Calibrating Multiple Robust Learning for Causal Recommendation

Shuxia Gong[1], **Chen Ma**[2]

[1]Mogo Inc. [2]Renmin University of China
silencesliver@hotmail.com machen2001@ruc.edu.cn

Abstract

The ratings in recommendation systems (RS) are missing not at random (MNAR) due to the biased selection of the items to rate, resulting in inaccurate rating prediction for all user-item pairs. Doubly robust (DR) learning has been studied in many tasks in RS with a single imputation or a single propensity model, in addition, multiple robust (MR) has been proposed with multiple imputation models and propensity models, and is unbiased when there exists a linear combination of these imputation models and propensity models is correct. However, we claim that the imputed errors and propensity scores are miscalibrated in the MR method. In this paper, we propose a calibrated multiple robust learning method to enhance the debiasing performance and reliability of the rating prediction model. Specifically, we propose to use bi-level optimization to solve the weights and model coefficients of each propensity and imputation model in the MR framework. Moreover, we adopt the differentiable expected calibration error as part of the objective to optimize the model calibration quality significantly. Experiments on three real-world datasets show that our method outperforms the state-of-the-art baselines.

Introduction

Recommendation systems (RS) is an effective tool to address the problem of information overload and has been widely used in e-commerce, social media, and entertainment (Ricci, Rokach, and Shapira 2010). RS aims to predict user preferences for items based on collected historical interaction data (Wang et al. 2019; Schnabel et al. 2016). However, the collected data cannot include all ratings from users to items, resulting in inevitably missing due to users' self-selection behavior, i.e., users can choose the item to rate freely, the missing is non-random, which is also known as selection bias problem (Chen et al. 2022; Wu et al. 2022). The selection bias indicates that the collected dataset is not representative for the target population of interest (all user-item pairs), and the training distribution differs from the target test distribution. Ignoring such distributional shift will inevitably lead to sub-optimal recommendation performance (Steck 2010; Schnabel et al. 2016; Wang et al. 2019). To address the selection bias, one line of previous research proposed to use error imputation-based (EIB) methods, which first impute the missing ratings and then train the prediction model based on both observed and imputed ratings (Chang et al. 2010; Steck 2010). Additionally, another category of methods leverages propensity scores, which computes the probability of an event being observed, to reweight the observed ratings and align the distribution of observed data with the target population (Saito et al. 2020; Schnabel et al. 2016). Furthermore, Doubly Robust (DR) method combines the error imputation and the inverse propensity re-weighting to achieve double robustness, which means the DR estimator achieves unbiasedness if either the imputed errors or the learned propensities are correct (Morgan and Winship 2015; Saito 2020; Wang et al. 2019). Furthermore, the Multiple Robust (MR) method is proposed to mitigate inaccuracies in single-model propensity scores or error imputations found in DR method (Li et al. 2023a). By considering multiple candidate propensity and imputation models, MR estimator achieves unbiasedness if any of the propensity models, imputation models, or a linear combination of these models accurately estimate the true propensities or prediction errors.

However, we argue that the imputed errors and estimated propensity scores are miscalibrated in the existing MR method, which cannot reflect the ground-truth likelihood of the correctness of the true error or true propensity. For instance, if we have 100 user-item pairs with estimated propensity scores equal to 0.2, there should be exactly 20 ratings being observed and 80 ratings being unobserved. Although previous study has proposed to adopt calibration experts to calibrate the single propensity model and imputation model in DR estimator (Kweon and Yu 2024), this approach cannot be directly extended to the MR estimator, as calibrating each model individually is expensive and unreasonable due to the unbiasedness condition of MR in terms of linear combinations is not considered. Furthermore, the calibration metric previously used in (Kweon and Yu 2024) is non-differentiable and cannot be directly optimized.

To fill this gap, we propose the calibrated multiple robust learning (Cali-MR) method to calibrate the linear combinations of multiple imputation models and propensity models using bi-level optimization, which aims to learn an ensemble model that simultaneously possesses strong prediction performance and calibration ability. In this bi-level optimization, we adopt differentiable expected calibration errors to quantify the calibration ability that allows it to be directly

optimized. The calibrated linear combination of propensity and imputation models is then used to train the prediction model based on a joint learning algorithm. The contributions of this paper are summarized as follows.

- We propose a novel MR calibration method using bi-level optimization via calibrating the ensemble imputation and propensity models and address the non-differentiable issue by adopting differentiable expected calibration errors.
- We further propose a bi-level calibrated multiple robust learning algorithm to update the calibrated imputation models and the prediction model. To the best of our knowledge, this is the first work to perform calibration for MR estimator.
- We conduct extensive experiments on three real-world datasets, showing the effectiveness of our method compared to the state-of-the-art debiasing methods.

Related Works

Debiased Recommendation

Selection bias is common in the recommender systems (RS) (Chen et al. 2022; Wu et al. 2022; Wang et al. 2023c), resulting in the distribution shift between the observed population and the target population. There are many methods proposed to address this issue (Wang et al. 2022b; Zou et al. 2023; Wang et al. 2023a, 2024). Specifically, the error imputed based (EIB) method is proposed to mitigate this issue (Steck 2010). However, these types of methods require an accurate imputation, which is hard in practice. The IPS method uses inverse propensity score to weight the observed sample but may suffer from large variance with small propensities (Schnabel et al. 2016; Wang et al. 2022a). DR methods combine the advantages of both the EIB and IPS methods, guaranteeing unbiasedness if either the error imputation model or propensity model is correctly specified. There have been quantities of variants of DR methods to improve the debiasing performance, such as SDR (Li, Zheng, and Wu 2023), TDR (Li et al. 2023b), CDR (Song et al. 2023), N-DR (Li et al. 2024b), DT-DR (Zhang et al. 2024), UIDR (Li et al. 2024c), and OME-DR (Li et al. 2024d). Besides, Multiple robust (MR) (Li et al. 2023a) combines multiple imputation models and propensity models, and is unbiased when there exists a linear combination of them is correct. In addition, Liu et al. (2023) use an information bottleneck-based method and Yang et al. (2021) and Wang et al. (2023b) use adversarial learning for debiasing. However, these methods fail to consider model calibration properties. To mitigate this issue, DCE-DR (Kweon and Yu 2024) is proposed to calibrate the propensity and imputation model in the DR method. However, calibrating each imputation model and propensity model in MR is expensive and unreasonable, due to the unbiasedness condition of MR based on linear combinations is not taken into account. In this paper, we propose the Cali-MR method to calibrate the linear combination of multiple propensity and imputation models using the bi-level optimization method to enhance the debiasing performance and reliability of MR.

