Nonparametric Preference Completion

Julian Katz-Samuels
University of Michigan

Clayton Scott
University of Michigan

Abstract

We consider the task of collaborative preference completion: given a pool of items, a pool of users and a partially observed item-user rating matrix, the goal is to recover the *personalized ranking* of each user over all of the items. Our approach is nonparametric: we assume that each item i and each user u have unobserved features x_i and y_u, and that the associated rating is given by $g_u(f(x_i, y_u))$ where f is Lipschitz and g_u is a monotonic transformation that depends on the user. We propose a k-nearest neighbors-like algorithm and prove that it is consistent. To the best of our knowledge, this is the first consistency result for the collaborative preference completion problem in a nonparametric setting. Finally, we demonstrate the performance of our algorithm with experiments on the Netflix and Movielens datasets.

1 Introduction

In the preference completion problem, there is a pool of items and a pool of users. Each user rates a subset of the items and the goal is to recover the personalized ranking of each user over all of the items. This problem is fundamental to recommender systems, arising in tasks such as movie recommendation and news personalization. A common approach is to first estimate the ratings through either a matrix factorization method or a neighborhood-based method and to output personalized rankings from the estimated ratings (Koren et al., 2009; Zhou et al., 2008; Ning et al., 2011; Breese et al., 1998). Recent research has observed a number of shortcomings of this approach (Weimer et al., 2007; Liu and Yang, 2008); for example, many ratings-oriented algorithms minimize the RMSE, which does not necessarily produce a good ranking (Cremonesi et al., 2010). This observation has sparked a number of proposals of algorithms that aim to directly recover the rankings (Weimer et al., 2007; Liu and Yang, 2008; Lu and Negahban, 2014; Park et al., 2015; Oh et al., 2015; Gunasekar et al., 2016). Although these ranking-oriented algorithms have strong empirical performance, there are few theoretical guarantees to date and they all make specific distributional assumptions (discussed in more detail below). In addition, these results have focused on low-rank methods, while ranking-oriented neighborhood-based methods have received little theoretical attention.

In this paper, we consider a statistical framework for nonparametric preference completion. We assume that each item i and each user u have unobserved features x_i and y_u, respectively, and that the associated rating is given by $g_u(f(x_i, y_u))$ where f is Lipschitz and g_u is a monotonic transformation that depends on the user. We make the following contributions. (i) We propose a simple k-nearest neighbors-like algorithm, (ii) we provide, to the best of our knowledge, the first consistency result for ranking-oriented algorithms in a nonparametric setting, and (iii) we provide a necessary and sufficient condition for the optimality of a solution (defined below) to the preference completion problem.

2 Related Work

The two main approaches to preference completion are matrix factorization methods (e.g., low-rank approximation) and neighborhood-based methods. Recently, there has been a surge of research with many theoretical advances in low-rank approximation for collaborative filtering, e.g., (Recht, 2011; Keshavan et al., 2010). These methods tend to focus on minimizing the RMSE even though applications usually use ranking measures to evaluate performance. While recent work has developed ranking-oriented algorithms that outperform ratings-oriented algorithms (Gunasekar et al., 2016; Liu and Yang, 2008; Rendle et al., 2009; Pessiot et al., 2007; Cremonesi et al., 2010; Weimer et al., 2007), many of these proposals lack basic theoretical guarantees such as consistency. A recent line of work

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

has begun to fill this gap by establishing theoretical results under specific generative models. Lu and Negahban (2014) and Park et al. (2015) provided consistency guarantees using a low-rank approach and the Bradley-Terry-Luce model. Similarly, Oh et al. (2015) established a consistency guarantee using a low rank approach and the MultiNomial Logit model. By contrast, our approach forgoes such strong parametric assumptions.

Neighborhood-based algorithms are popular methods, e.g. (Das et al., 2007), because they are straightforward to implement, do not require expensive model-training, and generate interpretable recommendations (Ning et al., 2011). There is an extensive experimental literature on neighborhood-based collaborative filtering methods. The most common approach is the user-based model; it is based on the intuition that if two users give similar ratings to items in the observed data, then their unobserved ratings are likely to be similar. This approach employs variants of k nearest-neighbors. Popular similarity measures include the Pearson Correlation coefficient and cosine similarity. There are a large number of schemes for predicting the unobserved ratings using the k nearest neighbors, including taking a weighted average of the ratings of the users and majority vote of the users (Ning et al., 2011).

Recently, researchers have sought to develop neighborhood-based collaborative filtering algorithms that aim to learn a personalized ranking for each user instead of each user's ratings (Liu and Yang, 2008; Wang et al., 2014, 2016). Eigenrank, proposed by Liu and Yang (2008), is structurally similar to our algorithm. It measures the similarity between users with the Kendall rank correlation coefficient, a measure of the similarity of two rankings. Then, it computes a utility function $\psi : [n_1] \times [n_1] \longrightarrow \mathbb{R}$ for each user that estimates his pairwise preferences over the items. From the estimated pairwise preferences, it constructs a personalized ranking for each user by either using a greedy algorithm or random walk model. In contrast, our algorithm uses the average number of agreements on pairs of items to measure similarity between users and a majority vote approach to predict pairwise preferences.

Neighborhood-based collaborative filtering has not received much theoretical attention. Kleinberg and Sandler (2003, 2004) model neighborhood-based collaborative filtering as a latent mixture model and prove consistency results in this specific generative setting. Recently, Lee et al. (2016), who inspired the framework in the current paper, studied rating-oriented neighborhood-based collaborative filtering in a more general nonparametric setting. Their approach assumes that each item i and each user u have unobserved features x_i and y_u, respectively, and that the associated rating is given by $f(x_i, y_u)$ where f is Lipschitz, whereas we assume that the associated rating is given by $g_u(f(x_i, y_u))$ where g_u is a user-specific monotonic transformation. As we demonstrate in our experiments, their algorithm is not robust to monotonic transformations of the columns, but this robustness is critical for many applications. For example, consider the following implicit feedback problem (Hu et al., 2008). A recommender system for news articles measures how long users read articles as a proxy for item-user ratings. Because reading speeds and attention spans vary dramatically, two users may actually have very similar preferences despite substantial differences in reading times.

Even though our method is robust to user-specific monotonic transformations, we do not require observing many more entries of the item-user matrix than Lee et al. (2016) in the regime where there are many more users than items (e.g., the Netflix dataset). If there are n_1 items and n_2 users, Lee et al. (2016) requires that there exists $\frac{1}{2} > \alpha > 0$ such that the probability of observing an entry is greater than $\max(n_1^{-\frac{1}{2}+\alpha}, n_2^{-1+\alpha})$, whereas we require that this probability is greater than $\max(n_1^{-\frac{1}{2}+\alpha}, n_2^{-\frac{1}{2}+\alpha})$.

