MATRIX CO-FACTORIZATION FOR COLD-START RECOMMENDATION

Olivier Gouvert[1] **Thomas Oberlin**[1] **Cédric Févotte**[1]
[1] IRIT, Université de Toulouse, CNRS, France
firstname.lastname@irit.fr

ABSTRACT

Song recommendation from listening counts is now a classical problem, addressed by different kinds of collaborative filtering (CF) techniques. Among them, Poisson matrix factorization (PMF) has raised a lot of interest, since it seems well-suited to the implicit data provided by listening counts. Additionally, it has proven to achieve state-of-the-art performance while being scalable to big data. Yet, CF suffers from a critical issue, usually called cold-start problem: the system cannot recommend new songs, i.e., songs which have never been listened to. To alleviate this, one should complement the listening counts with another modality. This paper proposes a multi-modal extension of PMF applied to listening counts and tag labels extracted from the Million Song Dataset. In our model, every song is represented by the same activation pattern in each modality but with possibly different scales. As such, the method is not prone to the cold-start problem, i.e., it can learn from a single modality when the other one is not informative. Our model is symmetric (it equally uses both modalities) and we evaluate it on two tasks: new songs recommendation and tag labeling.

1. INTRODUCTION

New albums and songs are released every day and are instantly available on streaming platforms. An important issue for streaming companies is therefore to develop recommender systems which are able to handle such new songs [13, 20]. More generally, additional information on those songs is needed to enrich the catalog, allowing the user to efficiently explore and find the songs he might like. In this perspective, tag labeling has proven to be very useful. The labels can be attributed by experts or by the user, and algorithms can complement this information with automatic labeling [7].

For both tasks (song recommendation and tag labeling), matrix factorization (MF) techniques [12, 17], and in particular Poisson MF (PMF), reach significant performance. Unfortunately, these techniques suffer from the well-known cold-start problem: such a recommender system cannot recommend songs which have never been listened to, and similarly it cannot labeled untagged songs. A joint modeling of both modalities can achieve cold-start recommendation, as soon as at least one modality is observed for every song [8, 22].

In this paper, we propose a new matrix co-factorization model based on PMF, which performs those two tasks jointly. Our model is robust to the cold-start problem for both modalities. It can recommend a song which has never been listened to, based on its associate tags. And symmetrically, it can associate tags on a song based on who listened to it. To do that, we separately model the scale (popularity) of each song according to each modality, while the patterns across the topics are shared.

The state of the art of co-factorization techniques is presented in Section 2, along with some background on PMF. Then, in Section 3 we will present our new model and explain its properties. In Section 4, we provide a majorization-minimization (MM) algorithm for solving our optimization problem and underline its scalability. Finally, in Section 5, we test our model on songs recommendation and tag labeling in various settings.

2. RELATED WORKS

In this paper, we will focus on works based on so-called hybrid techniques [1] and Poisson matrix factorization. Note that recommendation tasks can also be addressed with other techniques such as factorization machines [19].

2.1 Poisson matrix factorization

PMF is a non-negative MF (NMF) technique [14]. Let $\mathbf{Y}$ be a matrix of size $F \times I$, where each column represent an item (song) i according to F features. MF approximates the observed matrix $\mathbf{Y}$ by a low-rank product of two matrices: $\mathbf{Y} \approx \mathbf{W}\mathbf{H}^T$, where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ represents a dictionary matrix, and $\mathbf{H} \in \mathbb{R}_+^{I\times K}$ represents a matrix of attributes (activations), with $K \ll \min(F, I)$.

When observed data are in the form of counts, i.e., $\mathbf{Y} \in \mathbb{N}^{F\times I}$, a classical hypothesis is to assume that each observation is drawn from a Poisson distribution:

$$y_{fi} \sim \text{Poisson}([\mathbf{W}\mathbf{H}^T]_{fi}). \quad (1)$$

The maximum likelihood (ML) estimator of $\mathbf{W}$ and $\mathbf{H}$ is therefore obtained by minimizing the cost function de-

fined by:

$$\begin{aligned} C(\mathbf{W},\mathbf{H}) &= -\log p(\mathbf{Y}|\mathbf{W},\mathbf{H}) \\ &= D_{KL}(\mathbf{Y} \mid \mathbf{W}\mathbf{H}^T) + cst \\ &\text{s.t. } \mathbf{W} \geq 0,\ \mathbf{H} \geq 0, \end{aligned} \tag{2}$$

where cst is a constant w.r.t. $\mathbf{W}$ and $\mathbf{H}$, and where D_{KL} is the generalized Kullback-Liebler (KL) divergence defined by:

$$D_{KL}(\mathbf{Y}|\mathbf{X}) = \sum_{f,i} \left(y_{fi} \log \frac{y_{fi}}{x_{fi}} - y_{fi} + x_{fi} \right). \tag{3}$$

This low-rank approximation is known as KL non-negative matrix factorization (KL-NMF) [9, 15].

