Non-Compensatory Psychological Models for Recommender Systems

ID: 335

Abstract

The study of consumer psychology reveals two categories of consumption decision procedures: compensatory rules and non-compensatory rules. Existing recommendation models which are based on latent factor models assume the consumers follow the compensatory rules, i.e. the consumer evaluate an item over multiple aspects and compute a weighted or/and summated score which is further used to derive the rating or the rankings among items. However, it has been shown in the literature of consumer psychology that, consumers adopt non-compensatory rules more than compensatory rules. Our main contribution in this paper is to study the unexplored utilization of non-compensatory rules in recommendation models.

Our general assumptions are (1) there are K universal hidden aspects. In each evaluation session, only one aspect is chosen as the prominent aspect according to user preference. (2) Evaluations over prominent and non-prominent aspects are non-compensatory. Evaluation is manly based on item performance on the prominent aspect. For non-prominent aspects the user sets a minimal acceptable value. We give a conceptual model for these general assumptions and show how this model can be applied to a wide range of existing recommender systems, including point-wise rating prediction models and pair-wise ranking prediction models. We experimentally show that adopting non-compensatory rules constantly improve ranking performance of existing models on a variety of real-world recommendation data sets.

Introduction

The majority of state-of-the-art recommendation models are based on latent factor models. Generally, latent factor models transform both user preferences and item features into the same hidden feature spaces with K aspects. To recover the observations (i.e. ratings or rankings) in any recommender system, they adopt the inner product of the user preferences and the item features. There are fruitful successful applications of latent factor models in rating predictions (Koren, Bell, and Volinsky 2009; Koren 2010; Lee et al. 2014) and ranking reconstructions (Rendle et al. 2009; Steck 2015; Zhao et al. 2018; Shi, Larson, and Hanjalic 2010).

From the perspective of consumer decision making, all existing latent factor models fall into the category of *compensatory rules*. Consumers who adopt compensatory rules evaluate every item over multiple aspects and compute a weighted or/and summated score for each item. Then they will rate or rank items based on the score. The key property of compensatory rules is that a good performance on one aspect of an item compensates for poor performances on other aspects.

However, in the study of human choice behavior, it is well regarded that there are two categories of decision making procedures, namely *compensatory rules* and *non-compensatory rules* (Engel, Blackwell, and Miniard 1986) . Furthermore, it is found in many surveys that consumers more frequently make consumption related choices based on non-compensatory rules. For example, (Hauser, Ding, and Gaskin 2009) reviews 132 empirical studies in literature and concludes that more than 70% of consumers adopt non-compensatory rules when buying air-conditioners, automobiles, computers, cameras and so on.

Non-compensatory rules do not allow the shortcomings of a product to be balanced out by its attractive features. The literature has proposed different non-compensatory rules, among which *lexicographic rule* and *conjunctive rule* are the most common. For example, in a survey interviewing consumption decisions about beer brands and fast-food outlets (Laroche, Kim, and Matsui 2003), conjunctive rule has a success rate of 62.0% in predicting brand consideration and lexicographic rule has a success rate of 34.6% which is the second highest non-compensatory rule. We next illustrate *lexicographic rule* and *conjunctive rule* by a toy example.

Example. Alice wants to buy a smart phone and she ranks her alternatives over three relevant aspects: battery life, price and storage space. *Lexicographic rule* assumes that aspects of products can be ordered in terms of importance and alternative brands are evaluated sequentially from most prominent to least prominent aspects. If Alice's priority is long-lasting battery, then she will adopt lexicographic rule, to rank phones first based on battery life. Clearly Honor and iPhone will be ranked higher than Galaxy, the other benefits offered by Galaxy do not outweigh her desire for a long-life battery . *Conjunctive rule* establishes a minimally acceptable threshold for each aspect and evaluation is made on the basis of whether or not the products satisfy the threshold. If Alice wants the phone to be cheap and with plenty

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

of storage space, then she will adopt conjunctive rule, to set thresholds (e.g. 600$ and $64GB$ on the corresponding aspects). iPhone fails to meet the cut-off point, it will not outrank Honor which satisfies the minimal acceptable value on each aspect. In either case, adopting a compensatory rule based recommendation model is problematic.

Table 1: Illustrative example of non-compensatory rules

Item	Prominent aspect	Not prominent aspects	
	Battery life	Price	Memory
iPhone SE	13 hours	700$	64GB
Galaxy S8	9 hours	500$	128GB
Honor 10	24 hours	589$	128GB

Non-compensatory rules are widely used in many **D**ecision **S**upport **S**ystems (DSS) (Lee 2009). Despite of the commercial success, current computer support for non-compensatory rules is labor costly and user unfriendly, i.e. they ask consumers to control or manipulate the rules, e.g. specify the value of an aspect. To the best of our knowledge, no previous effort has been devoted to modeling and learning non-compensatory rules in recommender systems.

Our goal in this paper is to study this unexplored area. Two challenges need to be addressed. (1) There are different non-compensatory rules. Incorporating all of them is neither practical nor efficient. We must embed the most typical non-compensatory rules in one unified framework. (2) Most non-compensatory rules explain discrete choices which are computationally inconvenient. It is important to model non-compensatory rules in a manner that complies with existing efficient rating and ranking models.

Our primary contribution is to give a conceptual model of how users adopt non-compensatory rules in recommender systems. Our assumptions are based on the lexicographic and conjunction rules. We assume that, (1) there are $K-$ hidden aspects which user preferences and item features are transformed into, (2) in each evaluation session, the user picks a prominent aspect according to his/her preference, (3) the user adopts different evaluation strategies on prominent and non-prominent aspects. The evaluation is mainly based on item performance on the prominent aspect. The evaluation is less influenced by item performance with respect to a user-defined minimal acceptance value on non-prominent aspects.