Our work is also related to the problem of Monotonic Matrix Completion (MMC) where a single monotonic Lipschitz function is applied to a low rank matrix and the goal is rating estimation (Ganti et al., 2015). In contrast, we allow for distinct monotonic, possibly non-Lipschitz functions for every user and pursue the weaker goal of preference completion.

To the best of our knowledge, there is no theoretically supported, nonparametric method for preference completion. Our work seeks to address this issue.

3 Setup

Notation: Define $[n] = \{1, \ldots, n\}$. Let $\Omega \subset [n_1] \times [n_2]$. If $X \in \mathbb{R}^{n_1 \times n_2}$, let $\mathcal{P}_\Omega(X) \in (\mathbb{R} \cup \{?\})^{n_1 \times n_2}$ be defined as $[\mathcal{P}_\Omega(X)]_{i,j} = \begin{cases} X_{i,j} & \text{if } (i,j) \in \Omega \\ ? & \text{if } (i,j) \notin \Omega \end{cases}$. If f is some function and U a finite collection of objects belonging to the domain of f, let $\max^{(l)}_{u \in U} f(u)$ denote the lth largest value of f over U. Let $\text{Bern}(p)$ denote a realization of a Bernoulli random variable with parameter p. For a metric space $\mathcal{M}$ with metric $d_\mathcal{M}$, let $B_\epsilon(z) = \{z' \in \mathcal{M} : d_\mathcal{M}(z, z') < \epsilon\}$. We use bold type to indicate random variables. For example, $\boldsymbol{z}$ denotes a random variable and z a realization of $\boldsymbol{z}$.

Nonparametric Model: Suppose that there are n_1 items and n_2 users. Furthermore,

1. The items are associated with unobserved features $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1} \in \mathcal{X}$, and the users are associated with unobserved features $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_2} \in \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces with metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, respectively.

2. $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_2}$ are independent random variables such that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1} \overset{i.i.d.}{\sim} \mathcal{P}_{\mathcal{X}}$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_2} \overset{i.i.d.}{\sim} \mathcal{P}_{\mathcal{Y}}$ where $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ denote Borel probability measures over $\mathcal{X}$ and $\mathcal{Y}$, respectively. We assume that for all $\epsilon > 0$ and $y \in \mathcal{Y}$, $\mathcal{P}_{\mathcal{Y}}(B_\epsilon(y)) > 0$.

3. The complete ratings matrix is $H \coloneqq [h_u(x_i, y_u)]_{i \in [n_1], u \in [n_2]}$ where $h_u = g_u \circ f$, $f : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ is a Lipschitz function with respect to the induced metric $d_{\mathcal{X}\times\mathcal{Y}}((x_1, y_1), (x_2, y_2)) \coloneqq \max(d_{\mathcal{X}}(x_1, x_2), d_{\mathcal{Y}}(y_1, y_2))$ with Lipschitz constant 1,[1] i.e., $\forall y_1, y_2 \in \mathcal{Y}$ and $\forall x_1, x_2 \in \mathcal{X}$, $|f(x_1, y_1) - f(x_2, y_2)| \leqslant \max(d_{\mathcal{X}}(x_1, x_2), d_{\mathcal{Y}}(y_1, y_2))$, and g_u is a non-decreasing function. Note that each h_u need not be Lipschitz.

4. Each entry of the matrix H is observed independently with probability p. Let $\boldsymbol{\Omega} \subset [n_1] \times [n_2]$ be a random variable denoting the indices of the observed ratings.

Whereas Lee et al. (2016) considers the task of completing a partially observed matrix $F \coloneqq [f(x_i, y_u)]_{i \in [n_1], u \in [n_2]}$ when $\{x_i\}_{i \in [n_1]}$ and $\{y_u\}_{u \in [n_2]}$ are unobserved, we aim to recover the ordering of the elements in each column of H when $\{x_i\}_{i \in [n_1]}$ and $\{y_u\}_{u \in [n_2]}$ are unobserved. In our setup, we view F as an ideal preference matrix representing how much users like items and H as how those preferences are expressed based on user-specific traits (see the news recommender system example in Section 2).

This framework subsumes various parametric models. For example, consider a matrix factorization model that assumes that there is a matrix $H \in \mathbb{R}^{n_1 \times n_2}$ of rank $d \leqslant \min(n_1, n_2)$ such that user u prefers item i to item j if and only if $H_{i,u} > H_{j,u}$. Then, we can factorize H such that $H_{i,u} = x_i^t y_u$ where $x_i, y_u \in \mathbb{R}^d$ for all $i \in [n_1]$ and $u \in [n_2]$. In our setup, we have $f(x_i, y_u) = x_i^t y_u$ and $g_u(z) = z$.

Task: Let $\mathcal{S}^{n_1} = \{\sigma : \sigma : [n_1] \longrightarrow [n_1], \sigma \text{ is a permutation}\}$ denote the set of permutations on n_1 objects. We call $\sigma \in \mathcal{S}^{n_1}$ a *ranking*. Let $\mathcal{S}^{n_1 \times n_2} = (\mathcal{S}^{n_1})^{n_2}$. That is, $\sigma \in \mathcal{S}^{n_1 \times n_2}$ if $\sigma : [n_1] \times [n_2] \longrightarrow [n_1]$ and for fixed $u \in [n_2]$, $\sigma(\cdot, u)$ is a permutation on $[n_1]$. We call $\sigma \in \mathcal{S}^{n_1 \times n_2}$ a *collection of rankings*. Let $\epsilon > 0$. Our goal is to learn $\sigma \in \mathcal{S}^{n_1 \times n_2}$ that minimizes the number of pairwise ranking disagreements per user with some slack, i.e.,

$$\begin{aligned}\text{dis}_\epsilon(\sigma, H) = \sum_{u=1}^{n_2} \sum_{i<j} &\mathbf{1}\{|f(x_i, y_u) - f(x_j, y_u)| > \epsilon\} \\ &\times \mathbf{1}\{(h_u(x_i, y_u) - h_u(x_j, y_u)) \\ &\times (\sigma(i, u) - \sigma(j, u)) < 0\}.\end{aligned}$$

4 Algorithm

Our algorithm, Multi-Rank (Algorithm 1), has two stages: first it estimates the pairwise preferences of each user and, second, it constructs a full ranking for each user from its estimated pairwise preferences. In the first stage, Multi-Rank computes $A \in \{0, 1\}^{n_2 \times n_1 \times n_1}$ where $A_{u,i,j} = 1$ denotes that user u prefers item i to item j and $A_{u,i,j} = 0$ denotes that user u prefers item j to item i. If a user has provided distinct ratings for a pair of items, Multi-Rank fills in the corresponding entries of A. Otherwise, Multi-Rank uses a subroutine called Pairwise-Rank that we will describe shortly. Once Multi-Rank has constructed A, it applies the Copeland ranking procedure to the pairwise preferences of each user (discussed at the end of the section).