The cost function C is scale invariant, i.e., for any diagonal non-singular matrix $\Lambda \in \mathbb{R}_+^{K\times K}$, we have $C(\mathbf{W},\mathbf{H}) = C(\mathbf{W}\Lambda^{-1},\mathbf{H}\Lambda)$. To avoid degenerate solutions, a renormalization such that $\sum_f w_{fk} = F$ is often used, where $w_{fk} = [\mathbf{W}]_{fk}$.

Several extensions based on Bayesian formulations have been proposed in the literature [3,5,6,10,17]. In [10], the authors developed a hierarchical Poisson factorization (HPF) by introducing new variables: the popularity of the items and the activity of the users. These variables play a significant role in recommendation tasks.

2.2 Co-factorization

A way of circumventing the cold-start problem is to introduce new modalities [8, 11, 16]. Co-factorization frameworks have been developed to jointly factorize two matrices of observations (two modalities): $\mathbf{Y}^A \approx \mathbf{W}^A(\mathbf{H}^A)^T$ and $\mathbf{Y}^B \approx \mathbf{W}^B(\mathbf{H}^B)^T$, with shared information between the activation matrices: $\mathbf{H}^A \approx \mathbf{H}^B$.

2.2.1 Hard co-factorization

Hard co-factorization [8, 21] posits that the link between activations is an equality constraint: $\mathbf{H}^A = \mathbf{H}^B = \mathbf{H}$. This is equivalent to concatenate the observations $\mathbf{Y}^A$ and $\mathbf{Y}^B$, and the dictionaries $\mathbf{W}^A$ and $\mathbf{W}^B$:

$$\begin{aligned} &D_{KL}(\mathbf{Y}^A|\mathbf{W}^A\mathbf{H}^T) + \gamma D_{KL}(\mathbf{Y}^B|\mathbf{W}^B\mathbf{H}^T) \\ &\quad = D_{KL}\left(\begin{pmatrix} \mathbf{Y}^A \\ \gamma\mathbf{Y}^B \end{pmatrix} | \begin{pmatrix} \mathbf{W}^A \\ \gamma\mathbf{W}^B \end{pmatrix} \mathbf{H}^T \right), \end{aligned} \tag{4}$$

where $\gamma \in \mathbb{R}^+$ is a weighting hyperparameter.

As in Section 2.1, scale invariance issues can be solved by a renormalization step such that: $\sum_u w^A_{uk} + \gamma \sum_v w^B_{vk} = U + V$.

2.2.2 Soft co-factorization

Soft co-factorization [21] relaxes the equality constraint on the activations replacing it by a soft penalty controlled by an hyperparameter $\delta \in \mathbb{R}^+$:

$$\begin{aligned} D_{KL}(\mathbf{Y}^A|\mathbf{W}^A(\mathbf{H}^A)^T) + \gamma D_{KL}(\mathbf{Y}^B|\mathbf{W}^B(\mathbf{H}^B)^T) \\ +\delta\, \mathrm{Pen}(\mathbf{H}^A,\mathbf{H}^B). \end{aligned} \tag{5}$$

A popular choice for this penalty is the ℓ_1-norm: $Pen(\mathbf{H}^A,\mathbf{H}^B) = \left\|\mathbf{H}^A - \mathbf{H}^B\right\|_1$. It is adapted when both modalities are likely to share the same activations, except at some sparse locations where they can differ significantly.

2.2.3 Offset models

Bayesian formulations of the soft co-factorization problem have also been developed through the introduction of an offset latent variable [11,22]. The link between activations is therefore given by:

$$h^B_{ik} = h^A_{ik} + \varepsilon_{ik}, \tag{6}$$

where ε is a latent random variable.