Calibration

Calibration means that the probability associated with the predicted class label should reflect its ground truth correctness likelihood (Guo et al. 2017; Kull, Silva Filho, and Flach 2017), which plays an important role in building reliable, robust AI systems, especially in safety-critical fields such as medical diagnosis (Caruana et al. 2015; Huang et al. 2020) and self-driving (Bojarski 2016). Calibration methods can be divided into the following four categories (Wang 2023): post-hoc calibration, regularization methods, uncertainty estimation, and hybrid calibration methods. Post-hoc calibration methods aim to calibrate after model training, including non-parametric calibration (Zadrozny and Elkan 2001) and parametric methods such as Platt scaling (Platt et al. 1999). Regularization methods adopt penalty terms such as the entropy regularization (Pereyra et al. 2017) and calibration errors (Kumar, Sarawagi, and Jain 2018) to ensure the calibration property. Uncertainty Estimation aims to alleviate model miscalibration by injecting randomness using Bayesian neural networks (Blundell et al. 2015), model and Gumbel-softmax (Jang, Gu, and Poole 2017) based approaches. Hybrid calibration methods combine two or more methods to achieve calibration. For example, Zhang, Kailkhura, and Han (2020) combine ensemble and temperature scaling and Laves et al. (2019) adopts monte-carlo dropout with temperature scaling. In this paper, we adopt a differentiable expected calibration error as part of the objective to ensure the model calibration.

Preliminary

Debiased Recommendation

Let $\mathcal{U} = \{u_1, \cdots, u_m\}$ be the users set, $\mathcal{I} = \{i_1, \cdots, i_n\}$ be the item set, and $\mathcal{D} = \mathcal{U} \times \mathcal{I}$ be the set of all user-item pairs. The rating matrix is denoted as $\mathbf{R} \in \mathbb{R}^{m \times n}$ with $r_{u,i}$ as element. Let $o_{u,i} \in \{0, 1\}$ be the observation indicator, where $o_{u,i} = 1$ indicates the rating $r_{u,i}$ is observed, otherwise is not. Define $x_{u,i}$ be the observed features. We denote the prediction model as $f_\theta(\cdot)$ parameterized by θ and the predicted ratings as $\hat{r}_{u,i} = f_\theta(x_{u,i})$. The goal is to accurately predict $r_{u,i}$ for all user-item pairs, which can be achieved by minimizing the ideal loss

$$\mathcal{L}_{\text{ideal}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} \mathcal{L}(f_\theta(x_{u,i}), r_{u,i}) := \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} e_{u,i},$$

where $\mathcal{L}(\cdot, \cdot)$ is the training loss function such as cross-entropy loss. However, in practice, we cannot obtain the complete rating matrix. We denote the set of user-item pairs with observed ratings as $\mathcal{O} = \{(u, i) \mid o_{u,i} = 1\}$. Thus, the naive method optimizes the average loss over the observed user-item pairs

$$\mathcal{L}_{\text{naive}}(\theta) = \frac{1}{|\mathcal{O}|} \sum_{(u,i) \in \mathcal{O}} e_{u,i}.$$

Due to the selection bias, $\mathbb{E}[\mathcal{L}_{\text{naive}}(\theta)] \neq \mathcal{L}_{\text{ideal}}(\theta)$. Several methods were proposed to unbiasedly estimate the ideal loss, including the EIB, IPS, DR, and their variants. Because

EIB and IPS can be regarded as a special case of DR, we only introduce the DR methods here. The loss function of the vanilla DR method is formulated as

$$\mathcal{L}_{\mathrm{DR}}(\theta)=\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\left[\hat{e}_{u,i}+\frac{o_{u,i}(e_{u,i}-\hat{e}_{u,i})}{\hat{p}_{u,i}}\right],$$

where $\hat{p}_{u,i}\triangleq\pi(x;\hat{\alpha})$ is the estimation of propensity score $p_{u,i}=\Pr(o_{u,i}=1\mid x_{u,i})$, and $\hat{e}_{u,i}=\mathcal{L}(m(x_{u,i};\beta),\hat{r}_{u,i})$ is the imputed error, while the imputation model is denoted as $m(x_{u,i};\hat{\beta})$. In addition, the multiple robust (MR) considers J propensity models $\pi_1(x;\hat{\alpha}_1),\ldots,\pi_J(x;\hat{\alpha}_J)$ and K imputation models $m_1(x;\hat{\beta}_1),\ldots,m_K(x;\hat{\beta}_K)$. Let $\hat{p}^j_{u,i}\triangleq\pi_j(x_{u,i};\hat{\alpha}_j)$, $\hat{m}^k_{u,i}\triangleq m_k(x;\hat{\beta}_k)$, the loss function of MR is shown below:

$$\mathcal{L}_{\mathrm{MR}}(\theta)=\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\boldsymbol{u}^T\left(x_{u,i}\right)\cdot\hat{\boldsymbol{\eta}}(\theta),$$

where $\boldsymbol{u}\left(x_{u,i}\right)=\left(1/\hat{p}^1_{u,i},\cdots,1/\hat{p}^J_{u,i},\hat{m}^1_{u,i},\cdots,\hat{m}^K_{u,i}\right)^T$ and $\hat{\boldsymbol{\eta}}(\theta)$ is the solution by minimizing

$$\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}o_{u,i}\left\{e_{u,i}-\boldsymbol{u}^T\left(x_{u,i}\right)\cdot\boldsymbol{\eta}\right\}^2.$$

The MR estimator is unbiased when there exists a weight $W=(w_1,w_2,\ldots,w_J,0,0,\ldots,0)$ satisfying $W\boldsymbol{u}\left(x_{u,i}\right)=1/p_{u,i}$ or $V=(0,\ldots,0,v_1,v_2,\ldots,v_K)$ satisfying $V\boldsymbol{u}\left(x_{u,i}\right)=e_{u,i}$ for all user-item pairs.