Algorithm 1 Multi-Rank

Input: $\mathcal{P}_\Omega(H), \beta \geqslant 2, k > 0$
for $u \in [n_2]$, $i, j \in [n_1], i < j$ **do**
  if $(i, u) \in \Omega$, $(j, u) \in \Omega$ and $H_{i,u} \neq H_{j,u}$ **then**
    Set $A_{u,i,j} = \mathbf{1}\{H_{i,u} > H_{j,u}\}$
    Set $A_{u,j,i} = 1 - A_{u,i,j}$
  else
    Set $A_{u,i,j} = \text{Pairwise-Rank}(u, i, j, \beta, k)$
    Set $A_{u,j,i} = 1 - A_{u,i,j}$
  end if
end for
for $u \in [n_2]$ **do**
  $\widehat{\sigma}_u = \text{Copeland}(A_{u,:,:})$
end for
return $\widehat{\sigma} \coloneqq (\widehat{\sigma}_1, \ldots, \widehat{\sigma}_{n_2})$

The Pairwise-Rank algorithm predicts whether a user u prefers item i to item j or vice versa. It is similar to k-nearest neighbors where we use the forthcoming ranking measure as our distance measure. Let $N(u)$ denote the set of items that user u has rated, i.e.,

$$N(u) = \{l : (l, u) \in \Omega\},$$

and $N(u, v) = N(u) \cap N(v)$ denote the set of items that users u and v have both rated. Viewing $N(u, v)$ as an

[1] We could develop our framework with an arbitrary Lipschitz constant L, but for ease of presentation, we fix $L = 1$.

ordered array where $N(u,v)[\ell]$ denotes the $(\ell+1)$th element, let

$$I(u,v) = \{(s,t) : s = N(u,v)[\ell], t = N(u,v)[\ell+1] \text{ for some } \ell \in \{2k : k \in \mathbb{N} \cup \{0\}\}\}.$$

In words, $I(u,v)$ is formed by sorting the indices of $N(u,v)$ and selecting nonoverlapping pairs in the given order. Note that there is no overlap between the indices in the pairs in $I(u,v)$.[2] Fix $y_u, y_v \in \mathcal{Y}$. If $I(u,v) = \varnothing$, define $R_{u,v} = 0$ and if $I(u,v) \neq \varnothing$, let $R_{u,v} \coloneqq$

$$\frac{1}{|I(u,v)|} \sum_{(s,t)\in I(u,v)} \mathbf{1}\{(h_u(\boldsymbol{x}_s, y_u) - h_u(\boldsymbol{x}_t, y_u)) \times (h_v(\boldsymbol{x}_s, y_v) - h_v(\boldsymbol{x}_t, y_v)) \geqslant 0\}$$

denote the fraction of times that users u and v agree on the relative ordering of item pairs belonging to $I(u,v)$. In practice, one can simply compute this statistic over all pairs of commonly rated items. Observe that $\rho(y_u, y_v) \coloneqq$

$$\mathbb{E}[R_{u,v} | I(u,v) \neq \varnothing, \boldsymbol{y}_u = y_u, \boldsymbol{y}_v = y_v] = \Pr_{\boldsymbol{x}_s, \boldsymbol{x}_t \sim P_{\mathcal{X}}}([h_u(\boldsymbol{x}_s, y_u) - h_u(\boldsymbol{x}_t, y_u)] \times [h_v(\boldsymbol{x}_s, y_v) - h_v(\boldsymbol{x}_t, y_v)] \geqslant 0)$$

i.e., $\rho(y_u, y_v)$ is the probability that users u and v with features y_u and y_v order two random items in the same way.

We apply Pairwise-Rank (Algorithm 2) to a user u and a pair of items (i,j) if the user has not provided distinct ratings for items i and j. Pairwise-Rank(u,i,j,β,k) finds users that have rated items i and j, and have rated at least β items in common with u. If there are no such users, Pairwise-Rank flips a coin to predict the relative preference ordering. If there are such users, then it sorts the users in decreasing order of $R_{u,v}$ and takes a majority vote over the first k users about whether item i or item j is preferred. If the vote results in a tie, Pairwise-Rank flips a coin to predict the relative preference ordering.

Next, Multi-Rank converts the pairwise preference predictions of each user into a full estimated ranking for each user. It applies the Copeland ranking procedure (Algorithm 3)–an algorithm for the feedback arc set problem in tournaments (Copeland, 1951; Coppersmith et al., 2006) to each user-specific set of pairwise preferences. The Copeland ranking procedure simply orders the items by the number of times an item is preferred to another item. It is possible to use other approximation algorithms for the feedback arc set problem such as Fas-Pivot from Ailon et al. (2008).

[2] We select nonoverlapping pairs to preserve independence in the estimates for the forthcoming analysis.

Algorithm 2 Pairwise-Rank

Input: $u \in [n_2], i \in [n_1], j \in [n_1], \beta \geqslant 2, k \in \mathbb{N}$
$W_u^{i,j}(\beta) = \{v \in [n_2] : |N(u,v)| \geqslant \beta, (i,v), (j,v) \in \Omega\}$
Sort $W_u^{i,j}(\beta)$ in decreasing order of $R_{u,v}$ and let V be the first k elements.
if $V = \varnothing$ **then**
  return Bern$(\frac{1}{2})$
end if
$\forall v \in V$, set $P_v = \mathbf{1}\{h_v(x_i, y_v) > h_v(x_j, y_v)\} - \mathbf{1}\{h_v(x_i, y_v) < h_v(x_j, y_v)\}$
if $\sum_{v\in V} P_v > 0$ **then**
  return 1
else if $\sum_{v\in V} P_v < 0$ **then**
  return 0
else
  return Bern$(\frac{1}{2})$
end if

Algorithm 3 Copeland

Input: $A \in \{0,1\}^{n_1 \times n_1}$
for $j \in [n_1]$ **do**
  $I_j = \sum_{i=1, i\neq j}^{n_1} A_{j,i}$
end for
return $\sigma \in \mathcal{S}^{n_1}$ that orders items in decreasing order of I_j

5 Analysis of Algorithm

The main idea behind our algorithm is to use pairwise agreements about items to infer whether two users are close to each other in the feature space. However, this is not possible in the absence of further distributional assumptions. The Lipschitz condition on f only requires that if users u and v are close to each other, then $\max_z |f(z, y_u) - f(z, y_v)|$ is small. Proposition 1 shows that there exist functions arbitrarily close to each other that disagree about the relative ordering of almost every pair of points.

Proposition 1. *Let $\mathcal{X} = [0,1]$ and $\mathcal{P}_{\mathcal{X}}$ be the Lebesgue measure over $\mathcal{X}$. For every $\epsilon > 0$, there exist functions $f, g : \mathcal{X} \longrightarrow \mathbb{R}$ such that $\max_{x\in[0,1]} |f(x) - g(x)| = \|f - g\|_\infty \leqslant \epsilon$ and for almost every pair of points $(x, x') \in [0,1]^2$, $f(x) > f(x')$ iff $g(x) < g(x')$.*

Thus, we make the following mild distributional assumption.