In particular in [11], a co-factorization model is developed based on PMF, with $\varepsilon_{ik} \sim \mathrm{Gamma}(\alpha,\beta)$. This choice is motivated by the conjugacy propriety of the gamma distribution with the Poisson distribution. Nevertheless, the model is not symmetric with respect to (w.r.t.) the activations $\mathbf{H}^A$ and $\mathbf{H}^B$, as $h^B_{ik} > h^A_{ik}$ by construction. Thus, it can solve the cold-start problem only for the modality A and not for B.

3. PROPOSED MODEL

3.1 Notations

In this article, we work with two different modalities. The first modality, denoted by A, corresponds to the listening counts of U users on I songs. The second modality, denoted by B, corresponds to the tags assigned to these I songs, among a set of V tags. $\mathbf{W}^A$ and $\mathbf{W}^B$ thus denote the preferences of users and the atoms of tags across the K patterns, respectively.

3.2 Link between attributes

We propose an equality constraint on normalized activations. We denote by $n^A_i = \sum_k h^A_{ik}$ and $n^B_i = \sum_k h^B_{ik}$, the sum of the rows of the activations. We impose, for each item i:

$$\frac{h^A_{ik}}{n^A_i} = \frac{h^B_{ik}}{n^B_i} = d_{ik}, \tag{7}$$

when $n^A_i > 0$ and $n^B_i > 0$.

- The $I \times K$ matrix $\mathbf{D}$ with entries d_{ik} controls the attributes patterns subject to the constraint $\sum_k d_{ik} = 1$. This information is shared by activations of both modalities. For example, the K patterns can be related to genre information: we expect that experimental rock songs share the same patterns.
- $\mathbf{N}^A = \mathrm{diag}(n^A_i)$ controls the scale of songs across the modality A. It corresponds to the popularity of the song, in the sense that a lot of people listen to it.
- $\mathbf{N}^B = \mathrm{diag}(n^B_i)$ controls the scale of songs across the modality B. It corresponds to the fact that a song can have more or less tag labels.

Two songs can have the same attributes patterns $\mathbf{D}$ but different scales. For example, a song i can be a very popular song, known by a large panel of people: $n_i^A \gg 0$, but lack tag labeling: $n_i^B \approx 0$. On the contrary, another song i can be unpopular (because it is new or not well-received): $n_i^A \approx 0$, but have a lot of tag information (a set of experts may have labeled the song): $n_i^B \gg 0$.

The counterpart of Equation (2) is the following cost function C, which we aim to minimize:

$$\begin{aligned} C(\mathbf{W}^A, \mathbf{W}^B, \mathbf{D}, \mathbf{N}^A, \mathbf{N}^B) \\ = D_{KL}(\mathbf{Y}^A \mid \mathbf{W}^A(\mathbf{N}^A\mathbf{D})^T) \\ + \gamma D_{KL}(\mathbf{Y}^B \mid \mathbf{W}^B(\mathbf{N}^B\mathbf{D})^T) \\ \text{s.t. } \mathbf{W}^A \geq 0,\ \mathbf{W}^B \geq 0,\ \mathbf{D} \geq 0, \\ \operatorname{diag}(\mathbf{N}^A) \geq 0,\ \operatorname{diag}(\mathbf{N}^B) \geq 0. \end{aligned} \tag{8}$$

We denote by $\mathbf{Z} = \{\mathbf{W}^A, \mathbf{W}^B, \mathbf{N}^A, \mathbf{N}^B, \mathbf{D}\}$ the set of variables to infer.

3.3 Scale invariance

Let $\Theta = \operatorname{diag}(\theta_i)$ be a diagonal matrix of size $I \times I$ with non-negative entries. We have the following scale invariance:

$$\begin{aligned} C(\mathbf{W}^A, \mathbf{W}^B, \Theta^{-1}\mathbf{D}, \mathbf{N}^A\Theta, \mathbf{N}^B\Theta) \\ = C(\mathbf{W}^A, \mathbf{W}^B, \mathbf{D}, \mathbf{N}^A, \mathbf{N}^B). \end{aligned} \tag{9}$$

This scale invariance allows us to impose the constraint on $\mathbf{D}$, described in Section 3, by applying a renormalization step (see Section 4.2).

