Recommendations as Treatments: Debiasing Learning and Evaluation

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh
Cornell University, Ithaca, NY, USA
{TBS49, FA234, AS3354}@cornell.edu

Navin Chandak
Google, Mountain View, CA, USA
NAVINCHANDAK92@GMAIL.COM

Thorsten Joachims
Cornell University, Ithaca, NY, USA
TJ@CS.CORNELL.EDU

Abstract
Most data for evaluating and training recommender systems is subject to selection biases, either through self-selection by the users or through the actions of the recommendation system itself. In this paper, we provide a principled approach to handling selection biases, adapting models and estimation techniques from causal inference. The approach leads to unbiased performance estimators despite biased data, and to a matrix factorization method that provides substantially improved prediction performance on real-world data. We theoretically and empirically characterize the robustness of the approach, finding that it is highly practical and scalable.

1. Introduction
Virtually all data for training recommender systems is subject to selection biases. In a movie recommendation system, for example, users typically watch and rate those movies that they like, and rarely rate movies that they do not like (Pradel et al., 2012). Similarly, when an ad-placement system recommends ads, it shows ads that it believes to be of interest to the user, but will less frequently display other ads. Having observations be conditioned on the effect we would like to optimize (e.g. the star rating, the probability of a click) leads to data that is Missing Not At Random (MNAR) (Little & Rubin, 2002). This creates a widely-recognized challenge for training and evaluating recommender systems (Marlin & Zemel, 2009; De Myttenaere et al., 2014).

In this paper, we develop an approach to evaluating and training recommender systems that remedies selection biases in a principled, practical, and highly effective way. To this effect, we view the recommendation task from a causal inference perspective. In particular, we argue that exposing a user to an item in a recommendation system is an intervention analogous to exposing a patient to a treatment in a medical study. In both cases, the goal is to accurately estimate the effect of new interventions (e.g. a new treatment policy or a new set of recommendations) despite incomplete and biased data due to self-selection or experimenter-bias. By connecting recommendation to causal inference from experimental and observational data, the paper derives a principled framework for unbiased evaluation and learning of recommender systems under selection biases.

The main contribution of this paper is four-fold. First, we show how estimating the quality of a recommendation system can be approached with propensity-weighting techniques commonly used in causal inference (Imbens & Rubin, 2015), complete-cases analysis (Little & Rubin, 2002), and other problems (Cortes et al., 2008; Bickel et al., 2009; Sugiyama & Kawanabe, 2012). In particular, we derive unbiased estimators for a wide range of performance measures (e.g. MSE, MAE, DCG). Second, with these estimators in hand, we propose an Empirical Risk Minimization (ERM) framework for learning recommendation systems under selection bias, for which we derive generalization error bounds. Third, we use the ERM framework to derive a matrix factorization method that can account for selection bias while remaining conceptually simple and highly scalable. Fourth, we explore methods for estimating propensities in observational settings where selection bias is due to self-selection by the users, and we characterize the robustness of the framework against misspecified propensities.

Our conceptual and theoretical contributions are validated in an extensive empirical evaluation. For the task of evaluating recommender systems, we show that our performance estimators can be orders-of-magnitude more accurate than
standard estimators commonly used in the past (Bell et al., 2007). For the task of learning recommender systems, we show that our new matrix factorization method substantially outperforms methods that ignore selection bias, as well as existing state-of-the-art methods that perform joint-likelihood inference under MNAR data (Hernández-Lobato et al., 2014). This is especially promising given the conceptual simplicity and scalability of our approach compared to joint-likelihood inference. We provide software implementing of our method, as well as a new real-world dataset that can serve as a future benchmark.

2. Related Work

Past work that explicitly dealt with the MNAR nature of recommendation data approached the problem as missing-data imputation based on the joint likelihood of the missing data model and the rating model (Marlin et al., 2007; Marlin & Zemel, 2009; Hernández-Lobato et al., 2014). This has lead to sophisticated and highly complex methods. We take a fundamentally different approach that treats both models separately, making our approach modular and scalable. Furthermore, our approach is robust to misspecification of the rating model, and we characterize how the overall learning process degrades gracefully under a misspecified missing-data model. We empirically compare against the state-of-the-art joint likelihood model (Hernández-Lobato et al., 2014) in this paper.

Related but different from the problem we consider is recommendation from positive feedback alone (Hu et al., 2008; Dawen Liang & Blei, 2016). Alternative approaches to learning with MNAR data (Steck, 2010; Lim et al., 2015) aim to avoid the problem by considering performance measures less affected by selection bias.

