Non-negative Multiple Matrix Factorization

Koh Takeuchi, Katsuhiko Ishiguro, Akisato Kimura, and Hiroshi Sawada
NTT Communication Science Laboratories, Kyoto 619-0237, Japan
{takeuchi.koh, ishiguro.katsuhiko, hiroshi.sawada}@lab.ntt.co.jp, akisato@ieee.org

Abstract

Non-negative Matrix Factorization (NMF) is a traditional unsupervised machine learning technique for decomposing a matrix into a set of bases and coefficients under the non-negative constraint. NMF with sparse constraints is also known for extracting reasonable components from noisy data. However, NMF tends to give undesired results in the case of highly sparse data, because the information included in the data is insufficient to decompose. Our key idea is that we can ease this problem if complementary data are available that we could integrate into the estimation of the bases and coefficients. In this paper, we propose a novel matrix factorization method called Non-negative Multiple Matrix Factorization (NM2F), which utilizes complementary data as auxiliary matrices that share the row or column indices of the target matrix. The data sparseness is improved by decomposing the target and auxiliary matrices simultaneously, since auxiliary matrices provide information about the bases and coefficients. We formulate NM2F as a generalization of NMF, and then present a parameter estimation procedure derived from the multiplicative update rule. We examined NM2F in both synthetic and real data experiments. The effect of the auxiliary matrices appeared in the improved NM2F performance. We also confirmed that the bases that NM2F obtained from the real data were intuitive and reasonable thanks to the non-negative constraint.

1 Introduction

Non-negative matrix factorization (NMF) [Lee and Seung, 1999] is a matrix factorization method that is widely used in various fields including audio processing [Smaragdis and Brown, 2003], text mining [Xu *et al.*, 2003], image analysis [Lee and Seung, 1999; Hoyer, 2004] and brain signal analysis [Cichocki *et al.*, 2009]. The non-negative constraint of NMF is known to reveal intuitive and reasonable factorization results in many applications. NMF with sparse constraint [Hoyer, 2004; Cemgil, 2009] provides good performance for decomposing noisy data [Abdallah and Plumbley, 2004; Dikmen and Févotte, 2012]. In general, however, matrix factorization methods including NMF perform poorly when the data to be factorized are very sparse and not sufficiently informative [Aharon *et al.*, 2006].

Recently, the variety of sensored data and human generated data is increasing greatly. An event can be observed as, for example, sounds, movies, geo-locations, text messages, annotations, and action histories. These features can often be high dimensional, and the problem of data sparseness is not negligible [Blei *et al.*, 2003; Koren *et al.*, 2009; Lin *et al.*, 2011]. One convincing solution for this problem is to combine a different set of data and analyze them simultaneously. We expect these different data to be correlated with each other because they are different representations of real world events. Thus the analysis of these data could be improved by utilizing another set of data as complementary information, and vice versa.

Our key idea is that the factorization of the target data could be improved if complementary data are available. Actually, the effectiveness of this approach has been confirmed in item recommendation tasks based on collaborative filtering. In that field, matrix factorization methods have been used in many studies such as [Koren *et al.*, 2009]. To deal with sparsity, some researchers (e.g. [Ma *et al.*, 2008; Wang and Blei, 2011; Purushotham *et al.*, 2012]) increase the available information for factorization by adding auxiliary data related to the target data matrix. These methods improved recommendation accuracy, however, these previous studies ignored the non-negative constraint. As described above, the non-negative constraint plays an essential role for obtaining intuitive and understandable data decomposition results.

In this paper, we propose a novel matrix factorization method called Non-negative Multiple matrix factorization (NM2F) that utilizes complementary data for non-negative matrix factorization. NM2F integrates complementary data as auxiliary matrices that share rows or columns with the target data matrix. NM2F decomposes all matrices simultaneously, and this greatly improves the decomposition performance under sparse data scenarios. In contrast to previous work on collaborative filtering, NM2F is formulated under the non-negative constraint. Therefore we can expect more intuitive and understandable decomposition results. In this paper, we formulate NM2F as a generalization of the original NMF and derive parameter estimation rules to

obtain a local optimal solution. Moreover, we study deeper interpretations of NM2F as a probabilistic model and NMF with undefined regions. Experimental results revealed the good performance of the proposed NM2F under sparse data situations, both in synthetic and real-world data sets.