Definition 1. *Fix $y \in \mathcal{Y}$ and let $f_y(x) \coloneqq f(x,y)$. Let r be a positive nondecreasing function. We say y is r-*discerning *if $\forall \epsilon > 0$, $\Pr_{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{P}_{\mathcal{X}}}(|f_y(\boldsymbol{x}_1) - f_y(\boldsymbol{x}_2)| \leqslant 2\epsilon) < r(\epsilon)$.*

This assumption says that the probability that $f_y(\boldsymbol{x}_1)$ and $f_y(\boldsymbol{x}_2)$ are within ϵ of each other decays at some

rate given by r. In a sense, it means that users perceive some difference between most randomly selected items with different features, although the difference might be masked by the transformation g_u.

We also assume that if two users are not close to each other in the latent space, then they must have some disagreements. Definition 2 requires that the nonparametric model is economical (i.e., not redundant) in the sense that different parts of the feature space correspond to different preferences.

Definition 2. *Fix $y \in \mathcal{Y}$. Let $\epsilon, \delta > 0$. We say that y is* (ϵ, δ)-discriminative *if $z \in B_\epsilon(y)^c$ implies that $\rho(y, z) < 1 - \delta$.*

These assumptions are satisfied under many parametric models. Proposition 2 provides two illustrative examples under a matrix factorization model. We briefly note that, as we show in the supplementary material, $f(x, y) = x^t y$ and $f(x, y) = \|x - y\|_2$ are equivalent models by adding a dimension.

Proposition 2. *Consider $(\mathbb{R}^d, \|\cdot\|_2)$. Let $f(x, y) = \|x - y\|_2$ and $g_u(\cdot)$ be strictly increasing $\forall u \in [n_2]$.*

1. *Let $\mathcal{X} = \mathcal{Y} = \{x \in \mathbb{R}^d : \|x\|_2 \leqslant 1\}$, $\mathcal{P}_\mathcal{X}$ be the uniform distribution and for all $y \in \mathcal{Y}$ define $r_y(\epsilon) = \sup_{z \in [0,2]} \mathcal{P}_\mathcal{X}(B_z(y) \backslash B_{z-4\epsilon}(y))$. Then, for all $y \in \mathcal{Y}$, y is r_y-discerning. Further, define for all $\epsilon > 0$, $\delta_\epsilon = \inf_{v \in \mathcal{Y}} 2\mathcal{P}_\mathcal{X}(B_{\frac{\epsilon}{2}}(v))^2$. Then, for all $y_u \in \mathcal{Y}$ and for all $\epsilon > 0$, y_u is $(\epsilon, \delta_\epsilon)$-discriminative.*
2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a finite collection of points, $\mathcal{P}_\mathcal{X}$ be uniform over $\mathcal{X}$, and for all $y \in \mathcal{Y}$ define $r_y(\epsilon) = \frac{|\{(x,x') \in \mathcal{X} \times \mathcal{X} : |\|y-x\| - \|y-x'\|| \leqslant 2\epsilon\}|}{|\mathcal{X}|^2}$. Then, for all $y \in \mathcal{Y}$, y is r_y-discerning. Next, suppose $\mathcal{Y}$ is a finite collection of points and every pair of distinct $y, y' \in \mathcal{Y}$ disagree about at least C pairs of items. Let $\delta = \frac{C}{|\mathcal{X}|^2}$. For all $y_u \in \mathcal{Y}$ and for all $\epsilon > 0$, y_u is (ϵ, δ)-discriminative.*

Our analysis uses two functions to express problem-specific constants. First, let $\tau : \mathbb{R}_{++} \longrightarrow (0, 1]$ be defined as $\tau(\epsilon) = \inf_{y_0 \in \mathcal{Y}} \Pr_{\boldsymbol{y} \sim \mathcal{P}_\mathcal{Y}}(d_\mathcal{Y}(y_0, \boldsymbol{y}) \leqslant \epsilon)$. Second, let $\kappa : \mathbb{R}_{++} \longrightarrow (0, 1]$ be such that $\kappa(\epsilon) = \inf_{y_0 \in \mathcal{Y}} \Pr_{\boldsymbol{y} \sim \mathcal{P}_\mathcal{Y}}(d_\mathcal{Y}(y_0, \boldsymbol{y}) > \epsilon)$. Our assumption that for all $\delta > 0$ and $y \in \mathcal{Y}$, $\mathcal{P}_\mathcal{Y}(B_\delta(y)) > 0$ ensures that $\tau(\cdot) > 0$ and $\kappa(\cdot) < 1$ (see Lemma D.3). If $\mathcal{P}_\mathcal{Y}$ is uniform over the unit cube in $(\mathbb{R}^d, \|\cdot\|_\infty)$, then $\tau(\epsilon) = \min(1, \epsilon)^d$ and if $\mathcal{Y}$ is a finite collection of points, then $\tau(\epsilon) = \min_{y \in \mathcal{Y}} \mathcal{P}_\mathcal{Y}(y)$ (Lee et al., 2016).

Our model captures the intrinsic difficulty of a problem instance as follows. $r(\cdot)$ and $\tau(\cdot)$ together control the probability of sampling nearby users with similar preferences. (ϵ, δ)-discriminative captures how often users u and v must agree in order to infer that y_u and y_v are close in the latent space and, thus, $\max_z |f(z, y_u) - f(z, y_v)| \leqslant \epsilon$.

5.1 Continuous Ratings Setting

Our analysis deals with the case of continuous ratings and the case of discrete ratings separately. In this section, we prove theorems dealing with the continuous case and in the next section we give analogous results with similar proofs for the discrete case. Theorem 1 establishes that with probability tending to 1 as $n_2 \longrightarrow \infty$, Multi-Rank outputs $\widehat{\sigma} \in \mathcal{S}^{n_1 \times n_2}$ such that $\mathrm{dis}_{2\epsilon}(\widehat{\sigma}, H) = 0$.