Let $\Lambda = \operatorname{diag}(\lambda_k)$ be a diagonal matrix of size $K \times K$ with non-negative entries, $\bar{\mathbf{W}}^A = \mathbf{W}^A\Lambda^{-1}$, $\bar{\mathbf{W}}^B = \mathbf{W}^B\Lambda^{-1}$ and $\bar{\mathbf{D}} = \mathbf{D}\Lambda$. We also have the following scale invariance:

$$\begin{aligned} C(\bar{\mathbf{W}}^A, \bar{\mathbf{W}}^B, \bar{\mathbf{D}}, \mathbf{N}^A, \mathbf{N}^B) \\ = C(\mathbf{W}^A, \mathbf{W}^B, \mathbf{D}, \mathbf{N}^A, \mathbf{N}^B). \end{aligned} \tag{10}$$

In practice, this invariance is not an issue and we do not apply a renormalization step. However, this kind of invariance plays a role for the scores used in recommendation as discussed in Section 3.4.

3.4 Recommendation tasks

In recommender systems, a classical problem is to propose a ranked list of songs, users or tags. We develop how to construct this list on two tasks: in- and out-prediction.

3.4.1 In-matrix recommendation

In-matrix recommendation is a task of recommendation on users and items which do not suffer from the cold-start problem. For in-matrix recommendation, we propose a ranked list of songs for each user, based on the score defined by:

$$s_{ui}^A = \sum_k w_{uk}^A h_{ik}^A. \tag{11}$$

This score and our cost function C have the same scale invariance described in Eq. 10.

3.4.2 Cold-start (out-matrix) recommendation

Cold-start (or out-matrix) recommendation is a task of recommendation on items which suffer from the cold-start problem (on modality A or B). In this section, we take the example of a cold-start problem on modality A, i.e., the song has no information in the modality A (nobody has listened to this song yet) but has tags associated to it. The following remark would hold for a cold-start problem on modality B.

For cold-start (out-matrix) recommendation the score is defined by:

$$s_{ui}^A = \sum_k w_{uk}^A d_{ik} = \sum_k w_{uk}^A \frac{h_{ik}^A}{\sum_l h_{il}^A}. \tag{12}$$

Contrary to in-prediction, we use $\mathbf{D}$ and not $\mathbf{H}^A = \mathbf{N}^A\mathbf{D}$ since the popularity in the modality A is close to zero for songs with no information, i.e., $n_i^A \approx 0$.

This score and the cost function C do not have the same scale invariance described in Eq. 10. In fact, if we denote $\bar{w}_{uk}^A = \lambda_k w_{uk}^A$ and $\bar{h}_{ik}^A = \lambda_k h_{ik}^A$, we have:

$$\bar{s}_{ui}^A = \sum_k \bar{w}_{uk}^A \frac{\bar{h}_{ik}^A}{\sum_l \bar{h}_{il}^A} = s_{ui}^A \frac{\sum_k h_{ik}^A}{\sum_k \lambda_k h_{ik}^A} = s_{ui}^A c_i, \tag{13}$$

where $c_i = \frac{\sum_k h_{ik}^A}{\sum_k \lambda_k h_{ik}^A}$.

This means that, if we want to rank the different scores s_{ui}^A, we have to do it for a fixed item. Therefore, to properly evaluate the cold-start problem for songs, we will propose a ranked list of users (or tags), for a given item.

For a streaming company, it corresponds to obtaining a ranked list of users which are likely to listen to this new song, or a ranked list of tags which corresponds to the song.

4. OPTIMIZATION

4.1 Auxiliary function

The objective function C has no closed-form minimum and is not convex. We use a MM algorithm [9] to reach a local minimum. The MM algorithms start by designing a majorizing surrogate G of the objective function $C(\mathbf{Z}) \leq G(\mathbf{Z} \mid \tilde{\mathbf{Z}})$ which is tight at the current value $\tilde{\mathbf{Z}}$, i.e., $C(\tilde{\mathbf{Z}}) = G(\tilde{\mathbf{Z}} \mid \tilde{\mathbf{Z}})$.

We use Jensen inequality on terms of the form $\log(\sum_i x_i)$. We define:

$$\phi_{uik}^A = \frac{\tilde{w}_{uk}^A \tilde{d}_{ik}}{\sum_k \tilde{w}_{uk}^A \tilde{d}_{ik}},\ c_{uik}^A = y_{ui}^A \phi_{uik}^A, \tag{14}$$

$$\phi_{uik}^B = \frac{\tilde{w}_{uk}^B \tilde{d}_{ik}}{\sum_k \tilde{w}_{uk}^B \tilde{d}_{ik}},\ c_{uik}^B = y_{ui}^B \phi_{uik}^B. \tag{15}$$