The rest of this paper is organized as follows. After reviewing related work in Section 2, we formulate our problem and introduce the proposed model in Section 3. In Section 4, we evaluate the experimental results of our proposed method, and Section 5 concludes the paper.

2 Related Work

Non-negative Matrix Factorization (NMF) [Lee and Seung, 1999] is an unsupervised matrix factorization method. NMF is designed to minimize the loss (distance) between a non-negative observed data matrix and its low rank decomposition. In this decomposition, the observed data matrix is represented as the weighted linear sum of bases with a non-negative constraint. The non-negative constraint of decomposed bases and coefficients make the parameter estimation non-convex. Instead, we obtain sparse bases and coefficients that enable us to easily understandable their meaning in many applications.

There are many kinds of loss employed in NMF such as Euclidean distance, generalized Kullback-Leibler divergence, and Itakura-Saito distance [Cichocki *et al.*, 2009] which belong to the β divergence family. NMF with generalized Kullback-Leibler divergence has been proved to be equal to Probabilistic Latent Semantic Analysis (pLSA)[Hofmann, 1999; Ding *et al.*, 2006]. There are also a number of model estimation methods yielding a local optima of NMF, such as the multiplicative update rule [Lee *et al.*, 2000], Alternating Least Squares (ALS) [Cichocki *et al.*, 2007], and Gibbs sampling [Schmidt *et al.*, 2009]. Non-parametric Bayes extensions of NMF have also been proposed including Ga-P NMF [Hoffman *et al.*, 2011] and iFiHMM [Nakano *et al.*, 2011]. In many applications, NMF exhibits good performance and extracts understandable patterns thanks to its non-negativity [Cai *et al.*, 2011; Liu *et al.*, 2012]. Furthermore, the calculation could be safely accelerated with distributed processing [Liu *et al.*, 2010] and online learning methods [Cao *et al.*, 2007; Wang *et al.*, 2011]. In this paper, we formulate NM2F as the parameter estimation of generalized Kullback-Leibler divergence. We show a multiplicative rule for NM2F that guarantees the convergence of local optima.

Many matrix factorization techniques have been used in item recommendation tasks (e.g. [Koren *et al.*, 2009]). Among them, the Probabilistic Matrix Factorization (PMF) method developed by [Salakhutdinov and Mnih, 2008] has become a seminal technique in the field. PMF has been extended in two ways. One incorporates topic information using LDA [Wang and Blei, 2011]. The other augments PMF using link information available from social networks [Ma *et al.*, 2008; Mei *et al.*, 2008; Noel *et al.*, 2012]. In one recent paper [Purushotham *et al.*, 2012] the authors combine both ingredients in their model. Also, related models have been used in the document analysis research [Steyvers *et al.*, 2004]. Our model is closely related to these researches in a technical sense. However, the target domains and goals are very different. As explained in Section 1, these recommendation studies often ignore the non-negative constraint. In this paper, we propose a matrix factorization method with auxiliary matrices under the non-negative constraint for intuitive and reasonable decomposition results.

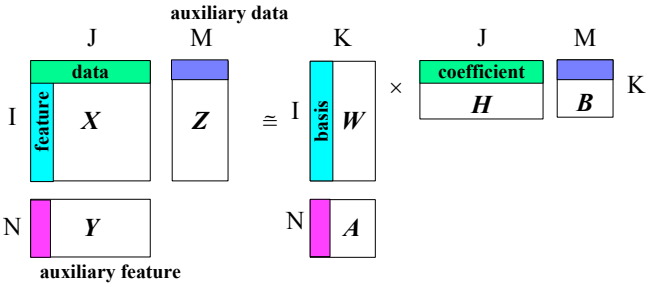

Figure 1: The target matrix $\boldsymbol{X}$ is factorized to the bases $\boldsymbol{W}$ and the coefficients $\boldsymbol{H}$ with the row wise auxiliary matrix $\boldsymbol{Y}$ and the column wise auxiliary matrix $\boldsymbol{Z}$.