Our second contribution is to realize the conceptual model in a wide range of recommendation frameworks, including point-wise rating prediction models such as the conventional Matrix Factorization (MF (Koren, Bell, and Volinsky 2009)), Matrix Factorization with neighborhood collaborative filtering (AMF (Koren 2008)), and locally low-rank matrix approximation (LLORMA (Lee et al. 2013)) and pairwise ranking reconstruction models such as BT model (Hu and Li 2016) and BPR style Thurstonian model (Rendle et al. 2009).

We conduct comprehensive experiments on a variety of real world data sets. We experimentally show that the non-compensatory versions of these models significantly improve ranking performances of the original models.

The paper is organized as follows. We start with surveying the most commonly adopted latent factor models in the community of recommendation research. We show that previous research work are compensatory models which are based on different rating prediction formulas and ranking models. Next, we describe our non-compensatory assumptions and develop non-compensatory versions of existing models. Then we experimentally show that the non-compensatory versions outperform the original versions of existing models on a variety of real-world data sets. Finally we conclude our work and future directions.

Compensatory Recommendation Models

In this section, we show that existing latent factor models are based on compensatory rules. We summarize and categorize related work based on the forms of rating prediction formulas and loss functions. We restrict our discussions to latent factor models, i.e. models where a universe of K factors is used to project user preferences and item features. Hereafter, unless stated otherwise, we use lower-case letters for indices, upper-case letters for universal constants, lower-case bold-face letters for vectors and upper-case bold-face letters for matrices. Specifically, $\mathbf{X} \in \mathcal{R}^{M\times N}$ denotes the rating matrix, $\hat{\mathbf{X}} \in \mathcal{R}^{M\times N}$ denotes the predicted rating matrix, $\mathbf{p}, \mathbf{q} \in \mathcal{R}^K$ denotes the item features, which are rows of item space $\mathbf{V} \in \mathcal{R}^{N\times K}$, $\mathbf{u} \in \mathcal{R}^K$ denotes the user preferences, which is a row of the user space $\mathbf{U} \in \mathcal{R}^{M\times K}$. $\mathbf{U}, \mathbf{V}$ are components of the model parameters $\Theta = \{\mathbf{U}, \mathbf{V}\}$.

Rating Prediction Formulas

One goal of recommendation research is to recover the rating matrix $\mathbf{X}$, by minimizing a loss function $\mathcal{L}(\Theta)$, which is usually defined as the regularized square loss between the predicted rating $\hat{\mathbf{X}}_{u,q}$ and the observed rating $\hat{\mathbf{X}}_{u,q}$ for each user u who has rated item q.

$$\mathcal{L}(\Theta) = \sum_{u,q}(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})^2 + \lambda(\|\mathbf{U}\| + \|\mathbf{V}\|) \quad (1)$$

We list some of the most successful rating prediction formulas for $\hat{\mathbf{X}}$.

Matrix Factorization. In conventional Matrix Factorization (MF) (Koren, Bell, and Volinsky 2009), the predicted rating can be computed as an inner product of user preferences and item features as follows.

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \mathbf{q}_k \mathbf{u}_k \quad (2)$$

For simplicity we ignore the user specific or item specific bias (Koren, Bell, and Volinsky 2009). A massive amount of techniques have been proposed based on Equ. 2. Most of them modified the loss function ,e.g. by incorporating prior distributions over $\mathbf{p}, \mathbf{u}$ (Salakhutdinov and Mnih 2008), adding priors over unknown values (Devooght, Kourtellis, and Mantrach 2015), weighing different samples (Pilászy, Zibriczky, and Tikk 2010) and so on.

Neighborhood Factorization. in traditional memory based collaborative filtering strategies, neighborhood information has been proved to be useful. It is possible to embed such neighborhood information in latent factor models. Instead of directly modeling user preferences $\mathbf{u}$, each user is represented by items that he/she gives explicit or implicit feedback. For example, if we consider explicit feedback only, then each item is associated with two types of vectors $\mathbf{p}, \mathbf{q}$, the rating prediction formula of Asymmetric Matrix Factorization (AMF) in (Koren 2008) is stated as follows.

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \mathbf{q}_k (\sum_{p \in R(u)} \mathbf{p}_k / \sqrt{|R(u)|}), \tag{3}$$

where $R(u)$ is the set of rated items for u. AMF has been extended to SVD++ (Koren 2008) with implicit feedback.

Local Low-Rank Matrix Approximation. The third type of rating prediction formula is Local Low-Rank Matrix Approximation (LLORMA) (Lee et al. 2013). The intuition is that the entire rating matrix $\mathbf{X}$ is not low-rank but a sub-matrix restricted to a neighborhood of similar users and items is low-rank. Therefore, the predicted rating is aggregated over S sub-matrices as follows:

$$\hat{\mathbf{X}}_{u,q} = \sum_{t=1}^{S} \sum_{k} \mathbf{u}_{t,k} \frac{K((\mathbf{u}_t, \mathbf{i}_t), (\mathbf{u}, \mathbf{q}))}{\sum_{s=1}^{S} K((\mathbf{u}_s, \mathbf{i}_s), (\mathbf{u}, \mathbf{q}))} \mathbf{q}_{t,k} \tag{4}$$

$\mathbf{u}_t, \mathbf{q}_t$ are the factorized user preferences and item features in the $t-$th sub-matrix, $\mathbf{i}_s, \mathbf{i}_t$ are anchor points in the corresponding matrix to locate a neighborhood for low-rank decomposition, $K(\cdot)$ is a smoothing kernel.