It leads to the following upper-bound:

$$G(\mathbf{Z} \mid \tilde{\mathbf{D}}, \tilde{\mathbf{W}}^A, \tilde{\mathbf{W}}^B) \quad (16)$$
$$= \sum_{uik} \left[-c^A_{uik} \log(w^A_{uk} n^A_i d_{ik}) + w^A_{uk} n^A_i d_{ik} \right]$$
$$+ \gamma \sum_{vik} \left[-c^B_{vik} \log(w^B_{vk} n^B_i d_{ik}) + w^B_{vk} n^B_i d_{ik} \right] + cst.$$

4.2 Updates

The auxiliary function G can be optimized by using a block descent algorithm. At each iteration, we optimize one latent variable, keeping all the others fixed. This technique leads to four update rules described in the following.

- Variables $\mathbf{W}^A$ and $\mathbf{W}^B$:

$$w^A_{uk} \leftarrow \frac{\sum_i c^A_{uik}}{\sum_i n^A_i d_{ik}}; \; w^B_{vk} \leftarrow \frac{\sum_i c^B_{vik}}{\sum_i n^B_i d_{ik}} \quad (17)$$

- Variables $\mathbf{N}^A$ and $\mathbf{N}^B$:

$$n^A_i \leftarrow \frac{\sum_u y^A_{ui}}{\sum_{uk} w^A_{uk} d_{ik}}; \; n^B_i \leftarrow \frac{\sum_v y^B_{vi}}{\sum_{vk} w^B_{vk} d_{ik}} \quad (18)$$

- Variable $\mathbf{D}$:

$$d_{ik} \leftarrow \frac{\sum_u c^A_{uik} + \gamma \sum_v c^B_{vik}}{n^A_i \sum_u w^A_{uk} + \gamma n^B_i \sum_v w^B_{vk}} \quad (19)$$

As discussed in Section 3.3, we add a renormalization step at the end of each iteration. The update is as follows:

$$\theta_i = \sum_k d_{ik} / I, \quad (20)$$
$$\mathbf{D} \leftarrow \Theta^{-1} \mathbf{D}; \; \mathbf{N}^A \leftarrow \mathbf{N}^A \Theta; \; \mathbf{N}^B \leftarrow \mathbf{N}^B \Theta. \quad (21)$$

4.3 Algorithm

The complete algorithm is summarized in Algorithm 1. Note that the inference only requires browsing the non-zero data $y^A_{ui} > 0$ and $y^B_{vi} > 0$, during the update of the local variables c^A_{uik} and c^B_{uik}. Hence, our algorithm has the same scalability as PMF, making it particularly well-suited for processing huge sparse matrices, as it is the case in recommender systems (see Table 1).

The algorithm is stopped when the relative increment of the cost function C is lower than a chosen parameter τ.

Algorithm 1: MM Algorithm

Input : $\mathbf{Y}^A, \mathbf{Y}^B, K, \gamma$

Initialize: $\mathbf{W}^A, \mathbf{W}^B, \mathbf{N}^A, \mathbf{N}^B, \mathbf{D}$

repeat

- for each pair (u, i) such that $y^A_{ui} > 0$: Eq. 14
- for each pair (v, i) such that $y^B_{vi} > 0$: Eq. 15
- for each user u and tag v: Eq. 17
- for each item i: Eq. 18-19
- normalization step: Eq. 21

until C *converges*;

5. EXPERIMENTS

5.1 Experimental Setup

5.1.1 Datasets

We use two datasets extracted from the Million Song Dataset (MSD) [2] and merge them on songs:

- The Taste Profile dataset provides listening counts of 1M users on 380k songs [18]. We select a subset of the users and pre-process the data to remove users and items with few information [16]. We keep only users who listened to at least 20 songs, and songs which have been listened to by at last 20 users.

	Taste Profile	**Last.fm**
# columns (songs)	15,667	15,667
# rows (users or tags)	16,203	620
# non-zeros	792,761	128,652
% non-zeros	0.31%	1.32%

Table 1. Datasets structure after pre-processing.

- The Last.fm dataset provides tag labels for around 500k songs. These tags were extracted from the Last.fm API [4]. Since the tags were collected via user annotation, they are quite noisy. To avoid miss-labeling in the train data, we pre-process it. We keep only the 1000 most used tags in the whole dataset. For each couple song-tag, a confidence rating is given by Last.fm, we keep only couples with confidence higher than 10. Finally, we keep only tags which appears at least in 20 songs. The top 10 of the tags in the dataset after the pre-processing are shown in Table 2.