3 Non-negative Multiple Matrix Factorization

3.1 Problem Formulation

Before describing the model, let us explain data structure, notations, and an overview the problem.

Our proposed method defines three matrices called the target matrix, row wise auxiliary matrix and column wise auxiliary matrix. Let us denote an index of features as $i \in \{1 \cdots I\}$ and an index of data as $j \in \{1, \cdots, J\}$. Let $x_{i,j}$ be the observed value of feature i in the j-th data. Then let $\boldsymbol{X} = \{x_{i,j}\} \in \mathbb{R}_+^{I \times J}$ be the target matrix. Let us denote an index of auxiliary features as $n \in \{1, \cdots, N\}$. Then $y_{n,j}$ denotes the auxiliary observed value of feature n in the j-th data. Then we denotes $\boldsymbol{Y} = \{y_{n,j}\} \in \mathbb{R}_+^{N \times J}$ as the row-wise auxiliary matrix. Let us denote an index of auxiliary data as $m \in \{1, \cdots, M\}$. Let $z_{i,m}$ be the observed value of feature i in the m-th auxiliary data. Then let $\boldsymbol{Z} = \{z_{i,m}\} \in \mathbb{R}_+^{I \times M}$ be the column-wise auxiliary matrix. Figure 1 shows matrices $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$.

The aim of this paper is to estimate appropriate bases and coefficients. Our method takes auxiliary matrices $\boldsymbol{Y}$ and $\boldsymbol{Z}$ into account when seeking bases and coefficients. Furthermore, our method simultaneously estimates patterns of $\boldsymbol{Y}$ and $\boldsymbol{Z}$. For example, in Section 4, we undertake an experiment to factorize a matrix $\boldsymbol{X}$ consisting of the number of users listening artists with $\boldsymbol{Y}$ comprising tags annotated on the artists and $\boldsymbol{Z}$ consisting of information about a user's network of friends.

3.2 Proposed Framework

This subsection describes our proposal, Non-negative Multiple Matrix Factorization (NM2F). NM2F factorizes the observed matrix into a set of bases and coefficients. This method requires the data, the bases and the coefficient matrix to be non-negative. Non-negative constraints yield sparse solutions for the bases and the coefficient matrix, which are preferable for interpretation.

The key idea of our method is to augment the available information by employing auxiliary matrices. NM2F employs two types of auxiliary data in order to factorize target data $\boldsymbol{X}$. The row-wise auxiliary matrix $\boldsymbol{Y}$ is assumed to share coefficients matrix $\boldsymbol{H}$ with $\boldsymbol{X}$, and the column-wise auxiliary matrix $\boldsymbol{Z}$ is assumed to share bases matrix $\boldsymbol{W}$ with $\boldsymbol{X}$.

Let K be the number of bases. Let $w_{i,k}$ be the value of the feature i in the k-th basis. We denote $\boldsymbol{W} = \{w_{i,k}\} \in \mathbb{R}_+^{I\times K}$ as the basis matrix. Let $h_{k,j}$ be the coefficient of the k-th basis in the j-th data. We denote the coefficient matrix as $\boldsymbol{H} = \{h_{k,j}\} \in \mathbb{R}_+^{K\times J}$. Let $a_{n,k}$ be the value of the k-th basis for auxiliary feature n. We denote $\boldsymbol{A} = \{a_{n,k}\} \in \mathbb{R}_+^{N\times K}$ as the auxiliary basis matrix. Let $b_{k,m}$ be the coefficient of the k-th basis in the m-th auxiliary data. We denote auxiliary coefficient matrix as $\boldsymbol{B} = \{b_{k,m}\} \in \mathbb{R}_+^{K\times M}$.