Propensity-based approaches have been widely used in causal inference from observational studies (Imbens & Rubin, 2015), as well as in complete-case analysis for missing data (Little & Rubin, 2002; Seaman & White, 2013) and in survey sampling (Thompson, 2012). However, their use within recommendation or matrix completion is new to our knowledge. Weighting approaches are also widely used in domain adaptation and covariate shift, where data from one source is used to train for a different problem (e.g., Huang et al., 2006; Bickel et al., 2009; Sugiyama & Kawanabe, 2012). We will draw upon this work, especially the learning theory related to weighting approaches in (Cortes et al., 2008; 2010). Weighted estimators are also used in Monte Carlo estimation via importance sampling (Owen, 2013).

3. Unbiased Performance Estimation for Recommendation

Consider the following toy example adapted from Steck (2010) to illustrate the disastrous effect that selection bias can have on conventional evaluation using a test set of held-out ratings. Denote with $u \in \{1,...,U\}$ the users and with $i \in \{1,...,I\}$ the items. Figure 1 shows the matrix of true ratings $Y \in \mathbb{R}^{U \times I}$ for our toy example, where a subset of users are “horror lovers” who rate all horror movies 1 and all romance movies 1. Similarly, there is a subset of “romance lovers” who rate just the opposite way. However, both groups rate dramas as 3.

The binary matrix $O \in \{0,1\}^{U \times I}$ in Figure 1 shows for which movies the users provided their rating to the system, $[O_{u,i} = 1] \leftrightarrow [Y_{u,i} \text{ observed}]$. Our toy example shows a strong correlation between liking and rating a movie, and the matrix $P$ describes the marginal probabilities $P_{u,i} = P(O_{u,i} = 1)$ with which each rating is revealed. For this data, consider the following two evaluation tasks.

3.1. Task 1: Estimating Rating Prediction Accuracy

For the first task, we want to evaluate how well a predicted rating matrix $\hat{Y}$ reflects the true ratings in $Y$. Standard evaluation measures like Mean Absolute Error (MAE) or Mean Squared Error (MSE) can be written as:

$$R(\hat{Y}) = \sum_{u=1}^{U} \sum_{i=1}^{I} \delta_{u,i}(Y, \hat{Y}), \quad (1)$$

for an appropriately chosen $\delta_{u,i}(Y, \hat{Y})$.

$$\text{MAE: } \delta_{u,i}(Y, \hat{Y}) = |Y_{u,i} - \hat{Y}_{u,i}|/(U \cdot I), \quad (2)$$

$$\text{MSE: } \delta_{u,i}(Y, \hat{Y}) = (Y_{u,i} - \hat{Y}_{u,i})^2/(U \cdot I), \quad (3)$$

$$\text{Accuracy: } \delta_{u,i}(Y, \hat{Y}) = 1 \{\hat{Y}_{u,i} = Y_{u,i}\}/(U \cdot I), \quad (4)$$

Since $Y$ is only partially known, the conventional practice is to estimate $R(\hat{Y})$ using the average over only the ob-

![Figure 1. MovieLovers toy example. Top row: true rating matrix $Y$, propensity matrix $P$, observation indicator matrix $O$. Bottom row: two rating prediction matrices $\hat{Y}_1$ and $\hat{Y}_2$, and intervention indicator matrix $Y_3$.](image-url)
served entries,
\[
\hat{R}_{naive}(\hat{Y}) = \frac{U \cdot I}{|O_{u,i} = 1|} \sum_{(u,i):O_{u,i} = 1} \delta_{u,i}(Y, \hat{Y}). \tag{5}
\]

We call this the naive estimator, and its naivety leads to a gross misjudgment for the \(\hat{Y}_1\) and \(\hat{Y}_2\) given in Figure 1. Even though \(\hat{Y}_1\) is clearly better than \(\hat{Y}_2\) by any reasonable measure of performance, \(\hat{R}_{naive}(\hat{Y})\) will reliably claim that \(\hat{Y}_2\) has better MAE than \(\hat{Y}_1\). This error is due to selection bias, since 1-star ratings are underrepresented in the observed data and \(\delta_{u,i}(Y, \hat{Y})\) is correlated with \(Y_{u,i}\). More generally, under selection bias \(\hat{R}_{naive}(\hat{Y})\) is not an unbiased estimate of the true performance \(R(Y)\) of \(\hat{Y}\):
\[
E_O[\hat{R}_{naive}(\hat{Y})] \neq R(\hat{Y}). \tag{6}
\]

Before tackling the problem of designing an improved estimator to replace \(\hat{R}_{naive}(\hat{Y})\), let’s turn to a closely related evaluation task.