Theorem 1. *Suppose $\forall u \in [n_2]$, $g_u(z)$ is strictly increasing. Let $\epsilon, \delta > 0$, $\eta \in (0, \frac{\epsilon}{2})$. Suppose that almost every $y \in \mathcal{Y}$ is $(\frac{\epsilon}{2}, \delta)$-discriminative. Let r be a positive nondecreasing function such that $r(\frac{\epsilon}{2}) \geqslant \delta$ and $r(\eta) < \frac{\delta}{2}$. Suppose that almost every $y \in \mathcal{Y}$ is r-discerning. Let $0 < \alpha < \frac{1}{2}$. If $p \geqslant \max(n_1^{-\frac{1}{2}+\alpha}, n_2^{-\frac{1}{2}+\alpha})$, $n_1 p^2 \geqslant 16$, and n_2 is sufficiently large, then Multi-Rank with $k = 1$ and $\beta = \frac{p^2 n_1}{2}$ outputs $\widehat{\sigma} \in \mathcal{S}^{n_1 \times n_2}$ such that*

$$\begin{aligned}
&\Pr_{\{\boldsymbol{x}_i\}, \{\boldsymbol{y}_u\}, \boldsymbol{\Omega}}(\mathrm{dis}_{2\epsilon}(\widehat{\sigma}, H) > 0) \\
\leqslant & n_2 \binom{n_1}{2} [2\exp(-\frac{(n_2-1)p^2}{12}) + (n_2 - 1)\exp(-\frac{n_1 p^2}{8}) \\
&+ \exp(-(\frac{(n_2-1)p^2}{2})\tau(\eta)) \\
&+ 3(n_2 - 1)p^2 \exp(-\frac{\delta^2 n_1 p^2}{20})].
\end{aligned}$$

A couple of remarks are in order. First, if ϵ and δ are small, then η must be correspondingly small. η represents how close a user y_v must be to a user y_u in the feature space to guarantee that the ratings of y_v can be used to make inferences about the ranking of user y_u. Second, whereas we require that $p \geqslant n_2^{-\frac{1}{2}+\alpha}$, Lee et al. (2016) require that $p \geqslant n_2^{-1+\alpha}$. We conjecture that this stronger requirement is fundamental to our algorithm since $v \in W_u^{i,j}(\beta)$ only if v has rated both items i and j, which v does with probability p^2. However, there may be another algorithm that circumvents this issue. Theorem 1 implies the following Corollary.

Corollary 1. *Assume the setting of Theorem 1. If $n_2 \longrightarrow \infty$, $p \geqslant \max(n_1^{-\frac{1}{2}+\alpha}, n_2^{-\frac{1}{2}+\alpha})$, and $n_2^{C_1} \geqslant n_1 \geqslant C_2 \log(n_2)^{\frac{1}{2\alpha}}$ for any constant $C_1 > 0$ and some constant $C_2 > 0$ depending on α, then $\Pr_{\{\boldsymbol{x}_i\}, \{\boldsymbol{y}_u\}, \Omega}(\mathrm{dis}_{2\epsilon}(\widehat{\sigma}, H) > 0) \longrightarrow 0$ as $n_2 \longrightarrow \infty$.*

Note that the growth rates of n_1, n_2 and p imply that the average number of rated items by each user pn_1 must grow as $C \log(n_2)^{\frac{1}{2}+\frac{1}{4\alpha}}$ for some universal constant $C > 0$.

Next, we sketch the proof. The main part of the analysis deals with establishing a probability bound of a

mistake by Pairwise-Rank for a specific user u and a pair of items i and j when $|f(\boldsymbol{x}_i, \boldsymbol{y}_u) - f(\boldsymbol{x}_j, \boldsymbol{y}_u)| > \epsilon$. First, we establish that w.h.p. $|W_u^{i,j}(\beta)|$ is large, i.e., there are many users that have rated i and j and many other items in common with u. Second, using standard concentration bounds, it is shown that for every $v \in W_u^{i,j}(\beta)$, $R_{u,v}$ concentrates around $\rho(u,v)$. Since $\beta \longrightarrow \infty$, this estimate converges to $\rho(u,v)$. Third, we show that eventually we sample a point from $B_\eta(\boldsymbol{y}_u)$. Further, if $\boldsymbol{y}_v \in B_\eta(\boldsymbol{y}_u)$ and $\boldsymbol{y}_w \in B_{\frac{\epsilon}{2}}(\boldsymbol{y}_u)^c$ (note $\eta \leqslant \frac{\epsilon}{2}$), then since $\boldsymbol{y}_u$ is $(\frac{\epsilon}{2}, \delta)$-discriminative w.p. 1, by our choice of η, $\rho(\boldsymbol{y}_u, \boldsymbol{y}_v) > \rho(\boldsymbol{y}_u, \boldsymbol{y}_w) + \frac{\delta}{2}$. Thus, by concentration bounds, $R_{u,v} > R_{u,w}$. Therefore, Pairwise-Rank with $k = 1$ uses the preference ordering of a user in $B_{\frac{\epsilon}{2}}(\boldsymbol{y}_u)$ on items i and j to make the prediction. The Lipschitzness of f and our assumption that g_v is strictly increasing imply that this prediction is correct. It is possible to extend this argument to handle the case when $k > 1$.

5.2 Discrete Ratings Setting

Let $N > 0$ and suppose that $|f(x,y)| \leqslant N \ \forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$. Suppose that there are L distinct ratings and let $\mathcal{G}$ denote the set of all step functions of the form

$$g_u(x) = \begin{cases} 1 & : x \in [-N, a_{u,1}) \\ 2 & : x \in [a_{u,1}, a_{u,2}) \\ \vdots & \\ L & : x \in [a_{u,L-1}, N] \end{cases}.$$

We assume that for all $u \in [n_2]$, $g_u \in \mathcal{G}$ and that the rating thresholds are random, i.e., $(\boldsymbol{a}_{1,1}, \ldots, \boldsymbol{a}_{1,L-1}), \ldots, (\boldsymbol{a}_{n_2,1}, \ldots, \boldsymbol{a}_{n_2,L-1}) \overset{i.i.d.}{\sim} \mathcal{P}_{[-N,N]^{L-1}}$. We write $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{n_2} \overset{i.i.d.}{\sim} \mathcal{P}_{\mathcal{G}}$ and we assume that $\{\boldsymbol{g}_u\}_{u \in [n_2]}$ is independent from $\{\boldsymbol{x}_i\}_{i \in [n_1]}$, $\{\boldsymbol{y}_u\}_{u \in [n_2]}$, and $\boldsymbol{\Omega}$. Let $\mathcal{P}_l$ denote the marginal distribution of $\boldsymbol{a}_{u,l}$ for all $u \in [n_2]$. We make the following assumption.

Definition 3. *We say that $\mathcal{P}_{\mathcal{G}}$ is* diverse *if for every open interval $I \subset [-N, N]$ there exists l such that $\mathcal{P}_l(I) > 0$.*

Let $d_{\mathbb{R}}$ denote a metric on $\mathbb{R}$; fix $u \in [n_2]$ and let $\gamma(\epsilon) = \inf_{z \in [-N,N]} \mathcal{P}_{\{\boldsymbol{a}_{u,l}\}_{l \in [L-1]}}(\exists l \in [L-1] : d_{\mathbb{R}}(z, \boldsymbol{a}_{u,l}) \leqslant \epsilon)$. The aforementioned assumption ensures via a measure theoretic argument that $\gamma(\epsilon) > 0$ for all $\epsilon > 0$ (see Lemma D.3 in the Appendix).

