An Interpretable Model for Collaborative Filtering Using an Extended Latent Dirichlet Allocation Approach

Florian Wilhelm[1,*], **Marisa Mohr**[1], **Lien Michiels**[2]
[1] inovex GmbH, 76131 Karlsruhe, Germany
[2] University of Antwerp, 2020 Antwerpen, Belgium
[*] Correspondence: florian.wilhelm@inovex.de

Abstract

With the increasing use of AI and ML-based systems, interpretability is becoming an increasingly important issue to ensure user trust and safety. This also applies to the area of recommender systems, where methods based on matrix factorization (MF) are among the most popular methods for collaborative filtering tasks with implicit feedback. Despite their simplicity, the latent factors of users and items lack interpretability in the case of the effective, unconstrained MF-based methods. In this work, we propose an extended latent Dirichlet Allocation model (LDAext) that has interpretable parameters such as user cohorts of item preferences and the affiliation of a user with different cohorts. We prove a theorem on how to transform the factors of an unconstrained MF model into the parameters of LDAext. Using this theoretical connection, we train an MF model on different real-world data sets, transform the latent factors into the parameters of LDAext and test their interpretation in several experiments for plausibility. Our experiments confirm the interpretability of the transformed parameters and thus demonstrate the usefulness of our proposed approach.

Introduction

In the field of recommendation systems, collaborative filtering is a fundamental technique that filters for patterns in user-item interactions to make predictions about future interactions. For decades, matrix factorization methods have been the state of the art for collaborative filtering. Even the rise of deep learning in this domain could not change this dominance (Rendle et al. 2020; Dacrema, Cremonesi, and Jannach 2019).

Although MF-based methods for recommender systems have been studied for a long time, the reason why they are so effective in finding a personalized ranking of items remains largely unclear. Since the idea of MF has its origin in linear algebra and not in a probabilistic generative process, there is no canonical interpretation of the learned latent factors. Some variants, such as non-negative matrix factorization (NMF), allow interpretation but often cannot rival the performance of general MF methods without constraints on the factors (Lee, Sun, and Lebanon 2012). As users of recommendations systems are increasingly sensitive to how their data is used and why certain recommendations are presented to them, the need for interpretability of MF-based methods is rising.

In this paper, we propose an approach that allows the interpretation of MF by transforming the latent factors into parameters of an interpretable model while keeping the MF-induced personalized ranking constant. We first introduce MF-based methods and how they solve the task of creating a user-specific ranking for a set of items based on the implicit feedback of a set of users. After that, we review the well-known Latent Dirichlet Allocation (LDA) model for the task of collaborative filtering and point out its shortcomings compared to MF. Therefore, LDA cannot be used as an interpretable alternative to MF directly. Subsequently, we propose an extended LDA (LDAext) model that remedies these shortcomings and show in a constructive proof the equivalence of MF and LDAext in the sense that the latent factors of MF can be transformed into the parameters of LDAext while maintaining the personalized ranking. We perform several experiments on different real-world data sets to evaluate the plausibility when interpreting the transformed latent factors of MF.

Related Work

With respect to the interpretability of MF-based methods, Zhang et al. (2006) propose NMF for collaborative filtering and interpret the latent user vector as an additive mixture of different user communities, i.e., cohorts. Hernando, Bobadilla, and Ortega (2016) replace the mixture of cohorts by a proper distribution over cohorts of users, which increases the interpretability. Both approaches require a non-negativity constraint on the factors, which reduces their performance in practical applications (Lee, Sun, and Lebanon 2012). Ding, Tao Li, and Jordan (2010) relieve these constraints by requiring only one of the factors to be non-negative in their semi NMF method and interpret it as a relaxation of K-means clustering. This interpretation as a clustering of items in the space of users is less versatile than the generative process of LDA with its inherent interpretability.

The direct application of LDA for collaborative filtering tasks is proposed by Blei, Ng, and Jordan (2003) in their original work and also in a slightly modified approach by

Xie, Dong, and Gao (2014). To incorporate content-based information about items, MF models are also used in conjunction with LDA to extract textual information (Wang and Blei 2011; Nikolenko 2015). While the use of content-based information also provides the possibility for some interpretability, it can only be considered an auxiliary help that is not available in pure collaborative filtering tasks.

An extension of LDA to interpret MF is proposed by Wilhelm (2021). In this work, we follow a similar approach but improves the interpretability by using Dirichlet distributions for the cohorts of item preferences and item popularity as well. Additionally, we demonstrate the plausibility of the interpretation in several experiments.