Ranking Models

Another goal of recommendation research is to reveal the observed rankings. We here consider pair-wise rankings $p \succ_u q$, where user u prefers item p over q. The pair-wise rankings can be generated from pre-processing the ratings, i.e. $\mathbf{X}_{u,p} > \mu, \mathbf{X}_{u,q} < \mu$ (Hu and Li 2017), or from explicit and implicit feedback, i.e. $\mathbf{X}_{u,p} \neq 0$ and $\mathbf{X}_{u,q}$ doesn't exist (Rendle et al. 2009).

A large body of previous research has been presented by employing a ranking aware loss function $\mathcal{L}(\Theta)$ on the observed pair-wise rankings. For example, the Bayesian posterior is expressed as:

$$\mathcal{L}(\Theta) = \sum_{u} \sum_{p,q} o(p \succ_u q) \log p(p \succ_u q) + \lambda(\|\mathbf{U}\| + \|\mathbf{V}\|). \tag{5}$$

where $p(p \succ_u q)$ is the predicted possibility and $o(p \succ_u q)$ is an indicator function of whether or not the ranking is observed.

To generate the probability of pair-wise rankings $p(p \succ_u q)$, each user-item combination is associated with a score, i.e. $\hat{\mathbf{X}}_{u,p}, \hat{\mathbf{X}}_{u,q}$. We list two most commonly adopted ranking models .

Thurstone Model The most frequently adopted ranking model in recommendation systems is the Thurstone model (Thurstone 1927) which uses a non-linear transformation of the predicted ratings.

$$p(p \succ_u q) = \frac{1}{1 + \exp[-(\hat{\mathbf{X}}_{u,p} - \hat{\mathbf{X}}_{u,q})]} \tag{6}$$

Bradley-Terry Model. The famous Bradley-Terry (BT) model (Hunter 2004) is extensively studied in learning to rank scenarios. BT models the generation of ranking pairs by a division.

$$p(p \succ_u q) = \frac{\exp \hat{\mathbf{X}}_{u,p}}{\exp \hat{\mathbf{X}}_{u,p} + \exp \hat{\mathbf{X}}_{u,q}} \tag{7}$$

In either ranking model, the score $\hat{\mathbf{X}}$ can be approximated by different ranking prediction formulas. We categorize existing ranking aware methods based on the combination of rating prediction formulas and the ranking models.

- BT model has been leveraged with MF prediction formula in (Hu and Li 2016);
- Thurstone model with standard matrix factorization prediction formula is first presented as BPR (Rendle et al. 2009), which maximizes the Bayesian posterior with respect to Thurstonian modeling of standard matrix factorization predictions. Abundant research has been carried out to improve BPR-style systems by modifying the sampling methods in optimization, including BTR++ (Lerche and Jannach 2014), WARP (Weston, Bengio, and Usunier 2011), DNS (Zhang et al. 2013), RankMBPR (Yu et al. 2016) and so on.
- Thurstone model with neighborhood factorized prediction formula AMF is first incorporated in a point-wise ranking framework In (Steck 2015), FSBPR (Zhao et al. 2018) implants AMF in a Thurstone model and maximizes its likelihood.
- Thurstone model with local low-rank factorization prediction formula is utilized in LCR (Lee et al. 2014).

The list is by no means exclusive. However, we believe that most of existing recommender systems are covered. It is worthy to point out that (1) we do not restrict the form of loss functions. For example, many ranking approaches consider Bayesian maximum posterior, cross entropy and other forms of loss functions. Nevertheless, the core ranking model is either BT or Thurstone. (2) Although we only study pair-wise ranking , the conclusion is insightful for other ranking-aware systems, i.e. point-wise and list-wise approaches. The reason is that, as shown in (Steck 2015), point-wise and list-wise loss functions can be decomposed to components which are directly based on each score $\hat{\mathbf{X}}_{u,p}$ and components that are not related to $\hat{\mathbf{X}}$. Thus our proposed strategy in the next section is also applicable to point-wise and list-wise ranking models.

Non-Compensatory Recommendation Models

We begin this section by reviewing the findings in consumer psychology study. We proceed to present a general framework for modeling the psychological assumptions about

decision rules. We show the universality of the proposed framework by realizing it in different rating prediction formulas and ranking models.

Ever since the dawn of consumption psychology study, psychologists have been studying how consumers adopt different heuristics to facilitate brand (or other consumption related) choices. Two distinct categories of decision rules are found (Engel, Blackwell, and Miniard 1986): compensatory rules and non compensatory rules. The decision rules can be naturally explained in the latent factor models. For example, compensatory rules are adopted if a consumer determines options in terms of each factor and computes a weighted and/or summated score for each item, then selects the item that scores the highest among the alternatives evaluated. It is clear that all related work that has been described in previous section is the application of compensatory rules.

A number of non-compensatory rules are discovered in human decision process (Engel, Blackwell, and Miniard 1986). The most common rules include *lexicographic* and *conjunctive* rules (Laroche, Kim, and Matsui 2003). The conjunctive rules are often used in conjunction with lexicographic rules (Laroche, Kim, and Matsui 2003). In a lexicographic rule the consumer first ranks the aspects. He/she will pick the item with highest score on the most important aspect, breaking ties using successively less important aspects. In a conjunctive rule an item to be chosen must have all of its aspects above a user-defined threshold of minimally acceptable value.