Theorem 2. *Let $\epsilon, \delta > 0$ and $\eta \in (0, \frac{\epsilon}{4})$. Suppose that $\mathcal{P}_{\mathcal{G}}$ is diverse and that almost every $y \in \mathcal{Y}$ is $(\frac{\epsilon}{4}, \delta)$-discriminative. Let r be a positive nondecreasing function such that $r(\frac{\epsilon}{4}) \geqslant \delta$ and $r(\eta) < \frac{\delta}{2}$. Suppose that almost every $y \in \mathcal{Y}$ is r-discerning. Let $\frac{1}{2} > \alpha > \alpha' > 0$. If $p \geqslant \max(n_1^{-\frac{1}{2}+\alpha}, n_2^{-\frac{1}{2}+\alpha})$, $n_1 p^2 \geqslant 16$, $n_1 \geqslant C_1 \log(n_2)^{\frac{1}{2\alpha}}$ for some constant C_1, and n_2 is sufficiently large, Multi-Rank with $k = n_2^{\alpha'}$ and $\beta = \frac{p^2 n_1}{2}$ outputs $\widehat{\sigma}$ such that*

$$\begin{aligned} &\Pr_{\{\boldsymbol{x}_i\},\{\boldsymbol{y}_u\},\{\boldsymbol{a}_{u,l}\},\boldsymbol{\Omega}}(\mathrm{dis}_{2\epsilon}(\widehat{\sigma}, H) > 0) \\ \leqslant & n_2 \binom{n_1}{2} [2\exp(-\frac{(n_2-1)p^2}{12}) + (n_2-1)\exp(-\frac{n_1 p^2}{8}) \\ & + 2\exp(-\gamma(\frac{\epsilon}{4}))k) \\ & + \frac{1}{1 - r(\frac{\epsilon}{2})}[3(n_2-1)p^2 \exp(-\frac{\delta^2 n_1 p^2}{20}) \\ & + \exp([1 - \kappa(\frac{\epsilon}{4}) + \tau(\eta) + \log(3\frac{(n_2-1)p^2}{2})]k \\ & - k\log(k) - \tau(\eta)\frac{(n_2-1)p^2}{2})]]. \end{aligned}$$

Corollary 2. *Assume the setting of Theorem 2. If $p \geqslant \max(n_1^{-\frac{1}{2}+\alpha}, n_2^{-\frac{1}{2}+\alpha})$, $k = n_2^{\alpha'}$, and $n_2^{C_1} \geqslant n_1 \geqslant C_2 \log(n_2)^{\frac{1}{2\alpha}}$ for any constant $C_1 > 0$ and some constant $C_2 > 0$ depending on α, then $\Pr_{\{\boldsymbol{x}_i\},\{\boldsymbol{y}_u\},\Omega}(\mathrm{dis}_{2\epsilon}(\widehat{\sigma}, H) > 0) \longrightarrow 0$ as $n_2 \longrightarrow \infty$.*

The bulk of the analysis for the discrete ratings setting is similar to the continuous rating setting and, once again, mainly deals with the analysis of Pairwise-Rank for a user u and items i and j. Since the ratings are discrete, although users that are sufficiently close to user u in the feature space agree about the ordering of items i and j, we need to show that at least one of these neighbors does not give the same rating to items i and j. To this end, we show that eventually k nearby points are sampled: $\boldsymbol{y}_{v_1}, \ldots, \boldsymbol{y}_{v_k} \in B_\eta(\boldsymbol{y}_u)$. Conditional on $|f(\boldsymbol{x}_i, \boldsymbol{y}_u) - f(\boldsymbol{x}_j, \boldsymbol{y}_u)| > \epsilon$, using the Lipschitzness of f, $(f(\boldsymbol{x}_i, \boldsymbol{y}_{v_q}), f(\boldsymbol{x}_j, \boldsymbol{y}_{v_q}))$ has length at least $\frac{\epsilon}{2}$. Finally, since $\mathcal{P}_{\mathcal{G}}$ is diverse, a concentration argument wrt $\boldsymbol{g}_{v_1}, \ldots, \boldsymbol{g}_{v_k}$ implies that w.h.p. there exists $q \in [k]$ and $l \in [L-1]$ such that $\boldsymbol{a}_{v_q,l} \in (f(\boldsymbol{x}_i, \boldsymbol{y}_{v_q}), f(\boldsymbol{x}_j, \boldsymbol{y}_{v_q}))$. Thus, user v_q provides distinct ratings for items i and j.

6 A Necessary and Sufficient Condition for $\mathrm{dis}_\epsilon(\sigma, H) = 0$

In this section, we characterize the class of optimal collections of rankings, i.e., $\sigma \in \mathcal{S}^{n_1 \times n_2}$ such that $\mathrm{dis}_\epsilon(\sigma, H) = 0$. We show roughly that a collection of rankings σ is optimal in the sense that $\mathrm{dis}_\epsilon(\sigma, H) = 0$ if and only if σ agrees with the observed data and σ gives the same ranking to users that are close to each other in the latent space $\mathcal{Y}$. To study this question, we consider the regime where the number of items n_1 is fixed, the probability of an entry being revealed p is fixed, and the number of users n_2 goes to infinity.

Consider the following notion, which is the main ingredient in our necessary and sufficient condition:

Definition 4. *Let $\epsilon > 0$ and $T \subset [n_1] \times [n_1] \times [n_2]$. $\sigma \in \mathcal{S}^{n_1 \times n_2}$ is an ϵ-consistent collection of rankings over T if $\forall i \neq j \in [n_1], u \neq v \in [n_2]$ such that $(i,j,u),(i,j,v) \in T$ and $d_{\mathcal{Y}}(y_u, y_v) \leqslant \epsilon$, it holds that $\sigma(i,u) < \sigma(j,u) \iff \sigma(i,v) < \sigma(j,v)$. If σ is an ϵ-consistent collection of rankings over $[n_1]\times[n_1]\times[n_2]$, then we simply say that σ is an ϵ-consistent collection of rankings.*

In words, a collection of rankings is ϵ-consistent if it gives the same ranking to users that are within ϵ of each other in the latent space.

We introduce the following objective function: $\widehat{\text{dis}}(\sigma, H) :=$

$$\sum_{u=1}^{n_2} \sum_{i<j:(i,u),(j,u)\in\Omega} \mathbf{1}\{(h_u(x_i, y_u) - h_u(x_j, y_u)) \times (\sigma(i,u) - \sigma(j,u)) < 0\}.$$

Once again, we analyze separately the continuous rating and discrete rating settings. With respect to the continuous rating setting, Theorems 3 and 5 *roughly* imply that with probability tending to 1 as $n_2 \longrightarrow \infty$, a collection of rankings $\sigma \in \mathcal{S}^{n_1 \times n_2}$ that minimizes $\widehat{\text{dis}}(\cdot, H)$ is $\frac{\epsilon}{2}$-consistent if and only if $\text{dis}_\epsilon(\sigma, H) = 0$. A similar statement holds for the discrete rating setting.

To begin, we present our sufficient conditions.