Notation and Terminology

Matrices are denoted by capital letters X, transposed matrices by X^t, vectors by bold letters $\mathbf{x}$, sets by calligraphic letters $\mathcal{X}$, and the cardinality of a set by $|\mathcal{X}|$. The scalar product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{n} x_i y_i$ and the l_1-norm is denoted by $\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|$, where n is the dimension of the vector space. A concatenation of two vectors $\mathbf{x}, \mathbf{z}$ is denoted by $[\mathbf{x}, \mathbf{z}]$. The i-th row vector of a matrix X is denoted by $\mathbf{x}_i$ and the j-th column vector as $\mathbf{x}_{*j}$. $\mathbb{R}_{\geq 0}$ denotes non-negative real numbers.

Let $\mathcal{U}$ be the set of all users and $\mathcal{I}$ the set of all items. With $\mathcal{S} \subset \mathcal{U} \times \mathcal{I}$ we denote the set of implicit feedback from users $u \in \mathcal{U}$ having interacted with items $i \in \mathcal{I}$. The task of personalized ranking is to provide each user u with a personalized total ranking $\succcurlyeq_u$ on $\mathcal{I}$ (Rendle et al. 2009).

Matrix Factorization

In MF-based methods, the sparse matrix of user-item interactions $X \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ is approximated by the product of two low-rank matrices $W \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{K}|}$ and $H \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{K}|}$, i.e.,

$$X \approx \hat{X} := W H^t,$$

where $\mathcal{K} = \{1, \ldots, |\mathcal{K}|\}$ is the index set of the latent dimensions. Commonly, one latent dimension of W is fixed to 1 for all users, which leads to an additional item bias term that has been shown to improve the predictive power of the model (Paterek 2007; Koren and Bell 2015). Therefore, we define the personalized score of a user u for an item i as

$$\hat{x}_{ui} = \langle \mathbf{w}_u, \mathbf{h}_i \rangle + b_i, \tag{1}$$

where $b_i \in \mathbb{R}$ is an item bias, which can be interpreted as the item's popularity. The personalized scores of a user then induce the personalized ranking $\succcurlyeq_u$ by virtue of $\hat{x}_{ui} \geq \hat{x}_{uj}$ for $i, j \in \mathcal{I}$.

The approximation $\hat{X}$ is highly dependent on the optimization loss $L(X, \hat{X})$, e.g. SVD++ (Koren 2009), WR-MF (Hu, Koren, and Volinsky 2008; Pan et al. 2008) and PMF (Salakhutdinov and Mnih 2007). In contrast to an approximation, Rendle et al. (2009) propose the Bayesian Personalized Ranking (BPR) loss to directly optimize for an optimal ranking $\succcurlyeq_u$.

Although the score $\hat{x}_{ui}$ is computed using a simple scalar product, it is hard to interpret the latent vectors $\mathbf{h}_i$ and $\mathbf{w}_u$. At first glance, they might quantify the prevalence of some latent feature in an item while the corresponding element of a user u quantifies the user's preference for this feature. The problem with this interpretation becomes apparent when considering negative elements, especially in $\mathbf{h}_i$. This observation motivates the usage of NMF methods that demand non-negativity for $\mathbf{w}_u$ and $\mathbf{h}_i$, which perform worse in direct comparison (Lee, Sun, and Lebanon 2012).

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative statistical model from the field of natural language processing. It is an instance of a *topic model* in that it explains the observations by assuming a set of unobserved groups or topics, where the observations within an assigned group share some common features. Simply speaking, LDA assumes that each document is a mixture of latent topics and each topic assigns a certain probability of occurrence to each word.

We reformulate the generative process of a smoothed LDA from Blei, Ng, and Jordan (2003) for the context of collaborative filtering. Given a set of items $i \in \mathcal{I}$ and $|\mathcal{K}|$ cohorts of users, each user $u \in \mathcal{U}$ has $\mathcal{S}_u = \{1, \ldots, |\mathcal{S}_u|\}$ interactions, assuming the following generative process:

1. choose $\boldsymbol{\theta}_u \sim \mathrm{Dir}(\boldsymbol{\alpha})$ for $u \in \mathcal{U}$,
2. choose $\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$ for $k \in \mathcal{K}$,
3. for each user $u \in \mathcal{U}$ and interactions $s \in \mathcal{S}_u$:
   (a) choose a cohort $z_{us} \sim \mathrm{Cat}(\boldsymbol{\theta}_u)$,
   (b) choose an item $i_{us} \sim p(i_{us} \mid \boldsymbol{\varphi}_{z_{us}}) := \mathrm{Cat}(\boldsymbol{\varphi}_{z_{us}})$.

The hyperparameters of this generative process are $\boldsymbol{\alpha} \in \mathbb{R}_{>0}^{|\mathcal{K}|}$ and $\boldsymbol{\beta} \in \mathbb{R}_{>0}^{|\mathcal{I}|}$ as well as the number of user interactions S_u, which is known from the interaction matrix X. The graphical model corresponding to the generative process is depicted in Figure 1.