Calibration

A model is calibrated if its output reflects the ground-truth likelihood of correctness (Kull, Silva Filho, and Flach 2017). For the propensity model $\pi(x;\hat{\alpha})$ and the observation indicator o, a formal definition is shown below:

$$\mathbb{E}[o\mid\pi(x;\hat{\alpha})=\hat{p}]=\hat{p}\quad\forall\hat{p}\in[0,1].$$

For instance, if we have 100 samples with estimated propensity scores equal to 0.2, there should be exactly 20 samples being observed. Similarly, the formal definition for the calibrated imputation model $m(x;\hat{\beta})$ is formulated below:

$$\mathbb{E}[e\mid m(x;\hat{\beta})=\hat{e}]=\hat{e}\quad\forall\hat{e}\in\mathbb{R}.$$

To measure the miscalibration of the model, the Expected Calibration Error (ECE) metric is proposed (Naeini, Cooper, and Hauskrecht 2015). For a propensity model $\pi(x;\hat{\alpha})$ and imputation model $m(x;\hat{\beta})$, the ECE is defined as follows:

$$\mathrm{ECE}(\hat{\alpha})=\mathbb{E}_{\hat{p}}[|\mathbb{E}[o\mid\pi(x;\hat{\alpha})=\hat{p}]-\hat{p}|],$$
$$\mathrm{ECE}(\hat{\beta})=\mathbb{E}_{\hat{e}}[|\mathbb{E}[e\mid m(x;\hat{\beta})=\hat{e}]-\hat{e}|].$$

Methodology

Distinctions from Previous Work

Previous studies have proposed to calibrate the single propensity model and imputation model in DR estimator (Kweon and Yu 2024), and they propose to use a binning strategy to estimate the ECE metric empirically, for example for the propensity model:

$$\widehat{\mathrm{ECE}}\left(\hat{p}\right)=\sum_{m=1}^{M}\frac{|B_m|}{N}\left|\frac{\sum_{(u,i)\in B_m}o_{u,i}}{|B_m|}-\frac{\sum_{(u,i)\in B_m}\hat{p}_{u,i}}{|B_m|}\right|,$$

where B_m is the predefined m-th bin and N is the corresponding number of samples in the bin.

However, how to properly calibrate multiple propensity and imputation models for MR estimators remains unexplored. A naive approach is to calibrate each propensity and imputation model in the MR estimator individually. However, this method is computationally expensive and overlooks the robust property of MR estimator, that is, the MR estimator achieves unbiased if a linear combination of multiple candidate models is accurate.

Inspired by this, we propose to calibrate the linear combination of multiple models instead of calibrating each model individually. In addition, note that the previously used $\widehat{\mathrm{ECE}}$ involves assigning each sample to a specific hard bin, making it non-differentiable and thus unsuitable for direct incorporation into the training objective. To address this issue, we employ a soft binning strategy to develop the differentiable expected calibration error metric and leverage it to construct a regularization term that constrains the model's calibration error, which can be used for model training. Next, we will introduce the proposed Cali-MR in detail.

Differentiable Expected Calibration Error

To address the non-differentiable problem of $\widehat{\mathrm{ECE}}$, we leverage the soft binning strategy (Bohdal, Yang, and Hospedales 2023), using the following differentiable expected calibration error (DECE) that allows directly optimize calibration quality to mitigate the model miscalibration. For example, the DECE for a propensity model $\pi(x;\hat{\alpha})$ is defined as:

$$\mathrm{DECE}\left(\hat{\alpha}\right)=\frac{1}{|\mathcal{D}|}\sum_{m=1}^{M}\left|\sum_{(u,i)\in\mathcal{D}}o_m(x_{u,i};\phi)(o_{u,i}-\hat{p}_{u,i})\right|,$$

where $o_m(x_{u,i};\phi)=\mathbb{P}(x_{u,i}\in B_m|\hat{p}_{u,i})$ denotes the probability that how likely it is that $\hat{p}_{u,i}$ belongs to the m-th bin. In practice, the $o_m(x_{u,i};\phi)$ can be logistic regression or any other model (Bohdal, Yang, and Hospedales 2023).

In our Cali-MR, we adopt DECE as a regularization term for model training. Specifically, to calibrate the linear combination of multiple models, we formalize the propensity DECE loss $\mathcal{L}^p_{\mathrm{DECE}}$ as follows:

$$\mathcal{L}^p_{\mathrm{DECE}}(\mathbf{w};\phi_p;\alpha_1,\ldots,\alpha_J)= \tag{1}$$
$$\frac{1}{|\mathcal{D}|}\sum_{m=1}^{M}\left|\sum_{(u,i)\in\mathcal{D}}o_m(x_{u,i};\phi_p)(o_{u,i}-\sum_{j=1}^{J}w_j\hat{p}^j_{u,i})\right|,$$

where $\mathbf{w}=(w_1,\ldots,w_J)$ is a given set of weight coefficients. This loss measures the calibration error of the combination model, and minimizing this loss improves the calibration ability of the current propensity model combination under the current combination coefficients. We use a one-layer

neural network with softmax activation function to model the propensity soft binning model $o_m(x_{u,i};\phi_p)$ with parameters ϕ_p, where m is a pre-defined hyper-parameter.

Similarly, the imputation DECE loss $\mathcal{L}^e_{\mathrm{DECE}}$ under weight coefficients $\mathbf{v}=(v_1,\ldots,v_K)$ is formalized below:

$$\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v};\phi_m;\beta_1,\ldots,\beta_K) = \frac{1}{|\mathcal{D}|}\sum_{m=1}^{M}\left|\sum_{(u,i)\in\mathcal{D}} o_m(x_{u,i};\phi_m)\left(\frac{o_{u,i}e_{u,i}}{\sum_{j=1}^{J} w_j\hat{p}^j_{u,i}} - \sum_{k=1}^{K} v_k\hat{e}^k_{u,i}\right)\right|, \tag{2}$$

where we similarly model the imputation soft binning model $o_m(x_{u,i};\phi_m)$ with parameters ϕ_m with a one-layer neural network with softmax activation function. Note that the $e_{u,i}$ is missing for user-item pairs with $o_{u,i}=0$, we reweight the observed $e_{u,i}$ using the inverse of the linear combination of the multiple propensity models.