We binarize the two datasets. Structure of both datasets is described in Table 1.

5.1.2 Evaluation metric: ranking prediction

In each experiment, we will propose a ranked list $\mathcal{L}$ of N items (which can be songs, tags or users) and evaluate its quality w.r.t. a ground-truth relevance. For this, we calculate the discounted cumulative gain (DCG) and its normalized version, the NDCG:

Tags	**Occ.**	**Tags**	**Occ.**
rock	6703	electronic	2413
alternative	4949	female vocalists	2407
indie	4151	indie rock	2171
pop	3853	Love	1875
alternative rock	2854	singer-songwriter	1786

Table 2. Occurences (Occ.) of the top tags in the dataset after pre-processing.

Experiment	OUT-A		OUT-B		IN-A	
Score	NDCG@20	NDCG@200	NDCG@1*	NDCG@10	NDCG@100	NDCG**
P-coNMF	0.0824 $\pm 1.48e^{-5}$	0.122 $\pm 1.33e^{-5}$	0.416 $\pm 5.85e^{-4}$	0.266 $\pm 1.59e^{-4}$	0.129 $\pm 4.24e^{-6}$	0.286 $\pm 2.82e^{-6}$
H-coNMF	0.0873 $\pm 1.39e^{-5}$	0.131 $\pm 2.21e^{-5}$	0.391 $\pm 1.73e^{-4}$	0.264 $\pm 1.00e^{-4}$	0.122 $\pm 5.72e^{-6}$	0.283 $\pm 2.96e^{-6}$
KL-NMF	.	.	.	.	0.163 $\pm 5.36e^{-7}$	0.313 $\pm 1.50e^{-7}$

Table 3. Performance of three models: P-coNMF, H-coNMF, KL-NMF, on three different tasks: out-matrix song recommendation (OUT-A), tag labeling (OUT-B), in-matrix recommendation (IN-A). Each algorithm is run 5 times, the mean and the variance of the NDCG metrics are displayed. * NDCG@1 corresponds to the percentage of success on the first predicted tag. ** NDCG is not truncated in this column, it is equivalent to chose $N = I$.

$$\text{DCG@N} = \sum_{n=1}^{N} \frac{\text{rel}(n)}{\log_2(n+1)}, \quad (22)$$

$$\text{NDCG@N} = \frac{\text{DCG@N}}{\text{IDCG@N}}, \quad (23)$$

where $\text{rel}(n)$ is the ground-truth relevance of the n-th item in the list $\mathcal{L}$. In the following, $\text{rel}(n) = 1$ if the item is relevant and $\text{rel}(n) = 0$ if not.

The denominator of the DCG penalizes relevant items which are at the end of the ranked list. It accounts for the fact that a user will only browse the beginning of the list, and will not pay attention to items which are ranked at the end. IDCG is the ideal DCG. It corresponds to the DCG score of an oracle which ranks perfectly the list, thus scaling the NDCG between 0 and 1.

5.1.3 Compared methods

For each experiment, we will compare the performance of our model, proportional co-factorization NMF (P-coNMF) with two other methods:

- KL-NMF, presented in Section 2.1. It can only be used for in-matrix prediction as it suffers from the cold-start problem.
- Hard co-factorization (H-coNMF), presented in Section 2.2.1), that use KL-NMF algorithm on concatenated matrix. For out-matrix prediction, we will use a mask that indicates what columns are missing. The objective function is then:

$$C(\mathbf{W}, \mathbf{H}) = D_{KL}(\mathbf{X} \otimes \mathbf{Y} \mid \mathbf{X} \otimes \mathbf{W}\mathbf{H}^T), \quad (24)$$

  where $\otimes$ is the elementwise multiplication, and $\mathbf{X}$ is the mask. Note that the masked H-coNMF is expected to perform as good as soft coNMF with the ℓ_1-norm, since it does not enforce common activation for new songs.

For both methods, we chose $K = 100$ latent factors. The hyperparameter is set such that $\gamma = \frac{U}{V}$, which allows to compensate for the size difference between the two datasets ($V \ll U$).