We can see that non-compensatory rules differ from compensatory rules in two key points. (1) *Distinguished factors*. In compensatory rules, different factors are essentially equivalent, while in non-compensatory rules factors are not interchangeable. (2) *Distinguished evaluation metrics on each factor*. In compensatory rules, the evaluations on each factor follow the same framework (i.e. a product of user preference and item feature on the specific factor), while in non-compensatory rules, the evaluations on each factor are dissimilar.

For computational convenience, inspired by the psychological findings, we present the following conceptual model based on lexicographic and conjunction rules. We assume that in each evaluation session[1], there is a prominent aspect. The choice of the prominent aspect is dependent on the user preferences. Two types of evaluation strategies are adopted, one for the prominent aspect and the other for other non-prominent aspects. The overall evaluation of the item is mainly based on the its performance on the prominent aspect. The overall evaluation is less influenced by the item's performance on non-prominent aspects, compared with the user-defined aspect-specific threshold.

Non-Compensatory Rating Prediction Formulas

Our goal here is to modify the rating prediction formulas as little as possible, while still preserving the most important properties of non-compensatory rules. Therefore, we follow the same notations for user preferences and item features. In each evaluation session, the hidden prominent aspect is sampled by $\frac{\exp \mathbf{u}_k}{\sum_{k'} \mathbf{u}_{k'}}$. We use a parameter θ to control the strength of prominent aspect, i.e. the evaluation on the prominent aspect is magnified by $\exp \theta$. The threshold on aspect k set by user u is denoted as $\mathbf{b}_{u,k}$. When the aspect k is chosen, the evaluation of user u on q is $\exp \theta \mathbf{q}_k + \sum_{k' \neq k} (\mathbf{q}_{k'} - \mathbf{b}_{u,k'})$. The prediction is generated across all possible hidden prominent aspects. This gives us the following non-compensatory versions of rating prediction formulas.

Matrix Factorization: MF-NCR

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \frac{\exp \mathbf{u}_k}{\sum_{k'} \exp \mathbf{u}_{k'}} [\exp \theta \mathbf{q}_k + \sum_{k' \neq k} (\mathbf{q}_{k'} - \mathbf{b}_{u,k'})]. \tag{8}$$

Neighborhood Factorization: AMF-NCR implements a similar scheme by setting $u_k = \sum_{p \in R(u)} \mathbf{p}_k / \sqrt{|R(u)|}$,

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \frac{\exp(\sum_{p \in R(u)} \mathbf{p}_k)}{\sum_{k'} \exp(\sum_{p \in R(u)} \mathbf{p}_{k'})} [\exp \theta \mathbf{q}_k + \sum_{k' \neq k} (\mathbf{q}_{k'} - \mathbf{b}_{u,k'})]. \tag{9}$$

LLORMA-NCR uses the same decomposition for each sub-matrix.

$$\hat{\mathbf{X}}_{u,q} = \sum_{t=1}^{S} \sum_{k} \frac{\exp \mathbf{u}_k}{\sum_{k'} \exp \mathbf{u}_{k'}} \frac{K((\mathbf{u}_t, \mathbf{i}_t), (\mathbf{u}, \mathbf{q}))}{\sum_{s=1}^{S} K((\mathbf{u}_s, \mathbf{i}_s), (\mathbf{u}, \mathbf{q}))} [\exp \theta \mathbf{q}_{t,k} + \sum_{k' \neq k} (\mathbf{q}_{t,k'} - \mathbf{b}_{u,k'})] \tag{10}$$

We can see that all these NCR versions are combinations of lexicographic and conjunction rules, where $\exp \theta \to \infty$ indicates that the user adopts lexicographical rules only. The threashold for a user on an aspect is static in the sense that $\mathbf{b}_{u,k}$ does not change by the nature of the items.

Non-Compensatory Ranking Models

Thurston-NCR. The modification of Thurston model is straightforward, as the ranking probability involves a subtraction component of $\hat{\mathbf{X}}_{u,q}$ which can be replaced by any NCR-version of rating prediction formulas. Note that the user-defined aspect specific threshold $\mathbf{b}_{u,k}$ cancels between $\hat{\mathbf{X}}_{u,p}$ and $\hat{\mathbf{X}}_{u,q}$.

Inference of Thurston models is easily extensible. For example, if we use the Bayesian maximum posterior estimator as in BPR (Rendle et al. 2009), the loss function is defined as:

$$\mathcal{L} = -\sum_{u} \sum_{p \succ_u q} \ln \frac{1}{1 + \exp -[\hat{\mathbf{X}}_{u,p} - \hat{\mathbf{X}}_{u,q}]} - \lambda \|\Theta\|, \tag{11}$$

where Θ is the set of all parameters. Thus the inference procedure is accomplished by stochastic gradient descent (SGD) with $\frac{\partial \mathcal{L}}{\partial \Theta} = \sum_u \sum_{p \succ_u q} \frac{\partial \mathcal{L}}{\partial \Delta \hat{\mathbf{X}}_{u,p,q}} \frac{\partial \Delta \hat{\mathbf{X}}_{u,p,q}}{\partial \Theta}$, where $\Delta \hat{\mathbf{X}}_{u,p,q} = \hat{\mathbf{X}}_{u,p} - \hat{\mathbf{X}}_{u,q}$.

[1]The evaluation session could be either a true user interaction session with multiple actions, or a pseudo session which contains one rating action. The impact of availability of session information is discussed in experiments.