NM2F minimizes the divergence between the given matrices $\{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ and their factored matrices $\{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}\}$ with scaling parameters α, β:

$$\begin{aligned}&\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{A},\boldsymbol{B}} \mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta)\\ &s.t.\ \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B} \geq 0,\ \alpha, \beta \geq 0,\end{aligned} \tag{1}$$

where $\mathcal{D}$ is a divergence. Let us denote $\hat{x}_{i,j}$,$\hat{y}_{n,j}$, and $\hat{z}_{i,m}$ as the approximated values of $x_{i,j}$, $y_{n,j}$, and $z_{i,m}$ obtained from sums of bases weighted by coefficients:

$$\hat{x}_{i,j} = \sum_{k=1}^{K} w_{i,k}h_{k,j},\ \hat{y}_{n,j} = \sum_{k=1}^{K} a_{n,k}h_{k,j},\ \hat{z}_{i,m} = \sum_{k=1}^{K} w_{i,k}b_{k,m}. \tag{2}$$

Then a divergence between the model and data is defined as

$$\begin{aligned}&\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta)\\ &= \mathcal{D}(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}) + \alpha\mathcal{D}(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{H}) + \beta\mathcal{D}(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{B})\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} d(x_{i,j}|\hat{x}_{i,j}) + \alpha\sum_{n=1}^{N}\sum_{j=1}^{J} d(y_{n,j}|\hat{y}_{k,j})\\ &+ \beta\sum_{i=1}^{I}\sum_{m=1}^{M} d(z_{i,m}|\hat{z}_{i,m}),\end{aligned} \tag{3}$$

where α and β satisfying $\alpha \geq 0$, $\beta \geq 0$ are scaling parameters to balance three divergences, and $d(\cdot)$ specifies an element-wise divergence. We employ the generalized Kullback-Leibler divergence (d_{gKL}) in our model.

$$d_{gKL}(x_{i,j}|\hat{x}_{i,j}) = x_{i,j}\log\frac{x_{i,j}}{\hat{x}_{i,j}} - x_{i,j} + \hat{x}_{i,j}. \tag{4}$$

The generalized Kullback-Leibler divergence is widely applied to NMF, but other distances such as the Euclidean distance and Itakura-Saito distance could be used as alternatives.

3.3 NMF Framework

The goal of NMF is to seek a basis matrix $\boldsymbol{W}$ and a coefficient matrix $\boldsymbol{H}$ that minimize the distance between the estimated parameters $\boldsymbol{W}, \boldsymbol{H}$ and the target matrix $\boldsymbol{X}$, subject to $\boldsymbol{W}, \boldsymbol{H} \geq 0$.

$$\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{D}(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H})\ s.t.\ \boldsymbol{W}, \boldsymbol{H} \geq 0 \tag{5}$$

If we set $\alpha = 0, \beta = 0$ in Eq. (3), we have

$$\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta) = \mathcal{D}(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}). \tag{6}$$

Therefore, NMF is regarded as a special case of NM2F.

3.4 Parameter Estimation

We derive multiplicative update rules of NM2F, similar to the case of NMF. The derivations of the parameter estimation method are detailed in Appendix A. In summary, we obtain the local minimum of Eq. (4) by iterating update rules of $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{A}$, and $\boldsymbol{B}$ as shown below:

$$w_{i,k}^{new} = w_{i,k}\frac{\left(\sum_{j\in\mathcal{J}_i}\frac{x_{i,j}}{\hat{x}_{i,j}}h_{k,j} + \beta\sum_{m\in\mathcal{M}_i}\frac{z_{k,m}}{\hat{z}_{k,m}}b_{i,m}\right)}{\sum_j h_{k,j} + \beta\sum_m b_{k,m}}, \tag{7}$$

