Reviewer 1

1. What is the rationale of the proposed non-compensatory conceptual model (figure 1)?

The idea of the proposed non-compensatory conceptual model comes from the study of consumer behavior. According to the famous textbook in marketing (Engel, Blackwell, and Miniard 1986), there are two categories of consumer decision rules: compensatory and non-compensatory rules. In addition, it is widely accepted that there are several types of non-compensatory rules.

In Figure 1, we consider a mixture of two most commonly adopted non-compensatory rules: lexicographic rule and conjunctive rule. To be specific, every part of Figure 1 is supported by study in consumer behavior to be a rational process. (1) Non-compensatory rules distinguish with compensatory rules in that they treat different aspects of an item differently. Therefore, we separate prominent and non-prominent aspects in Figure 1. (2) “Under lexicographic rules, brands are compared on the most important attribute. If one of the brands is perceived as superior based on that attribute, it is selected” (Engel, Blackwell, and Miniard 1986). Therefore, we select one prominent aspect. (3) The importance of attributes is determined by user. Therefore, we assume that prominent aspect is dependent on user preference. (4) “Under conjunctive rules, cutoff points are established for each attribute. If the brand meets the cutoffs for all the attributes, it is chosen“. Therefore, we introduce the threshold b.

We agree with you that more factors can be explored. However, as our work is the first attempt to introduce non-compensatory rules in recommendation models, and our goal is to present a general framework that (1) complies with both rating and ranking based recommendation models (2) universally improves recommendation performances, we believe abstraction and simplification is necessary. For example, allowing more than one prominent aspects raises the issue of determining the number of prominent aspects, which could be tricky.

1. Can you explain why some results are much improved, and others worst for the non-compensatory models?

Thank you for pointing out the anomality. We have to apologize here that we’ve made a silly mistake in typing the BPR’s AUC result on Filmtrust. In the period of rebuttal, we’ve double checked our logs. We also have re-run the BPR and BPR-N experiments. We found that we had mistakenly input one entry to calculate Table 5. The BPR has achieved an AUC of 0.6244 (instead of 0.7825 in our manuscript, which is its MRR result) on the Filmtrust dataset. Other entries are correct. Thus, the non-compensatory model BPR-N improves BPR by 7.75%.

Again, we are sorry for this unfortunate mistake. It is our duty to check everything before submitting. We will do our best to remove all errors including typos and grammar errors in the camera ready version. We will make our source codes and all results publicly available.

Reviewer 2

1. I would like to see a discussion on whether neighborhood models are compensatory or non-compensatory. Being possibly non-compensatory, they should be compared with the proposed non-compensatory latent factor models.

Neighborhood models are **compensatory**. We offer interpretations for two different cases. (1) Item-based neighborhood models such as item-KNN: the predicted rating is a weighted average of the user’s own ratings on neighboring items. Here, K neighboring items act as attributes of interest to the user, the user’s own rating on each neighboring item is the user’s evaluation on a relevant attribute. Item-KNN sums up evaluations on K relevant attributes, which is exactly the definition of compensatory rules. Hence, a high rating by a neighboring item cancels out a low rating by another neighboring item.

(2) User- based neighborhood models such as user-KNN: the predicted rating is computed as the weighted average values of K similar users’ ratings on the target item. Here, neighboring users act as attributes of interest to the user, each neighboring users’ score represents the user’s evaluation on the attribute, and user-KNN is compensatory as it computes an weighted summation over all attributes (neighboring users).

We did not compare our models with pure neighborhood models because it has been shown in many domains that latent factor models outperform pure neighborhood models. However, we did consider incorporating neighborhood information in a latent factor model. We explained in detail that AMF -- one of the earliest and most famous model of this kind (AMF incorporates neighborhood information from explicit feedback in the latent factor model, while SVD++ is the variant for implicit feedback) – is compensatory. We showed that AMF can be adapted to adopt non-compensatory rules and achieve better results.

1. The paper is not clear on how prominent features are selected.

We agree that the selection of prominent aspect is crucial. That is why the prominent aspect is not selected by some heuristics. Instead, the prominent aspect (indexed by k) is treated as a hidden variable in each evaluation session. The prominent aspect is sampled according to the user preference. Thus, the larger the user preference uk is, the more likely k is selected for any of u’s evaluation sessions.

For Bradley-Terry models, we adopt a stochastic expectation maximization algorithm to infer the hidden prominent aspect for each evaluation session.

For rating models and Thurstonian ranking models (i.e. BPR style models), we find it more convenient to integrate out all possible hidden prominent aspects in the rating prediction formulas (Equ. 8 ~Equ.10). Thus, the prominent aspect k is not explicitly inferred.