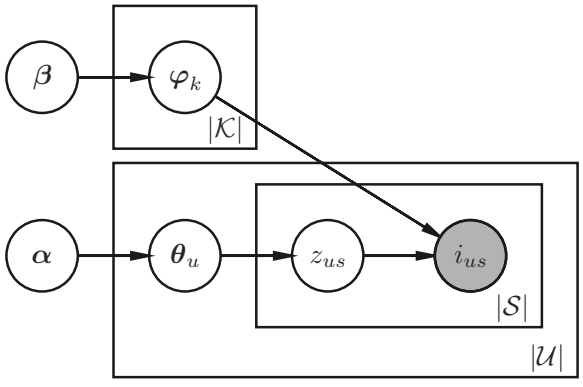

Figure 1: Graphical model of LDA in a collaborative filtering setting (Blei, Ng, and Jordan 2003). The upper plate represents the cohorts. The lower outer and inner plate represent the users and the repeated choice of cohorts and items within the interactions of a user (Step 3 of LDA), respectively.

With respect to interpretability, we can interpret $\boldsymbol{\varphi}_k$ as a cohort of users who have a certain preference for every item in $\mathcal{I}$. Each user then has an affiliation $\boldsymbol{\theta}_u$ with each of the cohorts. Together, they fully determine the probability of a user interacting with an item.

We can make the connection between LDA and MF explicit by interpreting these parameters as parameters of an MF model. We see that $\boldsymbol{\varphi}_{*i}$ roughly corresponds to the item vectors and $\boldsymbol{\theta}_u$ to the user vectors. However, LDA has some shortcomings in comparison to MF that can explain LDA's poor performance in the context of recommendation systems.

Firstly, there is no notion of item popularity in LDA's generative process. Thus, one item can be highly likely in one cohort and highly unlikely in another as there is no regularisation across the cohorts. Accounting for item popularity is however important in collaborative filtering, hence why it is included in (1) as b_i. We can remedy this by including an item bias term in the categorical distribution of Step 3b.

Secondly, if a bias term for item popularity is added to the process, the effect of this regularization will be the same for all users. There is no notion of more individualistic or more conformist users. MF-based methods, on the other hand, have this flexibility because the norm of $\mathbf{w}_u$ influences the regularization effect of the item biases. We address this by introducing a user-specific weighting factor for the item popularity in order to gain the same flexibility.

It is now clear that MF exhibits these inductive biases, whereas traditional LDA does not. Consequently, we propose an extended LDA model with the same flexibility as MF.

Extended Latent Dirichlet Allocation

We modify traditional LDA by incorporating the item popularity δ_i and the user's conformity λ_u. This results in the generative process of an extended LDA (LDAext) more suitable for recommender systems:

1. choose $\boldsymbol{\theta}_u \sim \mathrm{Dir}(\boldsymbol{\alpha})$ for $u \in \mathcal{U}$,
2. choose $\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$ for $k \in \mathcal{K}$,
3. choose $\boldsymbol{\delta} \sim \mathrm{Dir}(\boldsymbol{\gamma})$,
4. choose $\lambda_u \sim \mathrm{LogNormal}(\mu, \sigma^2)$ for $u \in \mathcal{U}$,
5. for each user $u \in \mathcal{U}$ and interactions $s \in \mathcal{S}_u$:
   (a) choose a cohort $z_{us} \sim \mathrm{Cat}(\boldsymbol{\theta}_u)$,
   (b) choose an item $i_{us} \sim p(\, i_{us} \mid \boldsymbol{\varphi}_{z_{us}}, \boldsymbol{\delta}, \lambda_u \,) := \mathrm{Cat}(\mathbf{p}_{z_{us}}(u))$ where $\mathbf{p}_{z_{us}}(u) = \|\mathbf{c}_{z_{us}}(u)\|_1^{-1} \mathbf{c}_{z_{us}}(u)$ with $\mathbf{c}_{z_{us}}(u) = \boldsymbol{\varphi}_{z_{us}} + \lambda_u \boldsymbol{\delta}$.

The additional hyperparameters μ, σ^2 and $\boldsymbol{\gamma} \in \mathbb{R}_{>0}^{|\mathcal{I}|}$ can be used to incorporate prior knowledge about the relations of $\boldsymbol{\lambda}$, $\boldsymbol{\delta}$ and $\boldsymbol{\varphi}_k$. In Step 5b, we see that the probability of a user interacting with an item not only depends on the preference assigned to the item by the cohort $\boldsymbol{\varphi}_{z_{us}}$, but also the popularity δ_i of the item and the conformity λ_u of the user. By virtue of the expected value $\mathbb{E}[i_{us}] = \langle \boldsymbol{\theta}_u, \mathbf{p}_{*i}(u) \rangle$ of this generative process, LDAext induces a personalized ranking $\geqslant_u$ similar to the personalized score of MF in (1). The graphical model of LDAext is illustrated in Figure 2.