BT-NCR. Finally we propose the non-compensatory version of BT ranking model. In order to treat prominent and non-prominent aspects differently, we define the probability of any ranking pair $p \succ_u q$ as the product of results by factor-wise comparisons, based on a variant of BT model with ties (Hunter 2004). Again, in each evaluation session, a hidden prominent aspect k is sampled by user preference $\mathbf{u}$. The overall prediction is aggregated over all possible hidden prominent aspect k.

$$p(p \succ_u q) = \prod_{k=1}^{K} \mathbf{u}_k [\frac{\mathbf{p}_k}{\mathbf{p}_k + \theta \mathbf{q}_k} \prod_{k' \neq k} \frac{\theta \mathbf{p}_{k'}}{\mathbf{q}_{k'} + \theta \mathbf{p}_{k'}}]. \quad (12)$$

where $\mathbf{u}_k > 0, \sum_k \mathbf{u}_k = 1, \mathbf{p}, \mathbf{q} > \mathbf{0}$ and $\theta > 1$. BT-NCR models the non-compensatory rules in a manner that (1) the evaluation is mainly based on the prominent aspect. The item p is more likely to be preferred than q by user u if p is significantly better than q on the prominent aspect, i.e. $p_k > \theta q_k, \theta > 1$. (2) The performance on other aspects are less important. Because p is considered to be as good as q, as long $\forall k' \neq k, \theta p_{k'} > q_{k'}, \theta > 1$. BT-NCR is also a combination of lexicographic rules and conjunction rules. An interpretation is that we dynamically set a minimal acceptance value for $p_{k'}$ on factor $k' \neq k$ based on the compared alternative $q_{k'}$, where the minimal acceptance value is $q_{k'}/\theta$. The parameter θ controls the tolerance range. When $\theta \to \infty$, the users adopt lexicographic rules only.

To infer the parameters of BT-NCR, we implement a stochastic expectation maximization (SEM) algorithm. In each E-step, we first draw the value of prominent aspect k for each evaluation session by

$$k \sim u_k^t \frac{\mathbf{p}_k^t}{\mathbf{p}_k^t + \theta^t \mathbf{q}_k^t} \prod_{k' \neq k} [\frac{\theta^t \mathbf{p}_{k'}^t}{\mathbf{q}_{k'}^t + \theta^t \mathbf{p}_{k'}^t}]. \quad (13)$$

where t indicates the value obtained from the $t-$th round of SEM algorithm. In each M-step, we incorporate the MM bound in (Hunter 2004) and maximize the log-likelihood of complete data.

Experiments

We conduct experiments to evaluate the performance of non-compensatory rules in recommendation models. We conduct three sets of experiments on real world datasets. The first set of experiments is conducted to examine whether the NCR versions of rating prediction models outperform the original versions on rating data sets. The second set of experiments is conducted to examine whether NCR versions of ranking aware models outperform the original versions on data sets with explicit rating feedback. The third set of experiments is conducted to examine whether NCR versions of ranking aware models outperform the original versions on data sets with graded implicit feedback. We also analyze the inferred parameters $\theta, \mathbf{b}$ in NCR models for further insights.

Comparative Results for Rating Prediction Models

Data Sets We use the standard benchmarking datasets with user-item ratings. (1) Movielens[2]: user-movie rating for movies collected from the MovieLens web site (2) FilmTrust (Guo, Zhang, and Yorke-Smith 2013): user-movie ratings crawled from the entire FilmTrust website. (3) CiaoDVD (Guo et al. 2014): user-movie ratings crawled from the entire category of DVDs from the UK Ciao website. Statistics of the datasets are described in Table. 2.

For each dataset, we reserve users with at least 5 ratings and randomly split 80% of the ratings as training and 20% as test set. We avoid cold-start users and items. We consider each rating as an individual evaluation session. The ratings are normalized to the range of $[0, 1]$. The reported results are averaged using 5-fold cross validation,

Table 2: Statistics of Datasets with ratings

Dataset	#users	#items	#ratings	#pairs
Movielens	942	1,650	80,000	4,641,262
FilmTrust	1,235	2,062	35,497	623,516
CiaoDVD	2,665	14,280	72,665	2,478,836

Comparative Methods. We compare the non-compensatory improved versions (with suffix "-NCR") with the original versions on three widely adopted rating prediction methods (1) MF (Koren, Bell, and Volinsky 2009): standard matrix factorization in Equ. 2; (2) AMF (Koren 2008): neighborhood factorization in Equ. 3; (3) LLORMA (Lee et al. 2013): local low-rank matrix factorization in Equ. 4. For all methods including the NCR versions, we set the number of aspects $K = 10$. The regularization coefficients for MF and MF-NCR is $\lambda = 0.01$. The number of local models for LLORMA and LLORMA-NCR is $S = 10$. The learning rate is self adapted as in (Wilson and Martinez 2003). We stop the learning process either when the improvement in training error is smaller than $1e^{-6}$ or when the algorithm reaches 1000 iterations. To reduce the number of parameters, we set the user-defined aspect-specific threshold for the NCR models $\mathbf{b}_{u,k} = 0$ for every u, k.

Evaluation Metrics. We evaluate different approaches based on the following metrics. (1) AUC: computes the area under precision-recall curve; (2) NDCG: another evaluation metric to measure the accuracy of item ranking by the predicted ratings v.s. the actual ranking; (3) MRR: computes the reciprocal of the position of the item with the largest observed rating in the predicted ranking for each user, averaged over all users.