$$h_{k,j}^{new} = h_{k,j}\frac{\left(\sum_{i\in\mathcal{I}_j}\frac{x_{i,j}}{\hat{x}_{i,j}}w_{k,j} + \alpha\sum_{n\in\mathcal{N}_j}\frac{y_{n,j}}{\hat{y}_{n,j}}a_{n,k}\right)}{\sum_i w_{k,j} + \alpha\sum_n a_{n,k}}, \tag{8}$$

$$a_{n,k}^{new} = a_{n,k}\frac{\sum_{n\in\mathcal{N}_j}\frac{y_{n,j}}{\hat{y}_{n,j}}h_{k,j}}{\sum_j h_{k,j}},\ b_{k,m}^{new} = b_{k,m}\frac{\sum_{m\in\mathcal{M}_i}\frac{z_{i,m}}{\hat{z}_{i,m}}w_{i,k}}{\sum_i w_{i,k}}. \tag{9}$$

For speeding up the calculation, let us denote $\mathcal{I}_j$ as sets of indices of non-zero entries on the j-th column. Let $\mathcal{J}_i$, $\mathcal{M}_i$, $\mathcal{N}_j$ similarly as sets of indices of non-zero entries on the i-th row in $\boldsymbol{X}$, the i-th row in $\boldsymbol{Z}$ and the j-th row in $\boldsymbol{Y}$. We denote $\sum_{i\in\mathcal{I}_j}$ as to sum up only non-zero entries of the j-th column, and so on. If $\alpha = \beta = 0$, the update rule of NM2F is the same as that of NMF.

3.5 NM2F as Probabilistic Generative Model

We can rewrite NM2F with generalized Kullback-Leibler divergence as a probabilistic generative model with a Poisson distribution, as with NMF. The log likelihood of NM2F can be rewritten as

$$\begin{aligned}&\ln p(\boldsymbol{X}, \tilde{\alpha}\boldsymbol{Y}, \tilde{\beta}\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \tilde{\alpha}\boldsymbol{A}, \tilde{\beta}\boldsymbol{B})\\ &= \ln p(\boldsymbol{X}, |\boldsymbol{W}, \boldsymbol{H}) + \ln p(\tilde{\alpha}\boldsymbol{Y}|\boldsymbol{H}, \tilde{\alpha}\boldsymbol{A}) + \ln p(\tilde{\beta}\boldsymbol{Z}|\boldsymbol{W}, \tilde{\beta}\boldsymbol{B})\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\ln p(x_{i,j}|\hat{x}_{i,j}) + \alpha\sum_{n=1}^{N}\sum_{j=1}^{J}\ln p(y_{n,j}|\hat{y}_{n,j})\\ &+ \beta\sum_{i=1}^{I}\sum_{m=1}^{M}\ln p(z_{i,m}|\hat{z}_{i,m})\\ &\simeq -\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta)\end{aligned} \tag{10}$$

where p denotes the Poisson distribution, $\tilde{\alpha} = \alpha/(NJK)$, and $\tilde{\beta} = \beta/(IMK)$. Minimizing Eq. (4) is equivalent to maximizing the log likelihood of probabilistic generative models. The graphical model of NM2F is shown in Figure 2.

3.6 NM2F as NMF with Undefined Region

Let us consider the problem of factorizing a matrix with an undefined (don't care) regions. Let $\boldsymbol{T} \in \mathbb{R}_+^{(I+N)\times(J+M)}$ as an observation matrix with undefined regions. We denote the bases matrix as $\boldsymbol{U} \in \mathbb{R}_+^{(I+N)\times K}$ and the coefficients matrix as $\boldsymbol{V} \in \mathbb{R}_+^{K\times(J+M)}$. We let ω_{ij} be the flag of undefined element.

$$\omega_{ij} = \begin{cases} 0 & t_{i,j}\ is\ undefined,\\ 1 & t_{i,j}\ is\ defined\end{cases}. \tag{11}$$

