing techniques. Our multifactorial debiasing approach could be an important contribution to the RS field, as multifactorial bias appears to better capture real-world forms of bias.

A limitation of our work is that we only consider multifactorial bias in explicit feedback and the rating prediction task. Future work could extend our multifactorial method to implicit feedback and other recommendation settings, *e.g.,* large language models as RSs.

ACKNOWLEDGMENTS

This research was partially supported by the Dutch Research Council (NWO) under project numbers, 024.004.022, NWA.1389.20.183, KICH3.LTP.20.006, and grant No. VI.Veni.222.269, and by the European Union's Horizon Europe program under No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Himan Abdollahpouri. 2019. Popularity bias in ranking and recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 529–530.

[2] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided exposure bias in recommendation. *arXiv preprint arXiv:2006.15772* (2020).

[3] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Over-specialization and Concentration Bias of Recommendations: Probabilistic Neighborhood Selection in Collaborative Filtering Systems. In *ACM Recsys*. 153–160.

[4] Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology* 5, 4 (2001), 323–370.

[5] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2017. Statistical Biases in Information Retrieval Metrics for Recommender Systems. *Information Retrieval Journal* 20 (2017), 606–634.

[6] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender Systems Survey. *Knowledge-based systems* 46 (2013), 109–132.

[7] Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-adapted Interaction* 12 (2002), 331–370.

[8] Rocío Cañamares and Pablo Castells. 2018. Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 415–424.

[9] Òscar Celma and Pedro Cano. 2008. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*. 1–8.

[10] 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.

[11] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k Off-policy Correction for a REINFORCE Recommender System. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 456–464.

[12] Joao Felipe Guedes da Silva, Natanael Nunes de Moura Junior, and Luiz Pereira Caloba. 2018. Effects of Data Sparsity on Recommender Systems Based on Collaborative Filtering. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[13] Thomas J DiCiccio and Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical science* 11, 3 (1996), 189–228.

[14] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 172–186.

[15] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research* 23, 1 (2022), 5232–5270.

[16] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A Fully-observed Dataset and Insights for Evaluating Recommender Systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 540–550.

[17] Jin Huang, Harrie Oosterhuis, and Maarten de Rijke. 2022. It Is Different When Items are Older: Debiasing Recommendations When Selection Bias and User Preferences are Dynamic. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 381–389.

[18] Jin Huang, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof. 2020. Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning based Recommender Systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 190–199.

[19] Nouhaila Idrissi, Ahmed Zellou, and Zohra Bakkoury. 2023. "Guess Why I Didn't Rate It": A New Preference-based Model for Enhanced Top-K Recommendation. *International Journal of Intelligent Engineering and Systems* 16, 3 (2023), 542–551.

[20] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

[21] 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.

[22] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*.

[23] Norman Knyazev and Harrie Oosterhuis. 2022. The Bandwagon Effect: Not Just Another Bias. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. 243–253.

[24] Sanjay Krishnan, Jay Patel, Michael J. Franklin, and Ken Goldberg. 2014. A Methodology for Learning, Analyzing, and Mitigating Social Influence Bias in Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender systems*. 137–144.

[25] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. 2023. Propensity matters: Measuring and enhancing balancing for recommendation. In *International Conference on Machine Learning*. PMLR, 20182–20194.

[26] 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.

[27] Zinan Lin, Dugang Liu, Weike Pan, Qiang Yang, and Zhong Ming. 2023. Transfer learning for collaborative recommendation with biased and unbiased data. *Artificial Intelligence* 324 (2023), 103992.

[28] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 831–840.

[29] Zhongzhou Liu, Yuan Fang, and Min Wu. 2024. Estimating Propensity for Causality-based Recommendation without Exposure Data. *Advances in Neural Information Processing Systems* 36 (2024).

[30] Christopher D. Manning, Raghavan Prabhakar, and Schütze Hinrich. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[31] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.

[32] 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*. ACM, 5–12.

[33] Paolo Massa and Paolo Avesani. 2007. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*. 17–24.

[34] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. 2004. Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods* 9, 4 (2004), 403.

[35] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web*. 677–686.