### 3.2. Task 2: Estimating Recommendation Quality

Instead of evaluating the accuracy of predicted ratings, we may want to more directly evaluate the quality of a particular recommendation. To this effect, let’s redefine \(\hat{Y}\) to now encode recommendations as a binary matrix analogous to \(O\), where \([Y_{u,i} = 1] \iff \{i\text{ is recommended to } u\}\), limited to a budget of \(k\) recommendations per user. An example is \(\hat{Y}_3\) in Figure 1. A reasonable way to measure the quality of a recommendation is the Cumulative Gain (CG) that the user derives from the recommended movies, which we define as the average star-rating of the recommended movies in our toy example\(^1\). CG can again be written in the form of Equation (1) with

\[
CG: \delta_{u,i}(Y, \hat{Y}) = \hat{Y}_{u,i} \cdot Y_{u,i} / (k \cdot U). \tag{7}
\]

However, unless users have watched all movies in \(\hat{Y}\), we cannot compute CG directly via (1). Hence, we are faced with the following counterfactual question: how well would our users have enjoyed themselves (in terms of CG), if they had followed our recommendations \(\hat{Y}\) instead of watching the movies indicated in \(O\)? Note that rankings of recommendations are similar to the set-based recommendation described above, and measures like Discounted Cumulative Gain (DCG), DCG@k, Precision at k (PREC@k), and others (Aslam et al., 2006; Yilmaz et al., 2008) also fit in this setting. For those, let the values of \(\hat{Y}\) in each row define the predicted ranking, then

\[
DCG: \delta_{u,i}(Y, \hat{Y}) = Y_{u,i} / (\log(\text{rank}(Y_{u,i})) \cdot U), \tag{8}
\]

\[
PRECK: \delta_{u,i}(Y, \hat{Y}) = Y_{u,i} \cdot I(\text{rank}(\hat{Y}_{u,i}) \leq k) / (k \cdot U). \tag{9}
\]

^1More realistically, \(Y\) would contain quality scores derived from behavioral indicator like “clicked” and “watched to the end.”

One approach, similar in spirit to condensed DCG (Sakai, 2007), is to use a naive estimator analogous to (5).
\[
\hat{R}_{naive}(\hat{Y}) = \frac{U \cdot I}{|O_{u,i} = 1|} \sum_{(u,i):O_{u,i} = 1} \delta_{u,i}(Y, \hat{Y}). \tag{10}
\]

But, again, this estimator is generally biased for \(R(\hat{Y})\).

The key to getting unbiased counterfactual estimates of recommendation quality despite missing observations lies in the following connection to estimating average treatment effects of a given policy in causal inference, already explored in the contextual bandit setting (Li et al., 2011; Dudík et al., 2011). If we think of making a recommendation as an intervention analogous to treating a patient with a specific drug, in both settings we want to estimate the effect of a new treatment policy (e.g. give drug A to women and drug B to men, or new recommendations \(\hat{Y}\)). The challenge in both cases is that we have only partial knowledge of how much certain patients (users) benefited from certain treatments (movies) (i.e., \(Y_{u,i}\) with \(O_{u,i} = 1\)), but the vast majority of potential outcomes in \(Y\) is unobserved.

### 3.3. Propensity-Scored Performance Estimators

The key to handling selection bias in both of the above-mentioned evaluation tasks lies in understanding the process that generates the observation pattern in \(O\). This process is typically called the Assignment Mechanism in causal inference (Imbens & Rubin, 2015) or the Missing Data Mechanism in missing data analysis (Little & Rubin, 2002). We differentiate the following two settings:

**Experimental Setting.** In this setting, the assignment mechanism is under the control of the recommendation system. An example is an ad-placement system that controls which ads to show to which user.

**Observational Setting.** In this setting, the users are part of the assignment mechanism that generates \(O\). An example is an online streaming service for movies, where users self-select the movies they watch and rate.

In this paper, we assume that the assignment mechanism is probabilistic, meaning that the marginal probability \(P_{u,i} = P(O_{u,i} = 1)\) of observing an entry \(Y_{u,i}\) is non-zero for all user/item pairs. This ensures that, in principle, every element of \(Y\) could be observed, even though any particular \(O\) reveals only a small subset. We refer to \(P_{u,i}\) as the propensity of observing \(Y_{u,i}\). In the experimental setting, we know the matrix \(P\) of all propensities, since we have implemented the assignment mechanism. In the observational setting, we will need to estimate \(P\) from the observed matrix \(O\). We defer the discussion of propensity estimation to Section 5, and focus on the experimental setting for the sake of exposition first.
Recommendations as Treatments: Debiasing Learning and Evaluation

**IPS Estimator.** The following defines the Inverse-Propensity-Scored (IPS) estimator (Thompson, 2012; Little & Rubin, 2002; Imbens & Rubin, 2015), which applies equally to the task of rating prediction evaluation as to the task of recommendation quality estimation.

\[
\hat{R}_{IPS}(\hat{Y}|P) = \sum_{(u,i):O_{u,i}=1} \frac{\delta_{u,i}(Y,\hat{Y})}{P_{u,i}} \quad .
\]

Unlike the naive estimator \(\hat{R}_{naive}(\hat{Y})\), the IPS estimator is unbiased for any probabilistic assignment mechanism.