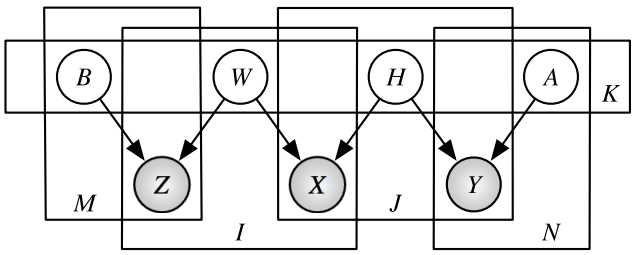

Figure 2: Graphical model of Non-negative Multiple Matrix Factorization. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are generated from the same bases $\boldsymbol{W}$ but different coefficients $\boldsymbol{H}$ and $\boldsymbol{B}$. $\boldsymbol{Y}$ is generated from bases A and coefficients $\boldsymbol{H}$ shared by $\boldsymbol{X}$.

Let $\boldsymbol{\Omega} \in \mathbb{R}_+^{(I+N)\times(J+M)}$ be a matrix consisting of ω_{ij}. We assume a case below.

$$\omega_{ij} = \begin{cases} 0 & \textit{if } i > I \textit{ and } j > J, \\ 1 & \textit{otherwise} \end{cases} . \tag{12}$$

Let us denote $\hat{\mathcal{D}}$ as the loss of NMF with an undefined region.

$$\hat{\mathcal{D}}(\boldsymbol{T}, \boldsymbol{\Omega}|\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=i)}^{I+N} \sum_{j=1}^{J+M} \omega_{ij} d(t_{ij}|\hat{t}_{ij}), \tag{13}$$

$\hat{\mathcal{D}}$ could be rewritten as:

$$\begin{aligned} \hat{\mathcal{D}}(\boldsymbol{T}, \boldsymbol{\Omega}|\boldsymbol{T}, \boldsymbol{V}) = &\sum_{i=I}^{I} \sum_{j=J}^{J} 1 \times d(t_{ij}|\hat{t}_{ij}) + \sum_{i=I+1}^{I+N} \sum_{j=1}^{J} 1 \times d(t_{ij}|\hat{t}_{ij}) \\ &+ \sum_{i=1}^{I} \sum_{j=J+1}^{J+M} 1 \times d(t_{ij}|\hat{t}_{ij}) + \sum_{i=I+1}^{I+N} \sum_{j=J+1}^{J+M} 0 \times d(t_{ij}|\hat{t}_{ij}). \end{aligned} \tag{14}$$

We set $\boldsymbol{T}$, $\boldsymbol{U}$, $\boldsymbol{V}$ as

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{X} & \tilde{\beta}\boldsymbol{Z} \\ \tilde{\alpha}\boldsymbol{Y} & \boldsymbol{O} \end{pmatrix}, \ \boldsymbol{U} = \begin{pmatrix} \boldsymbol{W} \\ \tilde{\alpha}\boldsymbol{A} \end{pmatrix}, \ \boldsymbol{V} = (\boldsymbol{H}\ \tilde{\beta}\boldsymbol{B}), \tag{15}$$

where $\boldsymbol{O} \in \mathbb{R}_+^{N\times M}$ is any non-negative matrix. Then $\hat{\mathcal{D}}$ could be written as:

$$\begin{aligned} &\hat{\mathcal{D}}(\boldsymbol{T}, \boldsymbol{\Omega}|\boldsymbol{T}, \boldsymbol{V}) \\ &= \sum_{i=I}^{I} \sum_{j=1}^{J} d(x_{ij}|\hat{x}_{ij}) + \alpha \sum_{n=1}^{N} \sum_{j=1}^{J} d(y_{nj}|\hat{y}_{nj}) + \beta \sum_{i=1}^{I} \sum_{m=1}^{M} d(z_{im}|\hat{z}_{im}) \\ &= \mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{B}; \alpha, \beta). \end{aligned} \tag{16}$$

According to these equations, NM2F is a special case of NMF with an undefined region. By adopting an appropriate undefined region, it is possible to factorizing more than three matrices simultaneously in NM2F frameworks.