Based on the DECE loss $\mathcal{L}^p_{\mathrm{DECE}}$ or $\mathcal{L}^e_{\mathrm{DECE}}$, we can measure the calibration quality of given multiple models and weight coefficients, and improve the calibration ability of the current combined model by minimizing the loss.

Calibrated Multiple Robust Learning

Note that in calculating the DECE loss $\mathcal{L}^p_{\mathrm{DECE}}$ or $\mathcal{L}^e_{\mathrm{DECE}}$, the coefficients $\mathbf{w}$ or $\mathbf{v}$ need to be explicitly specified. However, these coefficients are unknown during the training process of the existing MR method. Therefore, we propose using bi-level optimization to solve for the optimal coefficients, parameters of the soft binning model and multiple imputation and propensity models. In addition, we alternatively update the prediction model and the calibrated imputation model based on a joint learning algorithm.

Multiple Propensity Calibration For propensity models, the optimization objective can be formalized as follows:

$$\begin{aligned}&(\alpha_1^*,\ldots,\alpha_J^*) = \arg\min_{\alpha_1,\ldots,\alpha_J} \frac{1}{J}\sum_{j=1}^{J}\mathcal{L}_{p_j}(\alpha_j) + \lambda\mathcal{L}^p_{\mathrm{DECE}}(\mathbf{w}^*)\\ &\text{s.t.}\quad \mathbf{w}^*(\alpha_1,\ldots,\alpha_J) = \arg\min_{\mathbf{w}^*}\mathcal{L}_p(\mathbf{w}^*(\alpha_1,\ldots,\alpha_J)),\end{aligned} \tag{3}$$

where $\mathcal{L}^p_{\mathrm{DECE}}(\mathbf{w})$ loss is the calibration constraints defined in Equation 1, and $\mathcal{L}_{p_j}(\alpha_j)$ is the training loss for a single propensity model $\pi_J(x;\hat{\alpha}_J)$ ensuring the accuracy of each independent propensity model, which is shown below:

$$\mathcal{L}_{p_j}(\alpha_j) = \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}[-o_{u,i}\cdot\log p^j_{u,i} - (1-o_{u,i})\cdot\log(1-p^j_{u,i})].$$

$\mathcal{L}_p(\mathbf{w}(\alpha_1,\ldots,\alpha_J))$ is the loss for the combination coefficients $\mathbf{w}$, aiming to learn a set of coefficients such that the linear combination $\sum_{j=1}^{J} w_j\hat{p}^j_{u,i}$ can accurately predict observation indicator $o_{u,i}$, which is shown below:

$$\mathcal{L}_p(\mathbf{w}(\alpha_1,\ldots,\alpha_J)) = \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}[-o_{u,i}\cdot\log(\sum_{j=1}^{J} w_j\hat{p}^j_{u,i}) - (1-o_{u,i})\cdot\log(1-\sum_{j=1}^{J} w_j\hat{p}^j_{u,i})].$$

In this bi-level optimization, we aim to train the propensity models such that each model performs well and their linear combination is well calibrated, where the coefficients also ensure the strong prediction performance of the combined model. For practical implementation, we first assumed update coefficients $\mathbf{w}$ through optimizing $\mathcal{L}_p(\mathbf{w})$. Using these coefficients, we calculate the DECE loss $\mathcal{L}^p_{\mathrm{DECE}}(\mathbf{w})$ and combine it with the base prediction loss of each propensity model denoted as $\mathcal{L}_{p_j}(\alpha_j)$ to form the final loss. This final loss is then used to update the propensity models $\alpha_1,\ldots,\alpha_J$. After that, the loss $\mathcal{L}_p(\mathbf{w})$ and $\mathcal{L}^p_{\mathrm{DECE}}(\mathbf{w})$ are used to update the combination coefficients $\mathbf{w}$ and the soft binning model ϕ_p sequentially.

Multiple Imputation Calibration With the calibrated propensity model $\sum_{j=1}^{J} w_j\hat{p}^j_{u,i}$ obtained from Equation 3 by the bi-level optimization, we can further calibrate the linear

Algorithm 1: Bi-level Calibrated Multiple Robust Learning

Input: observed ratings $\mathbf{R}^o$, calibrated propensity model $\sum_{j=1}^{J} w_j\hat{p}^j_{u,i}$, and stabilization parameter λ

while *stopping criteria is not satisfied* **do**
  for *number of steps for training imputation model* **do**
    Sample a batch of user-item pairs $\{(u_l,i_l)\}_{l=1}^{L}$ from observed population $\mathcal{O}$
    Assumed update coefficients $\mathbf{v}'(\beta_1,\ldots,\beta_K) \leftarrow \mathbf{v} - \alpha_{\mathbf{v}}\nabla_{\mathbf{v}}\mathcal{L}_e\left(\mathbf{v}\mid\beta_1,\ldots,\beta_K\right)$
    Update imputation models $(\beta_1,\ldots,\beta_K) \leftarrow \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{e_k}(\beta_k) + \lambda\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v}'(\beta_1,\ldots,\beta_K))$
    Update coefficients $\mathbf{v} \leftarrow \mathcal{L}_e(\mathbf{v}(\beta_1,\ldots,\beta_K))$
    Update imputation soft binning model $\phi_m \leftarrow \mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v};\phi_m;\beta_1,\ldots,\beta_K)$
  end
  for *number of steps for training the prediction model* **do**
    Sample a batch of user-item pairs $\mathcal{D}'$ from $\mathcal{D}$
    Obtain the rated samples in $\mathcal{D}'$ as $\{(u_m,i_m)\}_{m=1}^{M} = \mathcal{O}' \subseteq \mathcal{O}$
    Update $\eta \leftarrow [\sum_{(u,i)\in\mathcal{O}'}\boldsymbol{u}(x_{u,i};\alpha,\beta)\cdot\boldsymbol{u}^T(x_{u,i};\alpha,\beta) + \lambda I]^{-1}[\sum_{(u,i)\in\mathcal{O}'}\boldsymbol{u}(x_{u,i};\alpha,\beta)\cdot e_{u,i}]$
    Sample a batch of user-item pairs $\{(u_n,i_n)\}_{n=1}^{N}$ from $\mathcal{D}\setminus\mathcal{D}'$
    Update prediction model $\theta \leftarrow \theta - \nabla_\theta\mathcal{L}_{\mathrm{MR}}(\theta;\alpha,\beta)$
  end
end