5.2 Cold-start recommendation

In this section, we evaluate our algorithm on cold-start recommendation tasks for both modalities A and B. For this, we artificially replace columns of $\mathbf{Y}^A$ and $\mathbf{Y}^B$ by columns full of zeros, in order to create the train datasets $\mathbf{Y}^A_{\text{train}}$ and $\mathbf{Y}^B_{\text{train}}$. It leads to 10% of songs with only listening counts information, 10% of songs with only tag information and 80% of songs with both informations. The removed columns form the test datasets $\mathbf{Y}^A_{\text{test}}$ and $\mathbf{Y}^B_{\text{test}}$.

For each song among the never-listened-to songs, we want to find a set of users that is likely to listen to it. We train all the algorithms on $\mathbf{Y}^A_{\text{train}}$ and $\mathbf{Y}^B_{\text{train}}$. For each song, we create a ranked list of users based on the score defined in Section 3.4.2. We evaluate its relevance based on the NDCG metrics with ground-relevance defined by: $rel(u, i) = \mathbb{1}(y^A_{\text{test},ui} > 0)$, where $\mathbb{1}(x)$ is the indicator function which is equal to 1 when x is true and 0 otherwise.

Similarly, for each song among the untagged songs, we want to find a set of tags that can annotate that song. Then we propose a ranked list of tags and calculate the NDCG score with ground-relevance defined by: $rel(v, i) = \mathbb{1}(y^B_{\text{test},vi} > 0)$.

The columns OUT-A and OUT-B of Table 3 present the results of P-coNMF and H-coNMF on the two cold-start problems. For recommending potential listeners (OUT-A), H-coNMF seems to be slightly better than our method. However, P-coNMF outperforms H-coNMF on tag labeling task. P-coNMF presents a success rate of 42% on the first predicted tag. This is an acceptable rate since the tag dataset is noisy: it has not been labeled by experts but by users and presents some incoherences. For example, the tag 'Hip-Hop' can also be written 'hip hop'. More details on tag labeling are provided in Section 5.4. Contrary to H-coNMF, P-coNMF does not need a mask to know which columns are missing. Additionally, the scale variables $\mathbf{N}^A$ and $\mathbf{N}^B$ are able to explain different scalings of the same song in the two datasets. This seems interesting because the amount of listening counts and tags for the same song is often highly different.

	FACTOR #94	FACTOR #29	FACTOR #30
Top tags	Hip-Hop hip hop classic rap Gangsta Rap	new wave post-punk Guilty Pleasures intense Post punk	experimental Experimental Rock Avant-Garde noise weird
Top songs based on $\mathbf{H}^A$	Eminem - "Mockingbird" Eminem - "Without Me" Kid Cudi - "Day 'N' Nite" Kid Cudi - "Up Up & Away" Kid Cudi - "Cudi Zone"	The Cure - "Boys Don't Cry" The Smiths - "There Is A Light [...]" The Smiths - "This Charming Man" The Smiths - "What Difference Does It Make?" Wolfsheim - "Once In A Lifetime"	Animal Collective - "Fireworks" Sigur Ros - "Staralfur" Sonic Youth - "Youth Against Fascism" Grizzly Bear - "Little Brother" TV On The Radio - "Crying"
Top songs based on $\mathbf{D}$	DMX - "Where The Hood At" Lil Jon - "Crunk Juice" 50 Cent - "Straight To The Bank" Eminem - "The Kiss" The Notorious B.I.G. - "Respect"	New Order - "The Perfect Kiss" Talking Heads - "Burning Down The House" Joy Division - "Disorder" Tears For Fears - "Goodnight Song" The Smiths - "Miserable Lie"	The Mars Volta - "Tira Me a Las Aranas" Cocorosie - "Gallows" The Mars Volta - "Concertina" The Mars Volta - "Roulette Dares" TV On The Radio - "Golden Age"

Table 4. Three examples of factors, with, for both of them, the 5 top tags associated to it, the 5 top songs associated to it, with or without the notion of popularity.

5.3 In-matrix song recommendation

We also evaluate our algorithm on in-matrix prediction. The goal is therefore to predict which songs a user is likely to listen. There is no cold-start recommendation here, and KL-NMF can be trained.

We artificially split the listening counts dataset in two. 20% of non-zero values of $\mathbf{Y}^A$ are removed to create the test set $\mathbf{Y}^A_{\text{test}}$. The 80% remaining form the train set $\mathbf{Y}^A_{\text{train}}$ on which the different models are trained. Each method is evaluated with NDCG metric. For each user, a list of songs is proposed based on the score defined in Section 3.4.1, among the songs he never listened to. The ground-truth relevance is defined by $rel(u,i) = \mathbb{1}(y^A_{\text{test},ui} > 0)$.