We can see from Table. 3 that overall adopting non-compensatory rules can improve model performance. We observe that for "simpler" models, the improvement is more significant. For example, MF-NCR outperforms MF on all three data sets in terms of higher AUC,NDCG and MRR results and lower RMSE and MAE results. AMF-NCR performs better than AMF on Filmtrust and CiaoDVD and generates better ranking results than AMF on Movielens. LLORMA-NCR achieves higher ranking related AUC,NDCG and MRR results, compared with LLORMA, while preserving comparable RMSE and MAE results. However, increasing the model complexity also leads to increased computation time and poor interpretability. Thus uti-

[2]http://www.grouplens.org

Table 3: Comparative performance for compensatory and non-compensatory rating prediction models, 'Improve' indicates the improvements of non-compensatory versions relative to the original models.

Dataset	Method	AUC	Improve	NDCG	Improve	MRR	Improve
Movielens	MF	0.6661		0.6856		0.8391	
	MF-NCR	0.6990	4.94%	0.7107	3.67%	0.8745	4.23%
	AMF	0.6043		0.5003		0.7506	
	AMF-NCR	0.6129	1.43%	0.5027	0.48%	0.7559	0.71%
	LLORMA			0.8990		0.5761	
	LLORMA-NCR			0.8994	0.04%	0.5761	0.01%
Filmtrust	MF	0.6056		0.5240		0.7522	
	MF-NCR	0.6166	1.81%	0.5252	0.24%	0.7624	1.35%
	AMF	0.6244		0.5055		0.7622	
	AMF-NCR	0.6436	3.07%	0.5098	0.85%	0.7717	1.24%
	LLORMA			0.8672		0.6481	
	LLORMA-NCR			0.8684	0.14%	0.6533	0.80%
CiaoDVD	MF	0.5915		0.6497		0.8427	
	MF-NCR	0.6948	17.45%	0.6872	5.77%	0.8969	6.43%
	AMF	0.6211		0.5048		0.7607	
	AMF-NCR	0.7993	28.69%	0.5657	12.05%	0.8950	17.67%
	LLORMA			0.7827		0.4883	
	LLORMA-NCR			0.7838	0.14%	0.4904	0.43%

lizing non-compensatory rules in simpler models, such as MF and AMF, generates recommendations with higher accuracy, efficiency and interpretability.

Comparative Results for Ranking Models

Data Sets. Next we evaluate the performance of models that target to ranking reconstruction. The datasets used are again Movielens, Filmtrust and CiaoDVD. We construct pair-wise ordering for each user between any higher rated item and lower rated item, i.e. $\mathbf{X}_{u,p} > \mathbf{X}_{u,q} \rightarrow p \succ_u q$. The number of ranking pairs on each dataset is shown in Table. 2

Comparative Methods. We compare the NCR improved versions with the original versions on four widely adopted ranking methods. (1) BT (Hu and Li 2016);: the Bradley-Terry ranking model with MF rating prediction formula, (2) BPR (Rendle et al. 2009): the Thurstonian ranking model with MF rating prediction formula, the optimization is through maximal Bayesian posterior, the regularization coefficient is $\lambda = 0.01$, (3) FSBPR (Zhao et al. 2018): the Thurstonian ranking model with AMF rating prediction formula, the optimization is through maximal Bayesian posterior, (4) LCR (Lee et al. 2014): the Thurstonian ranking model with local low-rank matrix factorization, the loss function for LLORMA and LLORMA-NCR is $log[M]$ which is the log-likelihood. The number of local models is $S = 10$. For all methods including the NCR versions, we set the number of aspects $K = 10$.

Evaluation Metrics. The goal is to reconstruct the observed rankings for each user. Hence we adopt different ranking evaluation metrics, including AUC, NDCG, MRR, MAP and Precision. MAP computes the mean average precision of the correctly ordered pairs. Precision is the concept borrowed from classifier evaluation metric which is based on the fraction of correctly ordered test pairs. To be specific, an item p is predicted to be a winner in a pair of p, q if the prediction favors p, i.e. $p(p \succ_u q) > p(q \succ_u p)$. Precision is the fraction with the number of correctly identified winners (i.e. p is predicted to be superior than q for user u while the actual rating $\hat{\mathbf{X}}_{u,p} > \hat{\mathbf{X}}_{u,q}$) as numerator, and the number of evaluation sessions (i.e. the pair-wise rating rankings) as denominator.

Table 4: Comparative ranking prediction performance

Dataset	Method	MAP	NDCG	Precision	AUC	MRR
Movielens	BT	0.7654	0.5070	0.5307		0.7654
	BT-NCR	0.8440	0.5425	0.6879		0.8440
	BPR	0.8478	0.5443	0.6956		0.8478
	BPR-NCR	0.8623	0.5508	0.7246		0.8623
	FSBPR	0.7474	0.4993	0.4968		0.7484
	FSBPR-NCR	0.7964	0.5205	0.5908		0.7954
	LCR					
	LCR-NCR					
Filmtrust	BT	0.7674	0.5070	0.5307		0.7654
	BT-NCR	0.8182	0.5312	0.6379		0.8190
	BPR	0.7825	0.5147	0.5649		0.7825
	BPR-NCR	0.8365	0.5392	0.6730		0.8365
	FSBPR	0.7484	0.4996	0.4980		0.7490
	FSBPR-NCR	0.7956	0.5205	0.5908		0.7954
	LCR					
	LCR-NCR					
CiaoDVD	BT	0.8009	0.5230	0.6016		0.8008
	BT-NCR	0.9394	0.5857	0.8787		0.9393
	BPR	0.7241	0.4883	0.4481		0.7240
	BPR-NCR	0.9537	0.5922	0.9074		0.9537
	FSBPR	0.7501	0.5001	0.5004		0.7502
	FSBPR-NCR	0.8906	0.5637	0.7815		0.8908
	LCR					
	LCR-NCR					

A general observation in Table. 4 is that embedding non-compensatory rules significantly improves existing ranking models. For example, in terms of precision on all data sets, the non-compensatory rules averagely improve BT's performance by 29% , BPR's performance by 42%, FSBPR by 32% and LCR by $\%$. This observation implies that non-compensatory rules better describe the consumer decision process in comparing between alternative items.