As we derived the parameters of LDAext from notions about the real world, they have an intuitive interpretation by design. An example use of this intuitive interpretation is finding similar users in terms of conformity or cohort affiliation. The canonical metric of distance for categorical distributions $\boldsymbol{\theta}_u$ is the *index of dissimilarity*, i.e.,

$$D(u, v) = \frac{1}{2} \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_1 \in [0, 1],$$

or equivalently the *overlap*, i.e.,

$$O(u, v) = 1 - D(u, v) = \sum_{k \in \mathcal{K}} \min(\theta_{uk}, \theta_{vk}) \in [0, 1], \quad (2)$$

for two users u and v. In the case of MF, we have two user vectors $\mathbf{w}_u, \mathbf{w}_v \in \mathbb{R}^{|\mathcal{K}|}$ and thus it is hard to argue if they should be compared in terms of $\|\mathbf{w}_u - \mathbf{w}_v\|_1$, $\|\mathbf{w}_u - \mathbf{w}_v\|_2$ or even a completely different metric.

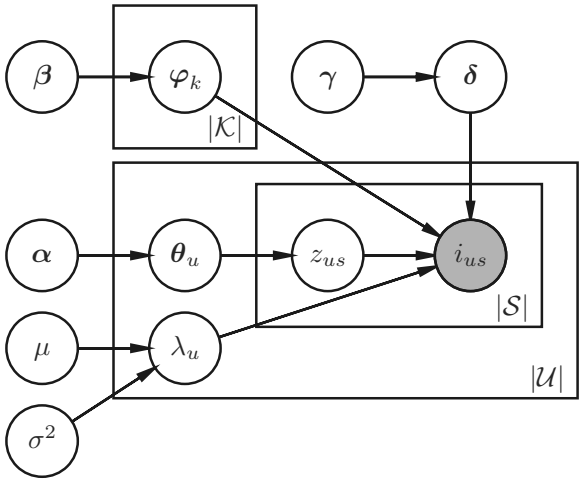

Figure 2: Graphical model of LDAext that also incorporates the popularity of items $\boldsymbol{\delta}$ as well as the user's conformity λ_u to these popularities.

In the following, we show that MF has an adjoint formulation that corresponds to the parameters $\boldsymbol{\varphi}_k$, $\boldsymbol{\theta}_u$, δ_i and λ_u of LDAext. Finally, this allows us to intuitively interpret the latent factors of MF.

LDAext Formulation of Matrix Factorization

To derive the adjoint LDAext formulation of MF, we use a lemma from Wilhelm (2021) that allows us to transform an MF into an NMF. For completeness, we reproduce the lemma and its short proof. We then prove our new theorem that allows us to transform the factors of MF into the parameters of LDAext.

Lemma. *Given personalized ranking scores* $\hat{x}_{ui} = \langle \mathbf{w}_u, \mathbf{h}_i \rangle + b_i$ *for users* $u \in \mathcal{U}$ *and items* $i \in \mathcal{I}$ *with* $\mathbf{w}_u \in \mathbb{R}^{|\mathcal{K}|}$, $\mathbf{h}_i \in \mathbb{R}^{|\mathcal{K}|}$ *and* $b_i \in \mathbb{R}$ *that induce a total ranking* $\geqslant_u$ *for all users. Then there exists* $x'_{ui} = \langle \mathbf{w}'_u, \mathbf{h}'_i \rangle + b'_i$

with $\mathbf{w}'_u \in \mathbb{R}_{\geq 0}^{|\mathcal{K}'|}$, $\mathbf{h}'_i \in \mathbb{R}_{\geq 0}^{|\mathcal{K}'|}$ *and* $b'_i \in \mathbb{R}_{\geq 0}$ *that induce the same total ranking* $\geqslant_u$ *for all users.*

Proof. We define $\mathbf{w}'_u = [\mathbf{w}^+{}_u, \mathbf{w}^-{}_u]$ where

$$\mathbf{w}^+{}_{uk} = \begin{cases} \mathbf{w}_{uk} & \text{if } \mathbf{w}_{uk} \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

$$\mathbf{w}^-{}_{uk} = \begin{cases} -\mathbf{w}_{uk} & \text{if } \mathbf{w}_{uk} < 0 \\ 0 & \text{otherwise} \end{cases},$$

for $k \in \mathcal{K}$. Also, we define analogously $\mathbf{h}'_i = [\mathbf{h}_i + \mathbf{s}, -\mathbf{h}_i + \mathbf{s}]$ with $\mathbf{s} = (s_i)_{i\in\mathcal{I}}$, $s_i = \max_{k\in\mathcal{K}} |h_{ik}|$ and $b'_i = b_i + \max_{i\in\mathcal{I}} |b_i|$. By construction, we have $\mathbf{w}'_u \in \mathbb{R}_{\geq 0}^{|\mathcal{K}'|}$, $\mathbf{h}'_i \in \mathbb{R}_{\geq 0}^{|\mathcal{K}'|}$ and $b'_i \in \mathbb{R}_{\geq 0}$ with $\mathcal{K}' = \{1, \ldots, 2|\mathcal{K}|\}$. Using these definitions, we trivially have $x'_{ui} \geq x'_{uj}$ if and only if $\hat{x}_{ui} \geq \hat{x}_{uj}$. Subsequently, x'_{ui} and $\hat{x}_{ui}$ induce the same total ranking $\geqslant_u$. □