\[
\mathbb{E}_O\left[\hat{R}_{IPS}(\hat{Y}|P)\right] = \sum_u \sum_i \mathbb{E}_{O_{u,i}} \left[\frac{\delta_{u,i}(Y,\hat{Y})}{P_{u,i}} O_{u,i}\right] = \sum_u \sum_i \delta_{u,i}(Y,\hat{Y}) = R(\hat{Y}) \quad .
\]

Note that the \(P_{u,i}\) are merely the marginal probabilities of observing an entry, and that unbiasedness is not affected by dependencies within \(O\). To characterize the variability of the IPS estimator, however, we assume that observations are independent given \(P\), which corresponds to a multivariate Bernoulli model where each \(O_{u,i}\) is a biased coin flip with probability \(P_{u,i}\). The following proposition (proof in appendix) provides some intuition about how the accuracy of the IPS estimator changes as the propensities become more "non-uniform".

**Proposition 3.1** (Tail Bound for IPS Estimator). Let \(P\) be the independent Bernoulli probabilities of observing each entry. For any given \(Y\), with probability \(1 - \eta\), the IPS estimator \(\hat{R}(\hat{Y}; P)\) does not deviate from the true \(R(\hat{Y})\) by more than:

\[
\left| \hat{R}_{IPS}(\hat{Y}|P) - R(\hat{Y}) \right| \leq \sqrt {-\log \frac{2\eta}{\rho^2}} \quad ,
\]

where \(\rho_{u,i} = \frac{\delta_{u,i}(Y,\hat{Y})}{P_{u,i}}\) if \(P_{u,i} < 1\), and \(\rho_{u,i} = 0\) otherwise.

To illustrate this bound, consider the case of uniform propensities \(P_{u,i} = p\). This means that \(n = pU \cap I\) elements of \(Y\) are revealed in expectation. In this case, the bound is \(O\left(\frac{1}{\sqrt{pU \cap I}}\right)\) for bounded \(\delta_{u,i}(Y,\hat{Y})\). If the \(P_{u,i}\) are non-uniform, the bound will be larger even if \(\sum P_{u,i}\) also equals \(n\). This indicates that we are paying for the unbiasedness of IPS in terms of variability, and we will evaluate whether this price is well spent throughout this paper.

**SNIPS Estimator.** One technique that can reduce variability is the use of control variates (Owen, 2013). Applied to the IPS estimator, we can use the knowledge that

\[
\mathbb{E}_O\left[ \sum_{(u,i):O_{u,i}=1} \frac{1}{P_{u,i}} \right] = U \cdot I.
\]

This yields the Self-Normalized Inverse Propensity Scoring (SNIPS) estimator (Trotter & Tukey, 1956; Swaminathan & Jouchsims, 2015)

\[
\hat{R}_{SNIPS}(\hat{Y}|P) = \frac{\sum_{(u,i):O_{u,i}=1} \delta_{u,i}(Y,\hat{Y})}{\sum_{(u,i):O_{u,i}=1} \frac{1}{P_{u,i}}} \quad .
\]

The SNIPS estimator often has lower variance than the IPS estimator but has a small bias (Hesterberg, 1995).

### 3.4. Empirical Illustration of Estimators

To illustrate the effectiveness of the proposed estimators in comparison to the naive estimator, we conducted an experiment on the semi-synthetic ML100K dataset described in more detail in Section 6.2. For this dataset, \(Y\) is completely known so that we can compute true performance via Equation (1). The probability \(P_{u,i}\) of observing a rating \(Y_{u,i}\) was chosen to mimic the observed marginals in ML100K (see Section 6.2) such that 1% of the \(O\) matrix was non-zero.

Table 1 shows the results for estimating rating prediction accuracy via MAE and recommendation quality via DCG@50 for the following five prediction matrices \(\hat{Y}_i\). Let \(|Y = r|\) be the number of \(r\)-star ratings in \(Y\).

**RECONES:** The prediction matrix \(\hat{Y}\) is identical to the true rating matrix \(Y\), except that \(|Y = 5|\) randomly selected true ratings of 1 are flipped to 5. This means half of the predicted fives are true fives, and half are true ones.

**RECFOURS:** Same as RECONES, but flipping 4-star ratings instead.

**ROTATE:** For each predicted rating \(\hat{Y}_{u,i} = Y_{u,i} - 1\) when \(Y_{u,i} \geq 2\), and \(\hat{Y}_{u,i} = 5\) when \(Y_{u,i} = 1\).

**SKEWED:** Predictions \(\hat{Y}_{u,i}\) are sampled via \(N(Y_{u,i}|\mu = Y_{u,i}, \sigma = \frac{\sigma_0 Y_{u,i}}{2})\) and clipped to the interval \([0, 6]\).