1. The additional parameters \theta and b may also be an obstacle. Latent factor models already ask for at least 4 user-defined hyper-parameters. Having to tune 2 additional ones may be a problem in practice. Moreover, the paper does not clearly specify what were the values of \theta and b used in each of the experiments and how they were obtained. The parameter analysis section does not have enough information to clarify this. Was this methodology used to obtain optimal values in the experiments? What are these optimal values?

Hyper-parameters of non-compensatory models such as the regularization coefficient \lambda in MF are tuned by 5-fold cross validation. Non-compensatory models use the same value for the hyper-parameters of their compensatory versions. For example \lambda=0.01 for MF and MF-N.

Parameters of non-compensatory models in our manuscript consist of the user preferences U, item features V, strength coefficient for prominent aspect \theta, and threshold for each user and each aspect b. They are to be learnt by stochastic gradient descent. Due to the limit of space, we did not include the derivation of updating equations for U,V,\theta and b. We will provide a supplementary material of derivations if necessary.

However, non-compensatory models do not necessarily have many additional parameters. (1) Firstly, the strength parameter \theta is a global parameter for all users. (2) Secondly, the threshold parameter b does not appear in non-compensatory ranking models. For example, it cancels between two items rated by the same user. (3) Thirdly, non-compensatory rating models without threshold parameters b also greatly outperform compensatory models. As mentioned in the first part of experiments (comparative results for rating prediction models), to reduce the number of parameters, we set b=0 for all users and aspects. This is equivalent as removing threshold parameters from the model (see Equ. 8~ Equ.10). The results in Table 4 validates the capability of non-compensatory models, even without threshold parameters b. (4) Finally, we also conduct the full model with threshold parameters b and strength parameters \theta learnt from training sets. The full model and the optimal parameters learnt are described in the last part of experiment (analysis of inferred parameters), which we will elaborate in the following paragraphs.

A full model with threshold parameters b outperforms the variant with b removed (i.e. the one used in the first part of the experiment). For example, non-compensatory MF is improved by 6.61%, 4.18%, 3.57% in terms of AUC on movielens, Filmtrust and CiaoDVD respectively with threshold parameters b. We did not include this experimental result because the performance improvement is predictable (i.e. setting b=0 clearly does not guarantee optimality). We will provide these results in a supplementary material.

Due to the page limit, we provide another analysis which we believe to be more important: the optimal values of \theta and b learnt in a full model. (1) The learnt \theta is moderate, indicating that users adopt a combination of lexicographical rules and conjunctive rules. (2) The optimal b learnt is not likely to be zero, suggesting that it would be better to adopt the full model when computational resource is affordable. (3) The not-small standard deviation of learned b within each user indicates that a user bias term varying among different latent feature dimensions is preferred. This suggests introducing a dimension-dependent user bias term is effective.

1. There is also the question of scalability. It is not clear if there a price to pay - computationally - in the adaptation for the non-compensatory model.

We make an effort to design the models so that the same algorithmic framework can be used for both compensatory models, so that the two counterparts are comparable computational wise. For example, the non-compensatory rating models and Thurstionian ranking models (BPR style models) are all solved by stochastic gradient descent algorithm, non-compensatory Bradley-Terryl model adopts a stochastic EM algorithm. Furthermore, one can adopt the accelerated version of rating model with threshold parameters b removed to obtain satisfying results. In that case, the computational cost between compensatory and non-compensatory models is subtal.

1. What exactly is the goal of "revealing the observed rankings"? When are rankings observed?

The observed rankings depend on the applications at hand. To be specific, in the experiments we consider observed rankings based on (1) ratings that can be compared. If for a user, the user’s rating on item I is greater than her rating on item j then the observed ranking is i>j. (2) implicit feedback that can be graded. For example, in Tmall-single, if in a user session, if an item I is purchased and another item j is clicked, then the observed ranking is i>j. In Tmall-hybrid, if an item I is purchased and another item j is either clicked, added to cart or added to favorite, then the observed ranking is i>j.

1. Notation problems

Reviewer #3

1. What was the formula used for calculating AUC?

Yes. To calculate AUC, we perform a binary classification. For the ground truth, a normalized rating that is above 0.5 is labeled as positive and otherwise negative. The predicted ratings by compensatory or non-compensatory models are used to determine the class label.

(2) Problems related to Equ. 6, 7 and 11

Thank you very much for pointing that out.

In this manuscript, we follow the non-exponential form of Bradley-Terry model. That is, the probability p(i > j) = i/(i + j), where I and j are positive.

Therefore for the compensatory Bradley-Terry model, Equation 7 should be Xui/(Xui + Xuj). It is worthy to note that, to make sure Xui and Xuj are positive, in implementation we add non-negative constraints for the user preferences and item features.

Equation 11 is correct.