Theorem. *Given personalized ranking scores* $\hat{x}_{ui} = \langle \mathbf{w}_u, \mathbf{h}_i \rangle + b_i$ *for users* $u \in \mathcal{U}$ *and items* $i \in \mathcal{I}$ *with* $\mathbf{w}_u \in \mathbb{R}^{|\mathcal{K}|}$, $\mathbf{h}_i \in \mathbb{R}^{|\mathcal{K}|}$ *and* $b_i \in \mathbb{R}$ *that induce a total ranking* $\geqslant_u$ *for all users. Then there exist* $\boldsymbol{\theta}_u$, $\boldsymbol{\varphi}_k$, $\boldsymbol{\delta}$, λ_u *and consequently* $\mathbf{p}(u)$*, such that the corresponding generative process of the extended LDA formulation induces the same total ranking* $\geqslant_u$ *by virtue of* $x'_{ui} = \langle \boldsymbol{\theta}_u, \mathbf{p}_{*i}(u) \rangle$ *for all users.*

Proof. By virtue of the lemma, we assume $\mathbf{w}_u \in \mathbb{R}_{\geq 0}^{|\mathcal{K}'|}$, $\mathbf{h}_i \in \mathbb{R}_{\geq 0}^{|\mathcal{K}'|}$ and $b_i \in \mathbb{R}_{\geq 0}$ without loss of generality. Let

$$\boldsymbol{\varphi}_k = \|\mathbf{h}_{*k}\|_1^{-1} \mathbf{h}_{*k}, \qquad \boldsymbol{\delta} = \|\mathbf{b}\|_1^{-1} \mathbf{b},$$
$$\theta_{uk} = \langle \hat{\mathbf{w}}_u, \mathbf{n}_u \rangle^{-1} \hat{w}_{uk} n_{uk}, \quad \lambda_u = \|\hat{\mathbf{w}}_u\|_1^{-1} \|\mathbf{b}\|_1,$$

where $\hat{w}_{uk} = \|\mathbf{h}_{*k}\|_1 w_{uk}$, $\mathbf{n}_u = (n_{uk})_{k\in\mathcal{K}'}$ with $n_{uk} = \sum_{i\in\mathcal{I}} \varphi_{ki} + \lambda_u \delta_i$. In the pathological case $\|\hat{\mathbf{w}}_u\|_1 = 0$, we have a trivial solution and $\geqslant_u$ only depends on $\boldsymbol{\delta}$. In case of $n_{uk} = 0$, we have $\sum_{i\in\mathcal{I}} \varphi_{ki} = 0$, consequently $\|\mathbf{h}_{*k}\|_1 = 0$ and thus the k-th latent vector can just be removed. Therefore, we assume now $\|\hat{\mathbf{w}}_u\|_1 > 0$ and $n_{uk} > 0$ and define $\mathbf{p}_k(u) = n_{uk}^{-1}(\boldsymbol{\varphi}_k + \lambda_u \boldsymbol{\delta})$ according to Step 5b in LDAext for each user u and cohort k. By construction, $\boldsymbol{\varphi}_k$, $\boldsymbol{\theta}_u$, $\boldsymbol{\delta}$ and $\mathbf{p}_k(u)$ are event probabilities of categorical distributions. Using these definitions, we have

$$\begin{aligned} x'_{ui} = \langle \boldsymbol{\theta}_u, \mathbf{p}_{*i}(u) \rangle &= \sum_{k\in\mathcal{K}'} \theta_{uk}\, p_{ki}(u) \\ &= \sum_{k\in\mathcal{K}'} \langle \hat{\mathbf{w}}_u, \mathbf{n}_u \rangle^{-1} \hat{w}_{uk} (\varphi_{ki} + \|\hat{\mathbf{w}}_u\|_1^{-1} b_i) \\ &= \langle \hat{\mathbf{w}}_u, \mathbf{n}_u \rangle^{-1} \sum_{k\in\mathcal{K}'} (\hat{w}_{uk} \varphi_{ki} + \hat{w}_{uk} \|\hat{\mathbf{w}}_u\|_1^{-1} b_i) \\ &= \langle \hat{\mathbf{w}}_u, \mathbf{n}_u \rangle^{-1} \Big(\sum_{k\in\mathcal{K}'} w_{uk}\, h_{ik} + b_i \Big) \\ &= \langle \hat{\mathbf{w}}_u, \mathbf{n}_u \rangle^{-1} (\langle \mathbf{w}_u, \mathbf{h}_i \rangle + b_i) = \langle \hat{\mathbf{w}}_u, \mathbf{n}_u \rangle^{-1} \hat{x}_{ui} \end{aligned}$$