Table 1: Performance on AUC, NDCG@K, and F1@K on **Coat**, **Yahoo! R3** and **KuaiRec**. The best and the second best results are bolded and underlined, where * means statistically significant results (p-value ≤ 0.05) using the paired-t-test.

	Coat			**Yahoo! R3**			**KuaiRec**		
Method	AUC	NDCG@5	F1@5	AUC	NDCG@5	F1@5	AUC	NDCG@20	F1@20
Naive	$0.703_{\pm 0.006}$	$0.605_{\pm 0.012}$	$0.467_{\pm 0.007}$	$0.673_{\pm 0.001}$	$0.635_{\pm 0.002}$	$0.306_{\pm 0.002}$	$0.753_{\pm 0.001}$	$0.449_{\pm 0.002}$	$0.124_{\pm 0.002}$
IPS	$0.717_{\pm 0.007}$	$0.617_{\pm 0.009}$	$0.473_{\pm 0.008}$	$0.678_{\pm 0.001}$	$0.638_{\pm 0.002}$	$0.318_{\pm 0.002}$	$0.755_{\pm 0.004}$	$0.452_{\pm 0.010}$	$0.131_{\pm 0.004}$
SNIPS	$0.714_{\pm 0.012}$	$0.614_{\pm 0.012}$	$0.474_{\pm 0.009}$	$0.683_{\pm 0.002}$	$0.639_{\pm 0.002}$	$0.316_{\pm 0.002}$	$0.754_{\pm 0.003}$	$0.453_{\pm 0.004}$	$0.126_{\pm 0.003}$
ASIPS	$0.719_{\pm 0.009}$	$0.618_{\pm 0.012}$	$0.476_{\pm 0.009}$	$0.679_{\pm 0.003}$	$0.640_{\pm 0.003}$	$0.319_{\pm 0.003}$	$0.757_{\pm 0.005}$	$0.474_{\pm 0.007}$	$0.130_{\pm 0.005}$
IPS-V2	$0.726_{\pm 0.005}$	$0.627_{\pm 0.009}$	$0.479_{\pm 0.008}$	$0.685_{\pm 0.002}$	$0.646_{\pm 0.003}$	$0.320_{\pm 0.002}$	$0.764_{\pm 0.001}$	$0.476_{\pm 0.003}$	$0.135_{\pm 0.003}$
KBIPS	$0.714_{\pm 0.003}$	$0.618_{\pm 0.010}$	$0.474_{\pm 0.007}$	$0.676_{\pm 0.002}$	$0.642_{\pm 0.003}$	$0.318_{\pm 0.002}$	$0.763_{\pm 0.001}$	$0.463_{\pm 0.007}$	$0.134_{\pm 0.002}$
AKBIPS	$0.732_{\pm 0.004}$	$0.636_{\pm 0.006}$	$0.483_{\pm 0.006}$	$0.689_{\pm 0.001}$	$0.658_{\pm 0.002}$	$0.324_{\pm 0.002}$	$0.766_{\pm 0.003}$	$0.478_{\pm 0.009}$	$0.138_{\pm 0.003}$
DR	$0.718_{\pm 0.008}$	$0.623_{\pm 0.009}$	$0.474_{\pm 0.007}$	$0.684_{\pm 0.002}$	$0.658_{\pm 0.003}$	$0.326_{\pm 0.002}$	$0.755_{\pm 0.008}$	$0.462_{\pm 0.010}$	$0.135_{\pm 0.005}$
DR-JL	$0.723_{\pm 0.005}$	$0.629_{\pm 0.007}$	$0.479_{\pm 0.005}$	$0.685_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.324_{\pm 0.002}$	$0.766_{\pm 0.002}$	$0.467_{\pm 0.005}$	$0.136_{\pm 0.003}$
MRDR-JL	$0.727_{\pm 0.005}$	$0.627_{\pm 0.008}$	$0.480_{\pm 0.008}$	$0.684_{\pm 0.002}$	$0.652_{\pm 0.003}$	$0.325_{\pm 0.002}$	$0.768_{\pm 0.005}$	$0.473_{\pm 0.007}$	$0.139_{\pm 0.004}$
DR-BIAS	$0.726_{\pm 0.004}$	$0.629_{\pm 0.009}$	$0.482_{\pm 0.007}$	$0.685_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.325_{\pm 0.003}$	$0.768_{\pm 0.003}$	$0.477_{\pm 0.006}$	$0.137_{\pm 0.004}$
DR-MSE	$0.727_{\pm 0.007}$	$0.631_{\pm 0.008}$	$0.484_{\pm 0.007}$	$0.687_{\pm 0.002}$	$0.657_{\pm 0.003}$	$0.327_{\pm 0.003}$	$0.770_{\pm 0.003}$	$0.480_{\pm 0.006}$	$0.140_{\pm 0.003}$
MR	$0.724_{\pm 0.004}$	$0.636_{\pm 0.006}$	$0.481_{\pm 0.006}$	$0.691_{\pm 0.002}$	$0.647_{\pm 0.002}$	$0.316_{\pm 0.003}$	$0.776_{\pm 0.005}$	$0.483_{\pm 0.006}$	$0.142_{\pm 0.003}$
TDR	$0.714_{\pm 0.006}$	$0.634_{\pm 0.011}$	$0.483_{\pm 0.008}$	$0.688_{\pm 0.003}$	$0.662_{\pm 0.002}$	$0.329_{\pm 0.002}$	$0.772_{\pm 0.003}$	$0.486_{\pm 0.005}$	$0.140_{\pm 0.003}$
TDR-JL	$0.731_{\pm 0.005}$	$0.639_{\pm 0.007}$	$0.484_{\pm 0.007}$	$0.689_{\pm 0.002}$	$0.656_{\pm 0.004}$	$0.327_{\pm 0.003}$	$0.772_{\pm 0.003}$	$0.489_{\pm 0.005}$	$0.142_{\pm 0.003}$
StableDR	$0.735_{\pm 0.005}$	$0.640_{\pm 0.007}$	$0.484_{\pm 0.006}$	$0.688_{\pm 0.002}$	$0.661_{\pm 0.003}$	$0.329_{\pm 0.002}$	$0.773_{\pm 0.001}$	$0.491_{\pm 0.003}$	$0.143_{\pm 0.003}$
DR-V2	$0.734_{\pm 0.007}$	$0.639_{\pm 0.009}$	$0.487_{\pm 0.006}$	$0.690_{\pm 0.002}$	$0.660_{\pm 0.005}$	$0.328_{\pm 0.002}$	$0.773_{\pm 0.003}$	$0.488_{\pm 0.006}$	$0.142_{\pm 0.004}$
KBDR	$0.730_{\pm 0.003}$	$0.631_{\pm 0.005}$	$0.482_{\pm 0.006}$	$0.682_{\pm 0.002}$	$0.648_{\pm 0.003}$	$0.323_{\pm 0.002}$	$0.765_{\pm 0.004}$	$0.460_{\pm 0.006}$	$0.138_{\pm 0.003}$
AKBDR	$\mathbf{0.745}_{\pm 0.004}$	$0.645_{\pm 0.008}$	$\underline{0.493}_{\pm 0.007}$	$0.692_{\pm 0.002}$	$0.661_{\pm 0.002}$	$0.328_{\pm 0.002}$	$0.782_{\pm 0.003}$	$0.498_{\pm 0.008}$	$0.147_{\pm 0.003}$
DCE-DR	$0.736_{\pm 0.006}$	$\underline{0.648}_{\pm 0.007}$	$0.489_{\pm 0.005}$	$\underline{0.698}_{\pm 0.002}$	$\underline{0.670}_{\pm 0.002}$	$\underline{0.333}_{\pm 0.003}$	$\underline{0.795}_{\pm 0.004}$	$\underline{0.512}_{\pm 0.005}$	$\underline{0.153}_{\pm 0.002}$
Cali-MR	$\underline{0.741}_{\pm 0.002}$	$\mathbf{0.658}^{*}_{\pm 0.004}$	$\mathbf{0.495}_{\pm 0.004}$	$\mathbf{0.703}^{*}_{\pm 0.002}$	$\mathbf{0.678}^{*}_{\pm 0.002}$	$\mathbf{0.338}^{*}_{\pm 0.004}$	$\mathbf{0.798}^{*}_{\pm 0.003}$	$\mathbf{0.521}^{*}_{\pm 0.005}$	$\mathbf{0.158}^{*}_{\pm 0.002}$