[36] Harrie Oosterhuis. 2023. Doubly Robust Estimation for Correcting Position Bias in Click Feedback for Unbiased Learning to Rank. *ACM Transactions on Information Systems* 41, 3 (2023), 1–33.

[37] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for Selection Bias in Learning-to-Rank Systems. In *Proceedings of The Web Conference 2020*. 1863–1873.

[38] Umberto Panniello, Shawndra Hill, and Michele Gorgoglione. 2016. The Impact of Profit Incentives on the Relevance of Online Recommendations. *Electronic Commerce Research and Applications* 20 (2016), 87–104.

[39] Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin.

[40] Kunwoo Park, Meeyoung Cha, and Eunhee Rhim. 2018. Positivity bias in customer satisfaction ratings. In *Companion Proceedings of the The Web Conference 2018*. 631–638.

[41] 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.

[42] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2010. Introduction to Recommender Systems Handbook. In *Recommender Systems Handbook*. Springer, 1–35.

[43] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender Systems: Introduction and Challenges. *Recommender Systems Handbook* (2015), 1–34.

[44] Yuta Saito. 2020. Asymmetric tri-training for debiasing missing-not-at-random explicit feedback. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 309–318.

[45] Yuta Saito, Hayato Sakata, and Kazuhide Nakata. 2019. Doubly Robust Prediction and Evaluation Methods Improve Uplift modeling for Observational Data. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 468–476.

[46] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased Recommender Learning from Missing-not-at-random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 501–509.

[47] Fatemeh Sarvi, Ali Vardasbi, Mohammad Aliannejadi, Sebastian Schelter, and Maarten de Rijke. 2023. On the Impact of Outlier Bias on User Clicks. In *SIGIR 2023: 46th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 18–27.

[48] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *International Conference on Machine Learning*. PMLR, 1670–1679.

[49] Harald Steck. 2011. Item Popularity and Recommendation Accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems*. 125–132.

[50] Harald Steck. 2013. Evaluation of Recommendations: Rating-prediction and

Ranking. In *Proceedings of the 7th ACM Conference on Recommender Systems*. 213–220.

[51] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. 2010. Learning from Logged Implicit Exploration Data. In *Advances in Neural Information Processing Systems*, Vol. 23.

[52] Adith Swaminathan and Thorsten Joachims. 2015. The Self-normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, Vol. 28.

[53] Gábor Takács and Domonkos Tikk. 2012. Alternating Least Squares for Personalized Ranking. In *Proceedings of the Sixth ACM Conference on Recommender Systems*. 83–90.

[54] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 115–124.

[55] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly Robust Joint Learning for Recommendation on Data Missing Not At Random. In *International Conference on Machine Learning*. PMLR, 6638–6647.

[56] Xinwei Wu, Hechang Chen, Jiashu Zhao, Li He, Dawei Yin, and Yi Chang. 2021. Unbiased Learning to Rank in Feeds Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 490–498.

[57] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased Offline Recommender Evaluation for Missing-not-at-random Implicit Feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 279–287.

[58] Sirui Yao, Yoni Halpern, Nithum Thain, Xuezhi Wang, Kang Lee, Flavien Prost, Ed H. Chi, Jilin Chen, and Alex Beutel. 2021. Measuring Recommender System Effects with Simulated Users. *arXiv preprint arXiv:2101.04526* (2021).

[59] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–20.

[60] Zhi Zheng, Zhaopeng Qiu, Tong Xu, Xian Wu, Xiangyu Zhao, Enhong Chen, and Hui Xiong. 2022. CBR: Context Bias Aware Recommendation for Debiasing User Modeling and Click Prediction. In *Proceedings of the ACM Web Conference 2022*. 2268–2276.

[61] Ziwei Zhu, Yun He, Yin Zhang, and James Caverlee. 2020. Unbiased implicit recommendation and propensity estimation via combinational joint learning. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 551–556.

[62] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. 2021. Popularity bias in dynamic recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2439–2449.

[63] Honglei Zhuang, Zhen Qin, Xuanhui Wang, Michael Bendersky, Xinyu Qian, Po Hu, and Dan Chary Chen. 2021. Cross-positional Attention for Debiasing Clicks. In *Proceedings of the Web Conference 2021*. 788–797.