Noting that the factor $\langle \hat{\mathbf{w}}_u, \mathbf{n}_u \rangle^{-1}$ only depends on u, we conclude that x'_{ui} and $\hat{x}_{ui}$ induce the same personalized ranking $\geqslant_u$. □

Evaluation

To show the implications for practical applications when interpreting the factors of MF as parameters of LDAext, we perform several experiments on public data sets. We evaluate to what extent $\boldsymbol{\varphi}_k$, $\boldsymbol{\theta}_u$, λ_u and δ_i can be interpreted as a cohort with preferences for items, a user's affiliation with different cohorts, the conformity of a user and the popularity of an item, respectively.

For our evaluation, we use three different data sets and prune users with less than 20 interactions. The pruned *MovieLens-1M* data set encompasses approximately 1 million movie ratings across 6,040 users and 3,706 movies, which accounts to a sparsity of 4.4% (Harper and Konstan 2016). After being pruned, *Goodbooks* has approximately 6 million interactions across 53,425 users and 10,000 books with a sparsity of 1.1% (Zajac 2017). The *Amazon* data set of ratings and reviews is further reduced by pruning items with less than 50 interactions, eventually yielding about 1.35 million ratings across 23,632 users and 27,028 items with a sparsity of 0.2% (Leskovec, Adamic, and Huberman 2007).

We adhere to the following evaluation protocol: For each of the three data sets, we split into a train and test set by randomly selecting 10 items for each user as test set. Then we train an MF model with BPR loss on each data set using 5 different random seeds. For MovieLens-1M we set $|\mathcal{K}| = 64$ and for Goodbooks and Amazon $|\mathcal{K}| = 256$. These numbers were determined beforehand to optimize for precision at 10, i.e., the fraction of known positives in the first 10 positions of the ranked list of prediction results. Eventually, we calculate for each trained MF model the adjoint LDAext formulation, then perform 4 statistical experiments defined below and report the mean as well as the standard deviation.

Experiments

1. **Cohort Allocation Test.** In this test, we evaluate the distributions of the cohorts $\boldsymbol{\varphi}_k$ by choosing for each user from $\boldsymbol{\theta}_u$ the cohort k_{max} with the maximum event probability and randomly choose one of the cohorts with event probability 0 as k_{min}. For every user u and the items $\mathcal{I}_u$ that the user interacted with, we calculate the sum of log probabilities for the two cohorts, i.e., $s_{uk} := \sum_{i\in\mathcal{I}_u} \log(\varphi_{ki})$ for $k \in \{k_{\text{min}}, k_{\text{max}}\}$. We then conduct a one-tailed, paired t-test on $t_{uk_{\text{max}}}$ and $t_{uk_{\text{min}}}$ with the null hypothesis that the user's interaction are not more likely in the cohort with which the user is most affiliated than in the cohort with which the user is least affiliated.

2. **Popularity Ranking Test.** To study if $\boldsymbol{\delta}$ can be interpreted as item popularity, we first calculate the empirical item popularity δ'_i as the number of user's that interacted with i and determine the Kendall τ_C correlation coefficient for $\boldsymbol{\delta}$ and $\boldsymbol{\delta}'$. Our null hypothesis is that they are not correlated.

3. **Conformity Ranking Test.** To test if λ_u can be interpreted as the conformity of a user towards item popularity, we calculate for the interactions of each user $\mathcal{I}_u$ the average item popularity, i.e., $\lambda'_u := \frac{1}{|\mathcal{I}_u|} \sum_{i\in\mathcal{I}_u} \delta_i$. Eventually, we determine the Kendall τ_C correlation coefficient

dataset	train	test
MovieLens-1M	0.608 ± 0.022	0.162 ± 0.018
Amazon	0.125 ± 0.004	0.059 ± 0.004
Goodbooks	0.178 ± 0.010	0.047 ± 0.005

Table 1: Results of experiment 1, i.e., cohort allocation test, reporting Cohen's d for the train and test set.

dataset	experiment 2	experiment 3
MovieLens-1M	0.520 ± 0.006	0.3749 ± 0.0090
Amazon	0.377 ± 0.003	0.0818 ± 0.0026
Goodbooks	0.265 ± 0.004	-0.0016 ± 0.0013

Table 2: Results of experiment 2, i.e., popularity ranking test, and experiment 3, i.e., conformity ranking test, showing Kendall τ_C coefficient.

for $\boldsymbol{\lambda}'$ and $\boldsymbol{\lambda}$ with the null hypothesis that they are not correlated.