combination of multiple imputation models:

$$(\beta_1^*, \ldots, \beta_K^*) = \arg\min_{\beta_1,\ldots,\beta_K} \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{e_k}(\beta_k) + \lambda \mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v}^*)$$

$$\text{s.t.} \quad \mathbf{v}^*(\beta_1, \ldots, \beta_K) = \arg\min_{\mathbf{v}^*} \mathcal{L}_e(v^*(\beta_1, \ldots, \beta_K)), \quad (4)$$

where $\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v})$ is the calibration constraints shown in Equation 2. The naive training loss for each imputation model $m_K(x; \hat{\beta}_K)$ is expressed as

$$\mathcal{L}_{e_k}(\beta_k; \theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \hat{e}^k_{u,i})^2}{\sum_{j=1}^{J} w_j \hat{p}^j_{u,i}},$$

and the loss for the combination coefficients $\mathbf{v}$ is

$$\mathcal{L}_e(\mathbf{v}(\beta_1, \ldots, \beta_K)) = \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \sum_{k=1}^{K} v_k \hat{e}^k_{u,i})^2}{|\mathcal{D}| \sum_{j=1}^{J} w_j \hat{p}^j_{u,i}},$$

which aims to learn a set of weight coefficient such that the linear combination of the imputation models $\sum_{k=1}^{K} v_k \hat{e}^k_{u,i}$ can unbiasedly estimate the prediction error $e_{u,i}$.

Similar to updating the propensity models, we first use $\mathcal{L}_e(\mathbf{v})$ to assumed update coefficients $\mathbf{v}'$. Based on that, we compute the DECE loss $\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v}')$, combine it with the base training loss of each imputation model $\mathcal{L}_{e_k}(\beta_k; \theta)$, and use the combined loss to update both the imputation model $\beta_1, \ldots, \beta_K$. Then we adopt loss $\mathcal{L}_e(\mathbf{v})$ and $\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v})$ to update the combination coefficients $\mathbf{v}$ and the soft binning model ϕ_e sequentially. After obtaining the updated imputation models and coefficients, we jointly train the prediction model using the standard multiple robust learning algorithm. Specifically, we use ridge regression to calculate $\hat{\eta}$ in the MR estimator, and update the prediction model based on the MR loss $\mathcal{L}_{\mathrm{MR}}$ using different samples. We summarize the above learning algorithm in Algorithm 1.