**COARSENED:** If the true rating \(Y_{u,i} \leq 3\), then \(\hat{Y}_{u,i} = 3\). Otherwise \(\hat{Y}_{u,i} = 4\).

As we can see in Table 1, the mean IPS estimate averaged over 50 samples of \(O\) from \(P\) perfectly matches the true
4. Propensity-Scored Recommendation Learning

The previous section derived effective performance estimators for evaluating a prediction $\hat{Y}$. In the following, we will use these estimators as part of an Empirical Risk Minimization (ERM) framework for learning, prove generalization error bounds for this ERM framework, and derive a matrix factorization method for rating prediction.

4.1. Empirical Risk Minimization for Recommendation with Propensities

Empirical Risk Minimization underlies many successful learning algorithms like SVMs (Cortes & Vapnik, 1995), Boosting (Schapire, 1990), and Deep Networks (Bengio, 2009), and weighted ERM approaches have been effective for cost-sensitive classification, domain adaptation and covariate shift (Zadrozny et al., 2003; Bickel et al., 2009; Sugiyama & Kawanabe, 2012). We adapt ERM to our setting by realizing that Equation (1) corresponds to an expected loss (i.e. Risk) over the data generating process $P(O|P)$. Given a sample from $P(O|P)$, we can think of the IPS estimator from Equation (11) as the Empirical Risk $\hat{R}(\hat{Y})$ that estimates $R(\hat{Y})$ for any $\hat{Y}$.

**Definition 4.1 (Propensity-Scored ERM for Recommendation).** Given training observations $O$ from $Y$ with marginal probabilities $P$, given a hypothesis space $H$ of predictions $\hat{Y}$, and given a loss function $\delta_{u,i}(Y,\hat{Y})$. ERM selects the $\hat{Y} \in H$ that optimizes:

$$\hat{Y}^{ERM} = \arg\min_{\hat{Y} \in H} \{ \hat{R}_{IPS}(\hat{Y}|P) \}.$$  

(15)

Note that using the SNIPS estimator does not change the argmax. To illustrate the validity of the propensity-scored ERM approach, we state the following generalization error bound (proof in appendix) similar to Cortes et al. (2010). We consider only finite $H$ for the sake of conciseness.

**Theorem 4.2 (Propensity-Scored ERM Generalization Error Bound).** For any finite hypothesis space of predictions $H = \{\hat{Y}_1, ..., \hat{Y}_|H|\}$ and loss $0 \leq \delta_{u,i}(Y,\hat{Y}) \leq \Delta$, the true risk $R(\hat{Y})$ of the empirical risk minimizer $\hat{Y}^{ERM}$ from $H$ using the IPS estimator, given training observations $O$ from $Y$ with independent Bernoulli propensities $P$, is bounded with probability $1 - \eta$ by:

$$R(\hat{Y}^{ERM}) \leq \hat{R}_{IPS}(\hat{Y}^{ERM}|P) + \Delta \sqrt{\frac{\log (2|H|/\eta)}{2\sum_{u,i} \frac{1}{P_{u,i}}}}. \quad (16)$$

4.2. Propensity-Scored Matrix Factorization

We now use propensity-scored ERM to derive a matrix factorization method for the problem of rating prediction. Assume a standard rank $d$-restricted and $L_2$-regularized matrix factorization model $\hat{Y}_{u,i} = u^T_{u}v_i + a_u + b_i + c$ with user, item, and global offsets as our hypothesis space $H$. Under this model, propensity-scored ERM leads to the following training objective:

$$\min_{U,V,A} \left[ \sum_{u,i} \delta_{u,i}(Y_{u,i}U^TV+A) + \lambda (\|U\|^2_F + \|V\|^2_F) \right]. \quad (17)$$

where $A$ encodes the offset terms and $\hat{Y}^{ERM} = U^TV + A$. Except for the propensities $P_{u,i}$ that act like weights for each loss term, the training objective is identical to the standard incomplete matrix factorization objective (Koren, 2008; Steck, 2010; Hu et al., 2008) with MSE (3) or MAE (2). This means that we can readily draw upon the existing arsenal of optimization algorithms (i.e., Gemulla et al., 2011; Yu et al., 2012), making the training problem highly scalable. For the experiments reported in this paper, we use Limited-memory BFGS (Byrd et al., 1995). Our implementation is available online at xxxxxxx.

Note that conventional incomplete matrix factorization is a special case of (17) for MCAR data, i.e., all propensities $P_{u,i}$ are equal. Solving this training objective for other $\delta_{u,i}(Y,\hat{Y})$ that are non-differentiable is more challenging, but possible avenues exist (Joachims, 2005; Chapelle & Wu, 2010). Finally, note that other recommendation methods (e.g., Weimer et al., 2007; Lin, 2007) can in principle be adapted to propensity scoring as well.

5. Propensity Estimation for Observational Data

We now turn to the Observational Setting where propensities need to be estimated, since users make self-directed choices of which observations to reveal. One might be worried that one needs to perfectly reconstruct all propensities. However, as we will show, we merely need estimated propensities that are “better” than the naive assumption of observations being revealed uniformly, i.e., $P_{u,i} = |O_{u,i} = 1| / (U \cdot I)$ for all users and items. The
following characterizes what constitutes “better” propensities both in terms of the bias they induce, as well as their effect on the variability of the learning process.