4. **User's Preferences Test.** In this test we examine if $\boldsymbol{\theta}_u$ is a good proxy for the preferences of a user. We randomly choose 2,000 users for each data set and determine for each user a good $\hat{t}_u$ and a bad twin $\check{t}_u$, such that the overlap $O(\theta_u, \theta_{\hat{t}_u})$ and $O(\theta_u, \theta_{\check{t}_u})$ defined in (2) is maximal and minimal, respectively. Independent of the user's preferences we also choose a random twin $\tilde{t}_u$ for each user. Using the Jaccard coefficient J, we determine $J(\mathcal{I}_u, \mathcal{I}_{\hat{t}_u})$, $J(\mathcal{I}_u, \mathcal{I}_{\check{t}_u})$ as well as $J(\mathcal{I}_u, \mathcal{I}_{\tilde{t}_u})$ and conduct two one-tailed, paired t-test with the null hypothesis that J of a good twin is not greater than the one of the bad twin and analogously for the good and the random twin.

Results

All results reported in this section have a p-value of less than 10^{-6} and thus the null hypothesis can confidently be rejected.

1. **Cohort Allocation Test.** We conduct the experiment on both the train and test set of each data set. In Table 1, we report Cohen's d for each case. In general, we see larger effect sizes on the train than on the test set, which is to be expected as the model was fitted on train. We note that the effect sizes are small to moderate. This is due to the fact that users are affiliated with many cohorts and rarely only with a few, thus the entropy of $\boldsymbol{\theta}_u$ is high.
2. **Popularity Ranking Test.** The results of the second experiment are shown in Table 2. We find that the effect sizes of τ_C are large, i.e. greater than 0.30. Only on Goodbooks we see a moderate effect. This supports our interpretation of $\boldsymbol{\delta}$ as item popularity.
3. **Conformity Ranking Test.** In Table 2 we also find the results of experiment 3. The effect size of the conformity test is large for MovieLens-1M but quite small for Amazon and even negative in the case of Goodbooks. Looking at the median of $\boldsymbol{\lambda}$, we have for Movielens-1M a value of 0.016, for Amazon 0.011 and Goodbooks 0.003. This explains the results, as the importance of the item popularity in Goodbooks is negligible in comparison to the other datasets. Thus, the slight negative correlation close to the standard deviation indicates that there is almost no correlation.
4. **User's Preferences Test.** In the results of the fourth experiment as shown in Table 3, we see very large effect sizes for train and small to large for test. As expected, the comparison of a good and bad twin has a larger effect than comparing a good to a random twin. Thus, we conclude that the user representation $\boldsymbol{\theta}_u$ with the metric O from (2) can be used for clustering users in an interpretable way.

dataset	other	train	test
MovieLens-1M	bad	1.88 ± 0.03	0.61 ± 0.02
Amazon	bad	1.48 ± 0.03	0.86 ± 0.03
Goodbooks	bad	2.88 ± 0.06	0.62 ± 0.04
MovieLens-1M	rnd	1.17 ± 0.01	0.45 ± 0.02
Amazon	rnd	1.44 ± 0.03	0.84 ± 0.03
Goodbooks	rnd	2.12 ± 0.04	0.43 ± 0.02

Table 3: Results of experiment 4, i.e., user's preferences test, showing Cohen's d when comparing the Jaccard coefficient $J(\mathcal{I}_u, \mathcal{I}_{\hat{t}_u})$ with the one of a bad twin $J(\mathcal{I}_u, \mathcal{I}_{\check{t}_u})$ and the one of a random (rnd) twin $J(\mathcal{I}_u, \mathcal{I}_{\tilde{t}_u})$ on train and test set.

Conclusion

In the theoretical part of this paper, we have reviewed MF-based methods as well as LDA and highlighted their differences for collaborative filtering. Based on these findings, we proposed a novel extended LDA (LDAext) approach that remedies the deficits of LDA for recommendation tasks by incorporating additional parameters for the popularity $\boldsymbol{\delta}$ of items and the conformity of users to the popularity $\boldsymbol{\lambda}$. Subsequently, we have proven that the factors of MF can be transformed into an adjoint LDAext formulation such that the induced personalized rankings of MF and LDAext are identical. Therefore, the adjoint LDAext formulation of an MF allows a simple interpretation of its parameters.

In the practical part, we have evaluated the proposed interpretation of LDAext's parameters on different data sets with the help of statistical tests in 4 experiments. The results of the experiments confirm our hypotheses about the relationships and interpretation of the parameters in LDAext. Although all statistical tests were significant with $p < 10^{-6}$, the effect sizes are small to medium in some experiments. The reason for this is most likely due to the inherent difficulty of collaborative filtering on sparse data sets. For further investigation, we would like to explore our approach on a single recommendation use-case together with domain experts to evaluate whether the adjoint LDAext formulation provides them with new insights.