Experiments

Datasets

To evaluate the debiasing performance, we conduct experiments on three benchmark datasets **Coat**[1] and **Yahoo! R3**[2], and **KuaiRec**[3] (Gao et al. 2022), which are widely used in debiased RS with both biased and unbiased data. **Coat** dataset consists of 6,960 biased training samples and 4,640 unbiased test samples derived from 290 users rating on 300 items. Each user self-selects 24 items to rate to consist of the training set and randomly rates 16 items to consist of the test set. The **Yahoo! R3** dataset includes 311,704 biased training samples and 54,000 unbiased test samples derived from 15,400 users rating on 1,000 items. Both datasets are five-scale, and following previous works (Chen et al. 2021; Li et al. 2024a, 2023c), we binarize the ratings greater than three to 1, and others to 0. The **KuaiRec** dataset is collected from a video-sharing platform and contains 4,676,570 video watching ratios derived from 1,411 users evaluating 3,327 videos. Following previous studies (Li et al. 2023d, 2024d; Kweon and Yu 2024), we binarize the continuous ratios greater than two to 1, otherwise to 0.

Baselines

We compare our method with the following baselines for comprehensive evaluations: **Naive** method (Koren, Bell, and Volinsky 2009), IPS-based methods including **IPS** (Schnabel et al. 2016), **SNIPS** (Swaminathan and Joachims 2015), **ASIPS** (Saito 2020), **IPS-V2** (Li et al. 2023d), **KBIPS** (Li et al. 2024d) and **AKBIPS** (Li et al. 2024d), and DR-based

[1] https://www.cs.cornell.edu/~schnabts/mnar/

[2] https://webscope.sandbox.yahoo.com

[3] https://github.com/chongminggao/KuaiRec

methods including **DR** (Saito 2020), **DR-JL** (Wang et al. 2019), **MRDR** (Guo et al. 2021), **DR-BIAS** (Dai et al. 2022), **DR-MSE** (Dai et al. 2022), **MR** (Li et al. 2023a), **TDR** (Li et al. 2023b), **TDR-JL** (Li et al. 2023b), **StableDR** (Li, Zheng, and Wu 2023), **DR-V2** (Li et al. 2023d), **KBDR** (Li et al. 2024d), **AKBDR** (Li et al. 2024d) and **DCE-DR** (Kweon and Yu 2024).

Experiment Protocols and Details

We evaluate the prediction performance with three widely adopted evaluation metrics: AUC, NDCG@K (N@K), and F1@K, and we set K = 5 on **Coat** and **Yahoo! R3** datasets, and K = 20 on **KuaiRec** dataset. In addition, we tune learning rate in $\{0.01, 0.05\}$ and weight decay in $\{1e-6, 5e-6, 1e-5, \ldots, 1e-3, 5e-3\}$. We implement our method on PyTorch with Adam as the optimizer and adopt NVIDIA A40 as computing resource. We use the same hyperparameter search space and follow the results in Li et al. (2024d).

Performance Analysis

The experimental results are shown in Table 1, and we find that all the debiasing methods including both IPS based and DR based baselines outperform the Naive method, which demonstrates the importance of debiasing. Besides, among all baselines, DCE-DR achieves the most competitive performance, indicating calibrating a single propensity model and imputation model in Doubly Robust estimator improves prediction performance. Furthermore, the Cali-MR method exhibits superior overall performance in all three datasets. In particular, on **Yahoo! R3** and **KuaiRec** datasets, we conduct statical significance tests using the paired t-test, validating that our method significantly outperforms the existing best baseline in all evaluation metrics. This shows that calibrating multiple propensity and imputation models in Multiple Robust estimator with weight coefficient learned in bi-level optimization further enhances the debiasing performance.

Conclusion

In this paper, we explore how to properly calibrate multiple propensity and imputation models in Multiple Robust (MR) estimator. First, we argue that calibrating each candidate model individually is too costly and unreasonable due to the unbiasedness condition of MR in terms of linear combinations is not considered. Based on this, we propose using bi-level optimization to calibrate the linear combination of multiple candidate models. Specifically, in the bi-level optimization, we first assumed to update the combination coefficients to obtain the best-performing combination coefficients under the current candidate model parameters. Then, based on these coefficients, we update the candidate model parameters to ensure that each model maintains good prediction performance, while also ensuring that the combination model achieves strong calibration ability, where a differentiable expected calibration error metric that allows it to be directly optimized is adopted. Experimental results on three real-world datasets demonstrate that the proposed multiple propensity and imputation calibration method further enhances the prediction performance.

References

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *ICML*.

Bohdal, O.; Yang, Y.; and Hospedales, T. 2023. Meta-Calibration: Learning of Model Calibration Using Differentiable Expected Calibration Error. *TMLR*.

Bojarski, M. 2016. End to end learning for self-driving cars. *arXiv:1604.07316*.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*.

Chang, Y.-W.; Hsieh, C.-J.; Chang, K.-W.; Ringgaard, M.; and Lin, C.-J. 2010. Training and testing low-degree polynomial data mappings via linear SVM. *JMLR*.

Chen, J.; Dong, H.; Qiu, Y.; He, X.; Xin, X.; Chen, L.; Lin, G.; and Yang, K. 2021. AutoDebias: Learning to Debias for Recommendation. In *SIGIR*.

Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; and He, X. 2022. Bias and Debias in Recommender System: A Survey and Future Directions. *TOIS*.

Dai, Q.; Li, H.; Wu, P.; Dong, Z.; Zhou, X.-H.; Zhang, R.; Zhang, R.; and Sun, J. 2022. A generalized doubly robust learning framework for debiasing post-click conversion rate prediction. In *KDD*.

Gao, C.; Li, S.; Lei, W.; Chen, J.; Li, B.; Jiang, P.; He, X.; Mao, J.; and Chua, T.-S. 2022. KuaiRec: A Fully-observed Dataset and Insights for Evaluating Recommender Systems. In *CIKM*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *ICML*.

Guo, S.; Zou, L.; Liu, Y.; Ye, W.; Cheng, S.; Wang, S.; Chen, H.; Yin, D.; and Chang, Y. 2021. Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation. In *SIGIR*.

Huang, Y.; Li, W.; Macheret, F.; Gabriel, R. A.; and Ohno-Machado, L. 2020. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inf. Assoc*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparametrization with Gumble-Softmax. In *ICLR*.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*.

Kull, M.; Silva Filho, T.; and Flach, P. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*.

Kumar, A.; Sarawagi, S.; and Jain, U. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *ICML*.