The results are presented in the third column (IN-A) of Table 3. P-coNMF is slightly better than H-coNMF, but we observe that KL-NMF achieves state-of-the-art performance. This is not surprising, since adding information on another modality (tags here) can be viewed as a regularizing term. We lose in precision in in-matrix recommendation task but we solve the cold-start problem. This seems an interesting trade-off.

5.4 Exploratory Analysis

In Table 4, we present for each of the three factors $k \in \{29, 30, 94\}$:

- in the first row, the tags which corresponds to the five highest values of $\mathbf{W}^B$.
- in the second row, the songs which corresponds to the five highest values of $\mathbf{H}^A = \mathbf{N}^A\mathbf{D}$.
- in the third row, the songs which corresponds to the five highest values of $\mathbf{D}$.

The top tags associated to each factor are consistent: for example, genre as 'new wave' and 'post-punk' are in the same factor. The model is also robust to the different spellings used by the users ('post-punk' and 'Post punk' for example). Then, we see that the top songs in each factor are related with the top tags. Eminem, 50 Cent and The Notorious B.I.G. are rap artists. The Cure, The Smiths and Joy Division are the leading figures of the new wave. TV On The Radio, The Mars Volta and Animal Collective are known to be experimental rock bands. Finally, we see that the popularity of songs $\mathbf{N}^A$ has an important influence on the diversity of the top songs in each factor. When this notion is removed (last row of the table), less popular songs and bands appear in the top songs.

6. CONCLUSION

In this paper, we proposed a new Poisson matrix co-factorization, in which the attributes of each modality are assumed proportional. Contrary to hard and ℓ_1-based soft co-factorization, in this new model each item may have different scaling (or popularity) in each modality. This is of particular interest when tackling cold-start recommendation, in which one scaling is close to zero. The benefits of the algorithm over standard co-factorization have been illustrated for song recommendation, with emphasis placed on cold-start situations.

This raised interesting short-term perspectives, such as the derivation of more involved Bayesian models, and inference or extensions to different, possibly non-binary datasets. Future works should also consider datasets with highly different dimensions or dynamics, by means of a tri-factorization.

7. ACKNOWLEDGMENTS

This work has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program under grant agreement No 681839 (project FACTORY).

8. REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender Systems Handbook*, pages 191–226. Springer, 2015.

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[3] John Canny. GaP: A factor model for discrete data. In *Proc. International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 122–129, 2004.

[4] Òscar Celma. Music recommendation. In *Music recommendation and discovery*, pages 43–85. Springer, 2010.

[5] Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

[6] O. Dikmen and C. Fevotte. Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10):5163–5175, 2012.

[7] Douglas Eck, Paul Lamere, Thierry Bertin-Mahieux, and Stephen Green. Automatic generation of social tags for music recommendation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 385–392, 2008.

[8] Yi Fang and Luo Si. Matrix co-factorization for recommendation with rich side information and implicit feedback. In *Proc. International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec)*, pages 65–69, 2011.

[9] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[10] Prem Gopalan, Jake M. Hofman, and David M. Blei. Scalable recommendation with hierarchical Poisson factorization. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 326–335, 2015.

[11] Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3176–3184. 2014.

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

[13] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In *Proc. International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 208–211, 2008.

[14] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[15] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562. 2001.

[16] Dawen Liang, Minshu Zhan, and Daniel PW Ellis. Content-aware collaborative music recommendation using pre-trained neural networks. In *Proc. International Society for Music Information Retrieval (ISMIR)*, pages 295–301, 2015.

[17] Hao Ma, Chao Liu, Irwin King, and Michael R. Lyu. Probabilistic factor models for web site recommendation. In *Proc. International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 265–274, 2011.

[18] Brian McFee, Thierry Bertin-Mahieux, Daniel PW Ellis, and Gert RG Lanckriet. The million song dataset challenge. In *Proc. International Conference on World Wide Web (WWW)*, pages 909–916, 2012.

[19] S. Rendle. Factorization machines. In *Proc. International Conference on Data Mining (ICDM)*, pages 995–1000, 2010.

[20] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proc. International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 253–260, 2002.

[21] N. Seichepine, S. Essid, C. Févotte, and O. Cappé. Soft nonnegative matrix co-factorization. *IEEE Transactions on Signal Processing*, 62(22):5940–5949, 2014.

[22] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 448–456, 2011.