Our work also contributes to the question about the importance and effectiveness of the scalar product for collaborative filtering tasks. In Neural Collaborative Filtering (NCF), the scalar product in MF is replaced by a learned similarity

with the help of a neural network (NN). Some recent reproducibility papers show that the scalar product outperforms several NCF-based methods and that it should thus be the default choice for combining latent factors (Rendle et al. 2020; Dacrema, Cremonesi, and Jannach 2019). Our work justifies these results in the sense that we can interpret the scalar product as the expected value of LDAext, which describes the underlying dynamics of users interacting with items in a simplified way. This inductive bias is beneficial especially in domains with inherent sparsity like collaborative filtering since learning a scalar product is possible in theory (Lin, Tegmark, and Rolnick 2017) for an NN but proves difficult in practice (Trask et al. 2018; Beutel et al. 2018; Rendle et al. 2020). For these reasons, MF-based methods will continue to be relevant in the future of collaborative filtering and are now also interpretable using the adjoint LDAext formulation.

References

Beutel, A.; Covington, P.; Jain, S.; Xu, C.; Li, J.; Gatto, V.; and Chi, E. H. 2018. Latent Cross: Making Use of Context in Recurrent Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 46–54. Marina Del Rey CA USA: ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:pp. 993–1022.

Dacrema, M. F.; Cremonesi, P.; and Jannach, D. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. *Proceedings of the 13th ACM Conference on Recommender Systems - RecSys '19* 101–109. arXiv: 1907.06902.

Ding, C.; Tao Li; and Jordan, M. 2010. Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(1):45–55.

Harper, F. M., and Konstan, J. A. 2016. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5(4):1–19.

Hernando, A.; Bobadilla, J.; and Ortega, F. 2016. A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. *Knowledge-Based Systems* 97:188–202.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*, 263–272. Pisa, Italy: Ieee.

Koren, Y., and Bell, R. 2015. Advances in collaborative filtering. In Ricci, F.; Rokach, L.; and Shapira, B., eds., *Recommender Systems Handbook*. Boston, MA: Springer US. 77–118.

Koren, Y. 2009. The bellkor solution to the netflix grand prize. *Netflix prize documentation* 81(2009):1–10.

Lee, J.; Sun, M.; and Lebanon, G. 2012. A Comparative Study of Collaborative Filtering Algorithms. *arXiv:1205.3193 [cs, stat]*. arXiv: 1205.3193.

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM Transactions on the Web* 1(1):5.

Lin, H. W.; Tegmark, M.; and Rolnick, D. 2017. Why does deep and cheap learning work so well? *Journal of Statistical Physics* 168(6):1223–1247. arXiv: 1608.08225.

Nikolenko, S. 2015. SVD-LDA: Topic Modeling for Full-Text Recommender Systems. In Pichardo Lagunas, O.; Herrera Alcántara, O.; and Arroyo Figueroa, G., eds., *Advances in Artificial Intelligence and Its Applications*, volume 9414. Cham: Springer International Publishing. 67–79. Series Title: Lecture Notes in Computer Science.

Pan, R.; Zhou, Y.; Cao, B.; Liu, N. N.; Lukose, R.; Scholz, M.; and Yang, Q. 2008. One-Class Collaborative Filtering. In *2008 Eighth IEEE International Conference on Data Mining*, 502–511. Pisa, Italy: IEEE.

Paterek, A. 2007. Improving regularized singular value decomposition for collaborative filtering. *Proceedings of KDD cup and workshop* vol. 2007:pp. 5–8.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. 452–461.

Rendle, S.; Krichene, W.; Zhang, L.; and Anderson, J. 2020. Neural collaborative filtering vs. matrix factorization revisited. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, 240–248. New York, NY, USA: Association for Computing Machinery.

Salakhutdinov, R., and Mnih, A. 2007. Probabilistic Matrix Factorization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, 1257–1264. Red Hook, NY, USA: Curran Associates Inc.

Trask, A.; Hill, F.; Reed, S. E.; Rae, J.; Dyer, C.; and Blunsom, P. 2018. Neural Arithmetic Logic Units. 10.

Wang, C., and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, 448. San Diego, California, USA: ACM Press.

Wilhelm, F. 2021. Matrix factorization for collaborative filtering is just solving an adjoint latent dirichlet allocation model after all. In *Fifteenth ACM Conference on Recommender Systems*, RecSys '21. New York, NY, USA: Association for Computing Machinery.

Xie, W.; Dong, Q.; and Gao, H. 2014. A Probabilistic Recommendation Method Inspired by Latent Dirichlet Allocation Model. *Mathematical Problems in Engineering* 2014:1–10.

Zajac, Z. 2017. Goodbooks-10k: a new dataset for book recommendations. *FastML*.

Zhang, S.; Wang, W.; Ford, J.; and Makedon, F. 2006. Learning from Incomplete Ratings Using Non-negative Matrix Factorization. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, 549–553. Society for Industrial and Applied Mathematics.