Kweon, W.; and Yu, H. 2024. Doubly Calibrated Estimator for Recommendation on Data Missing Not At Random. In *WWW*.

Laves, M.-H.; Ihler, S.; Kortmann, K.-P.; and Ortmaier, T. 2019. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv:1909.13550*.

Li, H.; Dai, Q.; Li, Y.; Lyu, Y.; Dong, Z.; Zhou, X.-H.; and Wu, P. 2023a. Multiple Robust Learning for Recommendation. In *AAAI*.

Li, H.; Lyu, Y.; Zheng, C.; and Wu, P. 2023b. TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations. In *ICLR*.

Li, H.; Wu, K.; Zheng, C.; Xiao, Y.; Wang, H.; Geng, Z.; Feng, F.; He, X.; and Wu, P. 2024a. Removing Hidden Confounding in Recommendation: A Unified Multi-Task Learning Approach. In *NeurIPS*.

Li, H.; Xiao, Y.; Zheng, C.; and Wu, P. 2023c. Balancing Unobserved Confounding with a Few Unbiased Ratings in Debiased Recommendations. In *WWW*.

Li, H.; Xiao, Y.; Zheng, C.; Wu, P.; and Cui, P. 2023d. Propensity Matters: Measuring and Enhancing Balancing for Recommendation. In *ICML*.

Li, H.; Zheng, C.; Ding, S.; Feng, F.; He, X.; Geng, Z.; and Wu, P. 2024b. Be Aware of the Neighborhood Effect: Modeling Selection Bias under Interference for Recommendation. In *ICLR*.

Li, H.; Zheng, C.; Wang, S.; Wu, K.; Wang, E.; Wu, P.; Geng, Z.; Chen, X.; and Zhou, X.-H. 2024c. Relaxing the Accurate Imputation Assumption in Doubly Robust Learning for Debiased Collaborative Filtering. In *ICML*.

Li, H.; Zheng, C.; and Wu, P. 2023. StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random. In *ICLR*.

Li, H.; Zheng, C.; Xiao, Y.; Wu, P.; Geng, Z.; Chen, X.; and Cui, P. 2024d. Debiased collaborative filtering with kernel-based causal balancing. In *ICLR*.

Liu, D.; Cheng, P.; Zhu, H.; Dong, Z.; He, X.; Pan, W.; and Ming, Z. 2023. Debiased representation learning in recommendation via information bottleneck. *TORS*.

Morgan, S. L.; and Winship, C. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, second edition.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv:1701.06548*.

Platt, J.; et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*.

Ricci, F.; Rokach, L.; and Shapira, B. 2010. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer.

Saito, Y. 2020. Asymmetric Tri-training for Debiasing Missing-Not-At-Random Explicit Feedback. In *SIGIR*.

Saito, Y. 2020. Doubly robust estimator for ranking metrics with post-click conversions. In *RecSys*.

Saito, Y.; Yaginuma, S.; Nishino, Y.; Sakata, H.; and Nakata, K. 2020. Unbiased recommender learning from missing-not-at-random implicit feedback. In *WSDM*.

Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; and Joachims, T. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *ICML*.

Song, Z.; Chen, J.; Zhou, S.; Shi, Q.; Feng, Y.; Chen, C.; and Wang, C. 2023. CDR: Conservative Doubly Robust Learning for Debiased Recommendation. In *CIKM*.

Steck, H. 2010. Training and testing of recommender systems on data missing not at random. In *KDD*.

Swaminathan, A.; and Joachims, T. 2015. The Self-Normalized Estimator for Counterfactual Learning. In *NeurIPS*.

Wang, C. 2023. Calibration in deep learning: A survey of the state-of-the-art. *arXiv:2308.01222*.

Wang, H.; Chang, T.-W.; Liu, T.; Huang, J.; Chen, Z.; Yu, C.; Li, R.; and Chu, W. 2022a. Escm2: Entire space counterfactual multi-task model for post-click conversion rate estimation. In *SIGIR*.

Wang, H.; Kuang, K.; Chi, H.; Yang, L.; Geng, M.; Huang, W.; and Yang, W. 2023a. Treatment effect estimation with adjustment feature selection. In *KDD*.

Wang, H.; Kuang, K.; Lan, L.; Wang, Z.; Huang, W.; Wu, F.; and Yang, W. 2024. Out-of-distribution generalization with causal feature separation. *TKDE*.

Wang, H.; Yang, W.; Yang, L.; Wu, A.; Xu, L.; Ren, J.; Wu, F.; and Kuang, K. 2022b. Estimating Individualized Causal Effect with Confounded Instruments. In *KDD*.

Wang, J.; Li, H.; Zhang, C.; Liang, D.; Yu, E.; Ou, W.; and Wang, W. 2023b. Counterclr: Counterfactual contrastive learning with non-random missing data in recommendation. In *ICDM*.

Wang, W.; Zhang, Y.; Li, H.; Wu, P.; Feng, F.; and He, X. 2023c. Causal Recommendation: Progresses and Future Directions. In *SIGIR*.

Wang, X.; Zhang, R.; Sun, Y.; and Qi, J. 2019. Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random. In *ICML*.

Wu, P.; Li, H.; Deng, Y.; Hu, W.; Dai, Q.; Dong, Z.; Sun, J.; Zhang, R.; and Zhou, X.-H. 2022. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In *IJCAI*.

Yang, M.; Dai, Q.; Dong, Z.; Chen, X.; He, X.; and Wang, J. 2021. Top-n recommendation with counterfactual user preference simulation. In *CIKM*.

Zadrozny, B.; and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*.

Zhang, H.; Wang, S.; Li, H.; Zheng, C.; Chen, X.; Liu, L.; Luo, S.; and Wu, P. 2024. Uncovering the Propensity Identification Problem in Debiased Recommendations. In *ICDE*.

Zhang, J.; Kailkhura, B.; and Han, T. Y.-J. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *ICML*.

Zou, H.; Wang, H.; Xu, R.; Li, B.; Pei, J.; Jian, Y. J.; and Cui, P. 2023. Factual Observation Based Heterogeneity Learning for Counterfactual Prediction. In *CLeaR*.