Theorem 3. *Assume the continuous rating setting. Let $\epsilon > 0$ and suppose that for all $u \in [n_2]$, $g_u(\cdot)$ is strictly increasing. With probability increasing to 1 as $n_2 \longrightarrow \infty$, if $\sigma \in \mathcal{S}^{n_1 \times n_2}$ is $\frac{\epsilon}{2}$-consistent and minimizes $\widehat{\text{dis}}(\cdot, H)$, then $\text{dis}_\epsilon(\sigma, H) = 0$.*

Theorem 4. *Assume the discrete rating setting and that $\mathcal{P}_{\mathcal{G}}$ is diverse. Let $\epsilon > 0$. With probability increasing to 1 as $n_2 \longrightarrow \infty$, if $\sigma \in \mathcal{S}^{n_1 \times n_2}$ is $\frac{\epsilon}{8}$-consistent and minimizes $\widehat{\text{dis}}(\cdot, H)$, then $\text{dis}_\epsilon(\sigma, H) = 0$.*

The proofs for the continuous and discrete cases are similar. We briefly sketch the argument for the continuous case. Since $\mathcal{Y}$ is compact, there is a finite subcover of $\mathcal{Y}$ with open balls with diameter at most $\frac{\epsilon}{2}$. As $n_2 \longrightarrow \infty$, with probability increasing to 1, for every open ball $\mathcal{O}$ belonging to the finite subcover and for every pair of distinct items $i, j \in [n_1]$, there is some user $u \in \mathcal{O}$ that has rated i and j. Then, on this event, it can be shown that if $\text{dis}_\epsilon(\sigma, H) > 0$, then $\widehat{\text{dis}}(\sigma, H) > 0$. Thus, using the contrapositive, the result follows.

Theorem 5 gives our necessary condition.

Theorem 5. *Let $\epsilon > 0$ and $\sigma \in \mathcal{S}^{n_1 \times n_2}$ such that $\text{dis}_\epsilon(\sigma, H) = 0$. Let $T = \{(i,j,u) \in [n_1] \times [n_1] \times [n_2] : |f(x_i, y_u) - f(x_j, y_u)| > \epsilon,\ h(x_i, y_u) \neq h(x_j, y_u)\}$. Then, σ is an ϵ-consistent collection of rankings over T.*

Theorem 5 shows that in our general setting, learning the correct collection of rankings requires giving the same ranking to nearby users. In particular, this provides an intuition on the kind of embedding that matrix factorization learns. Theorem 5 only applies to items i, j and user u if there is a large enough difference in the underlying values given by f. The proof follows by the Lipschitzness of f and algebra.

7 Experiments

In this section, we examine the empirical performance of Multi-Rank. It is well-known that matrix factorization methods tend to outperform neighborhood-based methods. Nevertheless, neighborhood-based methods remain popular in situations where practitioners want an easy-to-implement method, to avoid expensive model-building, and to be able to interpret predictions easily (Ning et al., 2011). Furthermore, it has been observed that for the task of matrix completion, (i) matrix factorization methods and neighborhood-based methods have complementary strengths and weaknesses and (ii) performance gains can be achieved by merging these methods into a single algorithm (Bell and Koren, 2007; Koren, 2008). Yet, it is non-trivial to generalize ideas for combining matrix factorization and neighborhood-based methods in the matrix completion setting to the preference completion setting. In light of this discussion, the purpose of our experiments is not to demonstrate the superiority of our method over matrix factorization methods, but to compare the performance of our algorithm with the state-of-the-art.

We compared the performance of our algorithm (MR) and a weighted version of our algorithm (MRW) where votes are weighted by $R_{u,v}$ against Alternating SVM (AltSVM) (Park et al., 2015), Retargeted Matrix Completion (RMC) (Gunasekar et al., 2016), and the proposed algorithm in (Lee et al., 2016) (LA). We chose AltSVM and RMC because they are state-of-the-art matrix factorization methods for preference completion and we chose LA because its theoretical guarantees are similar to our guarantees for Multi-Rank and it was shown to be superior to item-based and user-based neighborhood methods (Lee et al., 2016). We used grid search to optimize the hyperparameters for each of the algorithms using a validation set.

We use the ranking metrics Kendall Tau, Spearman Rho, NDCG@5, and Precision@5. Kendall Tau and Spearman Rho measure how correlated the predicted ranking is with the true ranking. The other metrics measure the quality of the predicted ranking at the

Dataset	Method	Kendall Tau	Spearman Rho	NDCG@5	Precision@5
	MRW	0.3156 (0.0021)	0.4012 (0.0029)	0.7104 (0.0010)	0.4492 (0.0018)
	MR	0.3105 (0.0021)	0.3963 (0.0030)	0.7063 (0.0038)	0.4457 (0.0044)
	LA	0.3271 (0.0018)	0.4153 (0.0022)	0.7136 (0.0026)	0.4570 (0.0041)
	AltSVM	0.3271 (0.0007)	0.4173 (0.0008)	0.7022 (0.0015)	0.4365 (0.0036)
Netflix	RMC	0.3288 (0.0017)	0.4178 (0.0020)	0.7204 (0.0006)	0.4581 (0.0048)
	MRW	0.3933 (0.0010)	0.5009 (0.0013)	0.7769 (0.0066)	0.6083 (0.0096)
	MR	0.3924 (0.0011)	0.4999 (0.0013)	0.7735 (0.0061)	0.6021 (0.0063)
	LA	0.3993 (0.0009)	0.5075 (0.0012)	0.7767 (0.0058)	0.6071 (0.0080)
	AltSVM	0.4099 (0.0008)	0.5219 (0.0010)	0.8002 (0.0042)	0.6417 (0.0067)
Movielens	RMC	0.4041 (0.0004)	0.5139 (0.0006)	0.8068 (0.0030)	0.6485 (0.0029)

Table 1: Netflix and MovieLens Results. On the Netflix dataset, MR usually used $\beta = 5$ and $k \in [13, 19]$. MRW usually used $\beta = 9$ and $k \in [16, 23]$. On the MovieLens dataset, MR usually used $\beta = 10$ and $k \in [7, 13]$. MRW usually used $\beta = 12$ and $k \in [13, 17]$.