**Lemma 5.1** (Bias of IPS Estimator under Inaccurate Propensities). Let \( P \) be the marginal probabilities of observing an entry of the rating matrix \( Y \), and let \( \hat{P} \) be the estimated propensities. The bias of the IPS estimator (1) using \( \hat{P} \) is:

\[
\text{bias} \left( \hat{R}_{IPS}(\hat{Y} | \hat{P}) \right) = \sum_{u,i} \delta_{u,i}(Y, \hat{Y}) \left[ 1 - \frac{P_{u,i}}{\hat{P}_{u,i}} \right].
\]

In addition to bias, the following generalization error bound (proof in appendix) characterizes the overall impact of the estimated propensities on the learning process.

**Theorem 5.2** (Propensity-Scored ERM Generalization Error Bound under Inaccurate Propensities). For any finite hypothesis space of predictions \( \mathcal{H} = \{ \hat{Y}_1, ..., \hat{Y}_{\mathcal{H}} \} \), the prediction error of the empirical risk minimizer \( \hat{Y}^{ERM} \) under the IPS estimator with estimated propensities \( \hat{P} \), given training observations \( O \) from \( Y \) with independent Bernoulli propensities \( P \), is bounded by:

\[
R(\hat{Y}^{ERM}) \leq \hat{R}_{IPS}(\hat{Y}^{ERM} | \hat{P}) + \Delta \sum_{u,i} \left| 1 - \frac{P_{u,i}}{\hat{P}_{u,i}} \right|
+ \Delta \sqrt{\frac{\log(2|\mathcal{H}|/\eta)}{2}} \sqrt{\sum_{u,i} \frac{1}{\hat{P}^2_{u,i}}}. \]

Note that the bound shows a bias-variance tradeoff that does not occur in conventional ERM learning. In particular, the bound suggests that it can be beneficial to overestimate small propensities, if this reduces the variability term more than it increases the bias term.

### 5.1. Propensity Estimation via Naive Bayes

For our assignment model, we need to estimate a collection of Bernoulli random variables \( O_{u,i} \), where the parameters \( P_{u,i} \) corresponds to the propensities we seek. In principle, the distribution

\[
P(O \mid X, X^{hid}, Y) \]  

can depend on some observable features \( X \) (e.g., the predicted rating displayed to the user), unobservable features \( X^{hid} \) (e.g., whether the movie was recommended by a friend), and the ratings \( Y \). It is reasonable to assume that \( O_{u,i} \) is independent of the new predictions \( \hat{Y} \) (and therefore independent of \( \delta_{u,i}(Y, \hat{Y}) \)) once the observable features are taken into account. The following outlines two propensity estimation methods, but there is a wide range of other techniques available (e.g., McCaffrey et al., 2004)

In our first approach to estimating \( P(O \mid X, X^{hid}, Y) \), we assume that dependencies from covariates \( X \) and \( X^{hid} \) and other ratings are negligible, reducing the model to \( P(O_{u,i} \mid Y_{u,i}) \) similar to Marlin & Zemel (2009). Note that we can treat \( Y_{u,i} \) as observed, since we only need to know the propensity for observed entries to compute IPS and SNIPS. This yields the Naive Bayes propensity estimator:

\[
P(O_{u,i} \mid Y_{u,i}) = \frac{P(Y_{u,i} \mid O_{u,i}) P(O_{u,i})}{P(Y_{u,i})} . \]

Of these quantities, we can estimate \( P(Y_{u,i} \mid O_{u,i} = 1) \) and \( P(O_{u,i}) \) even from MNAR data. However, we do need extra information to estimate \( P(Y_{u,i}) \), the proportion of different rating values in \( Y \). This can be estimated accurately from even a small experimental (e.g. uniform) sample.

### 5.2. Propensity Estimation via Logistic Regression

The second propensity estimation approach we explore in our experiments is based on logistic regression, as commonly used in statistics (Rosenbaum, 2002). It also starts from (20), but aims to find model parameters \( \phi \) such that \( O \) becomes independent of unobserved \( X^{hid} \) and \( Y \), i.e., \( P(O \mid X, X^{hid}, Y) = P(O \mid X, \phi) \). The main modeling assumption is that there exists a \( \phi = (w, \beta, \gamma) \) such that

\[
P(O_{u,i} = \sigma(w^T X_{u,i} + \beta + \gamma \delta_{u,i}) . \]

Here, \( X_{u,i} \) is a vector encoding all observable information about a user-movie pair (e.g., user demographics, whether a movie was promoted), and \( \sigma(\cdot) \) is the sigmoid function. \( \beta \) and \( \gamma \) are per-movie and per-user offsets.