Ranking Performance for Graded Implicit Feedback

In most recommender systems, users not only give explicit ratings but also implicit feedback that can be graded. For example, a purchase and a click are both implicit feedback that indicates user preference. A reasonable grading is that a purchase is "higher" than a click, as a purchase is a stronger indicator of user preference. Therefore, we conduct experiments on datasets with graded implicit feedback.

Data Sets We use three real world datasets, as shown in Table. 5. Tmall[3] is a collection of user shopping sessions, where in each session the user has four types of activities: click, add to cart, add to favorite and purchase. We build two data sets based on Tmall. (1) Tmall-single: a set of pairwise rankings where an item p purchased in u's session is considered to be superrior than an item q clicked in the same session. (2) Tmall-hybrid: the pairwise rankings are built by extracting purchased items in each session and all remaining items which are not purchased in the same session. Thus if an item p is purchased in the session, and an item q is either clicked, added to cart or added to favorite, we build $p \succ_u q$. (3) Yoochoose[4]: a collection of user shopping sessions with

[3]https://ijcai-15.org/index.php/repeat-buyers-prediction-competition

[4]http://2015.recsyschallenge.com

clicked and purchased items. In this data set, user information is not provided. To avoid the cold-start user problem, we assume all sessions are from users with similar preferences.

Table 5: Statistics of Datasets with graded implicit feedback

Dataset	#users	#items	#pairs	#sessions
Tmall-single	33,815	176,231	5,682,833	364,844
Tmall-hybrid	62,101	198,344	6,072,061	475,503
Yoochoose	-	30,852	3,044,572	341,396

Comparative Methods. We compare the NCR improved versions with the original versions on the same four ranking models. It is worthy to note that implementation of BT-NCR is different from previous sections. In our model the prominent aspect is associate with each evaluation session. In the previous experiments, an evaluation session is a rating or a pair of rating. Here we the user interaction session information is available. Thus in BT-NCR, we sample the prominent aspect for each session instead of a pair of actions.

Evaluation Metrics. The goal is to reconstruct the observed rankings of activities precisely. Hence we evaluate different approaches based on the following metrics: MAP, NDCG, MRR, Precision and Recall. are two evaluation metrics designed to evaluate sessional ranking performance. For each session, if there are N purchased items in a session, we compute the pairwise ranking probability between the purchased items and other items and generate a purchase item. The sessional precision is the the fraction of correctly identified items in all items that are predicted to be purchased, recall is the ratio of correctly identified items divided by the number of N. Reported precision and recall are averaged over all sessions.

Table 6: Comparative ranking prediction performance for sessional graded feedback

Dataset	Method	MAP	NDCG	Precision	Recall	MRR
Tmall-single	BT	0.4348	0.2814	0.2787	0.7253	0.4890
	BT-NCR	0.4408	0.2853	0.2810	0.7308	0.4973
	BPR	0.4359	0.2826	0.2789	0.7252	0.4932
	BPR-NCR	0.4410	0.2854	0.2810	0.7305	0.4977
	FSBPR	0.4163	0.2732	0.2734	0.7092	0.4717
	FSBPR-NCR	0.4193	0.2747	0.2749	0.7130	0.4740
	LCR					
	LCR-NCR					
Tmall-hybrid	BT	0.5015	0.3056	0.2934	0.7931	0.5458
	BT-NCR	0.5592	0.3305	0.3044	0.8249	0.6063
	BPR	0.5463	0.3248	0.3006	0.8132	0.5950
	BPR-NCR	0.5635	0.3324	0.3050	0.8267	0.6112
	FSBPR	0.4398	0.2770	0.2768	0.7431	0.4817
	FSBPR-NCR	0.4597	0.2865	0.2831	0.7624	0.5007
	LCR					
	LCR-NCR					
Yoochoose	BT	0.6368	0.4742	0.4569	0.8732	0.7156
	BT-NCR	0.7112	0.5166	0.4786	0.8966	0.7882
	BPR	0.6821	0.5019	0.4711	0.8934	0.7639
	BPR-NCR	0.7049	0.5144	0.4775	0.9030	0.7844
	FSBPR	0.5685	0.4379	0.4374	0.8405	0.6541
	FSBPR-NCR	0.6825	0.5445	0.5362	0.8839	0.7987
	LCR					
	LCR-NCR					

As shown in Table. 6.

Strength of Non-compensatory Rules

We next study how the two types of non-compensatory rules are combined. The parameter θ controls the strength of lexicographical rules. We report the values of θ in Table. 7. We have the following observations. (1) The obtained value $\theta > 0$ for all models on all datasets. Since $\exp\theta > 1$, the prominent aspect is more important than non-prominent aspect in user evaluations. This is consistent to our assumptions that lexicographical rules will evaluate item perforamance first on the most important aspect. (2) The optimal value of θ differs among models and data sets. In general, BT-NCR model relies more on the prominent aspect. (3) We observe positive association between rating model and ranking models with the same type of rating prediction formula. For example, MF-NCR and BPR-NCR adopt the same conventional matrix factorization prediction formula, and they generally derive a smaller θ, compared with AMF-NCR and FSBPR-NCR, which share the same form of neighborhood factorization prediction formula.