Dataset	Method	Kendall Tau	Spearman Rho	NDCG@5	Precision@5
	MRW	0.2736 (0.0017)	0.3327 (0.0021)	0.8063 (0.0031)	0.7849 (0.0034)
	MR	0.2677 (0.0019)	0.3255 (0.0023)	0.7980 (0.0034)	0.7764 (0.0008)
	LA	0.2786 (0.0022)	0.3387 (0.0027)	0.8024 (0.0024)	0.7843 (0.0018)
	AltSVM	0.2743 (0.0015)	0.3335 (0.0018)	0.7949 (0.0023)	0.7768 (0.0023)
Netflix	RMC	0.2856 (0.0017)	0.3473 (0.0021)	0.8052 (0.0038)	0.7861 (0.0032)
	MRW	0.3347 (0.0015)	0.4090 (0.0018)	0.8903 (0.0059)	0.8810 (0.0059)
	MR	0.3343 (0.0017)	0.4085 (0.0021)	0.8879 (0.0052)	0.8792 (0.0061)
	LA	0.3395 (0.0017)	0.4149 (0.0020)	0.8908 (0.0085)	0.8845 (0.0078)
	AltSVM	0.3451 (0.0016)	0.4217 (0.0020)	0.9070 (0.0056)	0.8982 (0.0056)
Movielens	RMC	0.3504 (0.0014)	0.4281 (0.0017)	0.9140 (0.0026)	0.9051 (0.0032)

Table 2: Quantized Netflix and MovieLens Results. On the Netflix dataset, MR usually used $\beta = 5$ and $k = 22$. MRW usually used $\beta \in [9, 10]$ and $k \in [27, 31]$. On the MovieLens dataset, MR usually used $\beta \in [10, 13]$ and $k \in [10, 19]$. MRW usually used $\beta \in [8, 11]$ and $k \in [16, 23]$.

Dataset	Method	Kendall Tau	Spearman Rho	NDCG@5	Precision@5
Netflix	LA	0.1798 (0.0034)	0.2300 (0.0040)	0.5962 (0.0022)	0.3322 (0.0053)
MovieLens	LA	0.2404 (0.0098)	0.3092 (0.0123)	0.6543 (0.0138)	0.4435 (0.0163)

Table 3: Monotonically Transformed Netflix and MovieLens Results. We only display the results for LA since the other methods are invariant to monotonic transformations of the columns.

top of the list. For Precision@5, we deem an item relevant if it has a score of 5. For all of these metrics, higher scores are better. See Liu (2009) for a more detailed discussion of these metrics. The numbers in parentheses are standard deviations.

We use the Netflix and MovieLens 1M datasets. We pre-process the data in a similar way to Liu and Yang (2008). For the Netflix dataset, we take the 2000 most popular movies and randomly selected 4000 users that had rated at least 100 of these movies. For both datasets, we randomly subsample the ratings 5 times in the following way: we randomly shuffled the (user-id, movie, rating) triples and split 40% into a training set, 15% into a validation set, and 45% into a test set. For the Netflix dataset, we drop users if they have fewer than 50 ratings in the training set and fewer than 10 ratings in either the validation set or the test set. For the MovieLens dataset, we drop users if they have fewer than 100 ratings in the training set and fewer than 50 ratings in either the validation set or the test set. Table 1 shows that although MRW does not have the best performance, it outperforms AltSVM on NDCG@5 and Precision@5 on the Netflix dataset and LA on NDCG@5 and Precision@5 on the MovieLens dataset.

In addition, we quantized the scores of both datasets to 1 if the true rating is less than or equal to 3 and to 5 otherwise (see Table 2). Here, MR and MRW have the same amount of information as LA and RMC. On the Netflix dataset, MRW performed the best on the NDCG@5 measure.

Finally, we considered a setting where a company performs A/B testing on various rating scales (e.g., 1-5, 1-10, 1-50, 1-100) and wishes to use all of the collected data to predict preferences. To model this situation, for each user, we randomly sampled a number $a \in \{1, 2, 10, 20\}$ and $b \in [a-1] \cup \{0\}$, and transformed the rating $r \mapsto a \cdot r - b$. Table 3 shows that on the monotonically transformed versions of the Netflix and MovieLens datasets, LA performs dramatically worse. This is unsurprising since it is well-known that the performance of rating-oriented neighborhood-based methods like LA suffers when there is rating scale variance (Ning et al., 2011).

Acknowledgements

This work was supported in part by NSF grant 1422157.

References

N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *Journal of the ACM*, 2008.

R. Bell and Y. Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, pages 75–79, 2007.

J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. *In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, 1998.

A. H. Copeland. A reasonable social welfare function. *Seminar on Mathematics in Social Sciences, University of Michigan*, 1951.

D. Coppersmith, L. Fleischer, and A. Rudra. Ordering by weighted number of wins gives a good ranking for weighted tournaments. *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2006.

P. Cremonesi, Y. Koren, and R. Turrin. Proceedings of the fourth acm conference on recommender systems. *Performance of recommender algorithms on top-n recommendation tasks*, 2010.

A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. *Proceedings of the 16th international conference on World Wide Web*, pages 271–280, 2007.

R. S. Ganti, L. Balzano, and R. Willett. Matrix completion under monotonic single index models. *Advances in Neural Information Processing Systems*, pages 1873–1881, 2015.

S. Gunasekar, O. Koyejo, and J. Ghosh. Preference completion from partial rankings. *Advances in Neural Information Processing Systems*, 2016.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. *Data Mining*, pages 263–272, 2008.

R. H. Keshavan, A. Montanai, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 2010.

J. Kleinberg and M. Sandler. Convergent algorithms for collaborative filtering. *Proceedings of the 4th ACM conference on Electronic commerce*, 2003.

J. Kleinberg and M. Sandler. Using mixture models for collaborative filtering. *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 2004.

Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, 2008.

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

C. E. Lee, Y. Li, D. Shah, and D. Song. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. *Advances in Neural Information Processing Systems*, 2016.

N. N. Liu and Q. Yang. Eigenrank: A ranking-oriented approach to collaborative filtering. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, page 8390, 2008.

T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, pages 225–331, 2009.

Y. Lu and S. Negahban. Individualized rank aggregation using nuclear norm regularization. *Technical Report, Department of Statistics Yale University*, 2014.

X. Ning, C. Desrosiers, and G. Karypis. *A comprehensive survey of neighborhood-based recommendation methods.* 2011.

S. Oh, K. Thekumparampil, and J. Xu. Collaboratively learning preferences from ordinal data. *In Advances in Neural Information Processing Systems*, pages 1909–1917, 2015.

D. Park, J. Neeman, J. Zhang, S. Sanghavi, and I. Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

J.-F. Pessiot, T.-V. Truong, N. Usunier, M.-R. Amini, and P. Gallinari. Learning to rank for collaborative filtering. *ICEIS*, 2007.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 2011.

S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr:bayesian personalized ranking from implicit feedback. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461, 2009.

S. Wang, J. Sun, B. Gao, and J. Ma. Vsrank: A novel framework for ranking-based collaborative filtering. *ACM Transactions on Intelligent Systems and Technology*, 2014.

S. Wang, S. Huang, T.-Y. Liu, J. Ma, Z. Chen, and J. Veijalainen. Ranking-oriented collaborative filtering: A listwise approach. *ACM Transactions on Information Systems*, 2016.

M. Weimer, A. Karatzoglou, Q. Viet Le, and A. Smola. Maximum margin matrix factorization for collaborative ranking. *Advances in neural information processing systems*, pages 1–8, 2007.

Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. *In International Conference on Algorithmic Applications in Management*, 2008.