### 6. Empirical Evaluation

We conduct extensive experiments to explore the empirical performance and robustness of the proposed methods in both the experimental and the observational setting. Furthermore, we compare against the state-of-the-art joint-likelihood method for MNAR data (Hernández-Lobato et al., 2014) on real-world datasets.

#### 6.1. Experiment Setup

In all experiments, we perform model selection for the regularization parameter \( \lambda \) and/or the rank of the factorization \( d \) via cross-validation as follows. We randomly split the observed MNAR ratings into four folds, training on three and evaluating on the remaining one using the IPS estimator. Note that reflecting this additional split in the propensities requires scaling the propensities in the training folds by \( \frac{k-1}{k} \) and those in the validation fold by \( \frac{k}{k-1} \). The parameters with the best average validation set performance are used to retrain over all MNAR data. We report test performance on the MCAR test samples that come with the real-world datasets, and via Equation (1) for our semi-synthetic data.
6.2. How is evaluation affected by sampling bias severity?

In this experiment we evaluate how different observation distributions impact the accuracy of the performance estimators. We compare the Naive estimators for MSE (5), MAE (5) and DCG (10) with their propensity-weighted analogues, IPS (11) and SNIPS (14) respectively. Since this experiment requires experimental control of sampling bias, we created a semi-synthetic dataset and observation model.

ML100K Dataset. The ML100K dataset\(^2\) provides 100K MNAR ratings for 1683 movies by 944 users. To allow ground-truth evaluation against a fully known rating matrix, we complete these partial ratings using a standard matrix factorization with \(d = 100\). The completed matrix, however, gives unrealistically high ratings to almost all movies. We therefore adjust ratings for the final \(Y\) to match a more realistic rating distribution \([p_1, p_2, p_3, p_4, p_5]\) for ratings 1 to 5 as given in Marlin & Zemel (2009) as follows: we assign the bottom \(p_1\) fraction of the entries by value in the completed matrix a rating of 1 and the next \(p_2\) fraction of entries by value a rating of 2 and so on.

ML100K Observation Model. If the underlying rating is 4 or 5, the propensity for observing the rating is equal to \(k\). For ratings \(r < 4\), the corresponding propensity is \(k\alpha^{4-r}\). For each \(k\), \(\alpha\) is set so that the expected number of ratings we observe is 1% of the entire matrix. By varying \(\alpha > 0\), we vary the MNAR effect: \(\alpha = 1\) is missing uniformly at random (MCAR), while \(\alpha \to 0\) only reveals 4 and 5 rated items. Note that \(\alpha = 0.25\) gives a marginal distribution of observed ratings that reasonably matches the observed MNAR rating marginals on ML100K \([0.06, 0.11, 0.27, 0.35, 0.21]\) in the real data vs. \([0.06, 0.10, 0.25, 0.42, 0.17]\) in our model.

Results. Table 1, described in Section 3.4, shows the estimated MAE and DCG@50 when \(\alpha = 0.25\). Next, we vary the severity of the sampling bias by changing \(\alpha \in (0, 1]\). Figure 2 reports the root mean squared estimation error (RMSE) when estimating MSE, MAE, and DCG respectively. These results are for the Experimental Setting where propensities are known. They are averages over the five prediction matrices \(Y\) given in Section 3.4 and across 50 trials. Shaded regions indicate a 95% confidence interval.

Over most of the range of \(\alpha\), in particular for the realistic value of \(\alpha = 0.25\), the IPS and SNIPS estimators are orders-of-magnitude more accurate than the conventional Naive estimator. Even for severely low choices of \(\alpha\), the gain in bias reduction of IPS and SNIPS still outweighs the added variability compared to Naive. Note that for \(\alpha = 1\), SNIPS is algebraically equivalent to Naive, while IPS pays a small penalty due to increased variability from propensity weighting. For MSE and MAE, SNIPS consistently reduces estimation error over IPS while both are tied for DCG.

6.3. How is learning affected by sampling bias severity?

The following experiment explores whether these gains in risk estimation accuracy translate into improved learning via ERM, again in the Experimental Setting. Using the same semi-synthetic ML100K dataset and observation model as above, we compare our matrix factorization MF-IPS with the traditional unweighted matrix factorization MF-Naive. Both methods use the same factorization model with separate \(\lambda\) selected via cross-validation and \(d = 20\). The results are plotted in Figure 3 (left), where shaded regions indicate 95% confidence intervals over 40 trials. The propensity-weighted matrix factorization MF-IPS outperforms the conventional matrix factorization by an order of magnitude in terms of MSE. We also conducted experiments for MAE, and the results are similar. Somewhat surprisingly, learning appears even less affected by severe selection bias (i.e., small \(\alpha\)) than evaluation. We conjecture that this is due to many of the critical \(Y\) sharing the variability of the IPS estimates, such that their ranking stays fairly stable.