Table 7: Scale of strength parameter θ

Method	Movielens	FilmTrust	CiaoDVD
MF-NCR			
AMF-NCR			
LLORMA-NCR			
BT-NCR			
BPR-NCR			
FSBPR-NCR			
LCR-NCR			
Method	Tmall-single	Tmall-hybrid	Yoochoose
BT-NCR			
BPR-NCR			
FSBPR-NCR			
LCR-NCR			

Effect of User-defined Aspect-specific Threshold

Finally, we study the effect of user-defined aspect-specific threshold $\mathbf{b}_{u,k}$. In the previous experiment on rating performance, we set $\mathbf{b}_{u,k} = 0$ for all users and aspects. Here we allow $\mathbf{b}_{u,k}$ to be infered. For a fair comparison, we compare the $MF - NCR - b$ with MF with user and item biases.

Table 8: Comparative rating prediction performance

Dataset	Method	AUC	NDCG	RMSE	MAE	MRR
Movielens	MF-biased					
	MF-NCR-b					
Filmtrust	MF-biased					
	MF-NCR-b					
CiaoDVD	MF-biased					
	MF-NCR-b					

Table 9: Scale of user-defined aspect-specific threshold $\mathbf{b}_{u,k}$

Method	Movielens	FilmTrust	CiaoDVD
$std(\mathbf{b}_{u,k})$			

Conclusion

References

Devooght, R.; Kourtellis, N.; and Mantrach, A. 2015. Dynamic matrix factorization with priors on unknown values. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 189–198. New York, NY, USA: ACM.

Engel, J. F.; Blackwell, R. D.; and Miniard, P. W. 1986. *Consumer Behavior*. The Dryden Press.

Guo, G.; Zhang, J.; Thalmann, D.; and Yorke-Smith, N. 2014. Etaf: An extended trust antecedents framework for trust prediction. In *Proceedings of the 2014 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 540–547.

Guo, G.; Zhang, J.; and Yorke-Smith, N. 2013. A novel bayesian similarity measure for recommender systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2619–2625.

Hauser, J. R.; Ding, M.; and Gaskin, S. P. 2009. Non-compensatory (and compensatory) models of consideration-set decisions. In *2009 Sawtooth Software Conference Proceedings, Sequin WA*.

Hu, J., and Li, P. 2016. Improved and scalable bradley-terry model for collaborative ranking. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 949–954.

Hu, J., and Li, P. 2017. Decoupled collaborative ranking. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1321–1329. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Hunter, D. R. 2004. Mm algorithms for generalized bradley-terry models. *The Annals of Statistics* 32(1):384–406.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37.

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

Koren, Y. 2010. Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data.* 4:1:1–1:24.

Laroche, M.; Kim, C.; and Matsui, T. 2003. Which decision heuristics are used in consideration set formation? *Journal of Consumer Marketing* 20(3):192–209.

Lee, J.; Kim, S.; Lebanon, G.; and Singer, Y. 2013. Local low-rank matrix approximation. In *International Conference on Machine Learning*, 82–90.

Lee, J.; Bengio, S.; Kim, S.; Lebanon, G.; and Singer, Y. 2014. Local collaborative ranking. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, 85–96. New York, NY, USA: ACM.

Lee, I. 2009. *Transforming E-Business Practices and Applications: Emerging Technologies and Concepts: Emerging Technologies and Concepts*. IGI Global.

Lerche, L., and Jannach, D. 2014. Using graded implicit feedback for bayesian personalized ranking. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, 353–356. New York, NY, USA: ACM.

Pilászy, I.; Zibriczky, D.; and Tikk, D. 2010. Fast als-based matrix factorization for explicit and implicit feedback datasets. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, 71–78. New York, NY, USA: ACM.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, 452–461. Arlington, Virginia, United States: AUAI Press.

Salakhutdinov, R., and Mnih, A. 2008. Probabilistic matrix factorization. *Advances in neural information processing systems* 20:1257–1264.

Shi, Y.; Larson, M.; and Hanjalic, A. 2010. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, 269–272. New York, NY, USA: ACM.

Steck, H. 2015. Gaussian ranking by matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, 115–122. New York, NY, USA: ACM.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological review* 34(4):273.

Weston, J.; Bengio, S.; and Usunier, N. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, 2764–2770.

Wilson, D. R., and Martinez, T. R. 2003. The general inefficiency of batch training for gradient descent learning. *Neural Networks* 16(10):1429–1451.

Yu, L.; Zhou, G.; Zhang, C.; Huang, J.; Liu, C.; and Zhang, Z.-K. 2016. Rankmbpr: Rank-aware mutual bayesian personalized ranking for item recommendation. In Cui, B.; Zhang, N.; Xu, J.; Lian, X.; and Liu, D., eds., *Web-Age Information Management*, 244–256. Cham: Springer International Publishing.

Zhang, W.; Chen, T.; Wang, J.; and Yu, Y. 2013. Optimizing top-n collaborative filtering via dynamic negative item sampling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, 785–788. New York, NY, USA: ACM.

Zhao, Q.; Zhang, Y.; Ma, J.; and Duan, Q. 2018. Factored item similarity and bayesian personalized ranking for recommendation with implicit feedback. *Arabian Journal for Science and Engineering*.