4 Experiments

In this section, we present experimental validations of the proposed method on synthetic data set and real world data set.

4.1 Evaluation Measures

Though our primal objective is to obtain intuitive decomposition of the data matrix, a quantitative measure would help us understand and compare behaviors of several models. In the sparse data scenario, we are often only interested in non-zero entries of the data matrix, as well non-zero highest values in decomposed bases and coefficients. Therefore, modeling precision of non-zero entries is a reasonable measure for the factorization models.

Specifically, we employ the average log likelihood for a test set, randomly picked up from non-zero entries of the target data matrix $\boldsymbol{X}$ as our quantitative measure. We define the average log likelihood as:

$$\frac{1}{M} \sum_{m=1}^{M} \log p(x_m|\theta), \tag{17}$$

where $m = (1, \cdots, M)$ is a number of non-zero elements in the test set and θ is estimated parameters of a model. Model parameters were estimated by a 5-fold cross validation. A higher average test log likelihood indicates better modeling of the data structure.

4.2 Synthetic Data Experiment

This experiment evaluates the performance of NMF, VB-NMF (NMF with sparse constraint) [Cemgil, 2009], PMF [Salakhutdinov and Mnih, 2008] and NM2F. These models are examined for several sparseness data sets. We also conducted a grid search for α and β to assess the effect of the scaling parameters α and β,

A synthetic data set was obtained from the probabilistic model with Poisson distribution shown in Section 3.5. We set the size of matrices $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{A}$, and $\boldsymbol{B}$ as $I = J = N = M = 100$. We used values of $0\%, 0.9\%, 9\%$, and 99% for the sparseness of the synthetic data set. We compared the means of the average test log likelihood for each model. The parameters α and β were set to at the values that present the highest log likelihood for the training set. Before the verification, we iterated the parameter estimation 50 times. Note that, if $\alpha = \beta = 0$ then NM2F is reduced to the original NMF.

Results:

The average test log likelihood values are presented in Table 1. As is evident from these results, the proposed model improved the average test log likelihood for all data sets.

The results in Table 1 confirm that NM2F achieves good performance by making use of auxiliary matrices. Figure 3 shows the result of grid searches over α and β. With a dense $\boldsymbol{X}$ (0% sparse), the scaling parameters α and β scarcely contributed to the improvement of the log likelihood. This result means that if $\boldsymbol{X}$ is dense then auxiliary matrices are not so useful for improvement. On the other hand, the log likelihood is affected by the choice of scaling parameters as the target data become more sparse. Therefore we conclude that when the observed matrix is highly sparse, auxiliary matrices and appropriate scaling parameters yield better factorization.

4.3 Real Data Experiment

Next, we evaluate the proposed model on two real data sets. VBNMF and PMF are omitted from these experiments due to

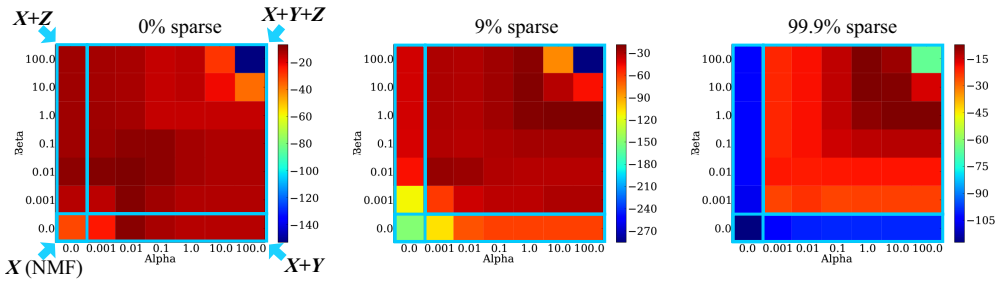

Figure 3: The average test log likelihood versus the scaling parameters α (horizontal) and β (vertical).