6.4. How robust is evaluation and learning to inaccurately learned propensities?

We now switch from the Experimental Setting to the Observational Setting, where propensities need to be estimated. To explore robustness to propensity estimates of varying accuracy, we use the ML100K data and observation model with \(\alpha = 0.25\). To generate increasingly bad propen-

\(^2http://grouplens.org/datasets/movielens/\)
Figure 4. RMSE of IPS and SNIPS as propensity estimates degrade. IPS with true propensities and Naive are given as reference.

6.5. Performance on Real-World Data

Our final experiments concern performance on real-world datasets. We use the following two datasets, which both have a separate test set where users were asked to rate a uniformly drawn sample of items.

Yahoo! R3 Dataset. This dataset\(^3\) (Marlin & Zemel, 2009) contains user-song ratings. The MNAR training set provides over 300K ratings for songs that were self-selected by 15400 users. The test set contains ratings by a subset of 5400 users who were asked to rate 10 randomly chosen songs. For this data, we estimate propensities via Naive Bayes. As a uniform sample for eliciting the marginal rating distribution, we set aside 5% of the uniform sample and only report results on the remaining 95% of the test set.

\(^3\)http://webscope.sandbox.yahoo.com/

Table 2. Test set MAE and MSE on the Yahoo and Coat datasets.

<table>
<thead>
<tr>
<th></th>
<th>YAHOO</th>
<th>COAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>MF-IPS</td>
<td>0.810</td>
<td>0.989</td>
</tr>
<tr>
<td>MF-Naive</td>
<td>1.154</td>
<td>1.891</td>
</tr>
<tr>
<td>HL MNAR</td>
<td>1.177</td>
<td>2.175</td>
</tr>
<tr>
<td>HL MAR</td>
<td>1.179</td>
<td>2.166</td>
</tr>
</tbody>
</table>

Coat Shopping Dataset. We collected a new dataset\(^4\) simulating MNAR data of customers shopping for a coat in an online store. The training data was generated by giving Amazon Mechanical Turkers a simple web-shop interface with facets and paging. They were asked to find the coat in the store that they wanted to buy the most. Afterwards, they had to rate 24 of the coats they explored (self-selected) and 16 randomly picked ones on a five-point scale. The self-selected ratings are the training set, the uniformly selected ratings are the test set. We learn propensities via logistic regression based on user covariates (gender, age group, location, and fashion-awareness) and item covariates (gender, coat type, color, and was it promoted). A standard regularized logistic regression (Pedregosa et al., 2011) was trained using all pairs of user and item covariates as features and cross-validated to optimize log-likelihood of the self-selected observations.

Results. Table 2 shows that our propensity-scored matrix factorization MF-IPS with learnt propensities substantially and significantly outperforms the conventional matrix factorization approach, as well as the Bayesian imputation models from (Hernández-Lobato et al., 2014), abbreviated as HL-MNAR and HL-MAR (paired t-test, \(p < 0.001\) for all). This holds for both MAE and MSE. Furthermore, the performance of MF-IPS beats the best published results for Yahoo in terms of MSE (1.115) and is close in terms of MAE (0.770) (the CTP-v model of (Marlin & Zemel, 2009) as reported in the supplementary material of Hernández-Lobato et al. (2014)). For MF-IPS and MF-Naive all hyperparameters (i.e., \(\lambda \in \{10^{-6}, ..., 1\}\) and \(d \in \{5, 10, 20, 40\}\)) were chosen by cross-validation. For the HL baselines, we explored \(d \in \{5, 10, 20, 40\}\) using software provided by the authors\(^5\) and report the best performance on the test set for efficiency reasons. Note that our performance numbers for HL on Yahoo closely match the values reported in (Hernández-Lobato et al., 2014).

Compared to the complex HL models, we conclude that MF-IPS performs robustly and efficiently on real-world data. We conjecture that this strength is a result of not requiring any generative assumptions about the validity of

\(^4\)www.XXXXXX.xxx/XXXXXXX/XXXXX/
\(^5\)https://bitbucket.org/jmh233/missingdataicml2014
the rating model. Furthermore, note that there are several promising directions for further improving performance, like propensity clipping (Srethl et al., 2010), doubly-robust estimation (Dudík et al., 2011), and the use of improved methods for propensity estimation (McCaffrey et al., 2004).

7. Conclusions

The paper proposed an effective and robust approach to handling selection bias in the evaluation and learning of recommender systems based on propensity scoring. The modularity of the approach—separating the estimation of the assignment model from the rating model—also makes it very practical. In particular, any conditional probability estimation method can be plugged in as the propensity estimator, and we conjecture that many existing rating models can be retrofit with propensity weighting without sacrificing scalability.

References


Recommendations as Treatments: Debiasing Learning and Evaluation


Owen, Art B. Monte Carlo theory, methods and examples. 2013.