Sparse	NMF	VBNMF	PMF	NM2F
0 %	$\mathbf{-1.24 \pm 0.21}$	-2.72 ± 0.03	-2.47 ± 0.239	$\mathbf{-1.03 \pm 0.09}$
9 %	-19.39 ± 2.82	-8.49 ± 0.60	-13.00 ± 2.74	$\mathbf{-0.99 \pm 0.08}$
99 %	-42.45 ± 6.30	-14.55 ± 1.40	-16.25 ± 6.05	$\mathbf{-1.07 \pm 0.25}$
99.9 %	-43.25 ± 33.45	-15.30 ± 6.30	-12.85 ± 11.20	$\mathbf{-0.86 \pm 0.55}$

Table 1: Comparison on the average test log likelihood of non-zero entries among different sparseness. Means and one standard deviations of four data sets are presented.

their high computational costs.

Last.fm Data Set

The Last.fm data set is provided by *Hetrec 2011*[1]. This data set consists of the user action history of the Last.fm service: it includes the user's listening history, tag information about artists and friend link information among users.

The data set contains $1,892$ unique users, $17,632$ unique artists and $11,946$ unique tags. We denote the indices of users, artists, tags and friend users as i, j, n, and m, respectively. Let $x_{i,j}$ be a count of the listening of user i to the artist j. Let $\boldsymbol{X} \in \mathbb{R}_+^{1,892 \times 17,632}$ be the target matrix. Let $y_{n,j}$ denote the count of tags n appearing for the j-th artist. Let $\boldsymbol{Y} \in \mathbb{R}_+^{11,946 \times 17,632}$ be the column-wise auxiliary matrix. Then $z_{i,m}$ denotes the link state between the ith user and the m-th user, if users are friends then $z_{i,m} = 1$ else $z_{i,m} = 0$. Let $\boldsymbol{Z} \in \mathbb{R}_+^{1,892 \times 1,892}$ be the row-wise auxiliary matrix. We evaluated the performance of NMF and NM2F by factorizing $\boldsymbol{X}$. Note that the data set is highly sparse: only 0.25% ($3,687$ elements) of $x_{i,j}$ are non-zero. The number of bases was set to $K = 20$ based on preliminary experiments.

Social Curation Data Set

In this experiment, we employ a data set of the curating service Togetter[2] that was established for summarizing Twitter messages into a *story* [Duh *et al.*, 2012; Ishiguro *et al.*, 2012]. In Figure 4, we present an example of a curated story. This story is entitled "20110311 JAPAN MEGA QUAKE M8.8 -ENGLISH NEWS TL", and includes multiple Twitter messages. Words related to the title such as "quake" and "tsunami" appear several times. Multiple Twitter users are mentioned in the story; some of them appear multiple times (multiple posts are included in this story). Our goal with this experiment was to detect structures of users and stories with auxiliary information.

The data set contained $1,823,184$ unique users, $235,086$ stories including $23,859,294$ tweets, and $165,046$ unique words. Let us denote the indices of users, stories, words, and auxiliary features as i, j, n, and m, respectively. Let $x_{i,j}$ be the count of the appearances (equivalently, a number of messages) of user i in a story j. Let $\boldsymbol{X} \in \mathbb{R}_+^{1,823,184 \times 235,086}$ be the core matrix of user story observation. Let $y_{n,j}$ denote the count of word n appearing in the j-th story. Let $\boldsymbol{Y} \in \mathbb{R}_+^{165,046 \times 1,823,184}$ be the row wise auxiliary matrix. We additionally utilize the Twitter based features of each user. One is the number of followers and the other is the number of Lists. We scale these features $\tilde{z}$ by $z = \log(\tilde{z} + 1)$. Let $z_{i,m}$ be a score of the user i in a feature m. Let $\boldsymbol{Z} \in \mathbb{R}_+^{235,086 \times 2}$ be a column-wise auxiliary matrix. Note that only 0.0018% ($7,714,891$ elements) of $x_{i,j}$ are non-zero in the data set. Furthermore, the data set is very large, but parameter estimations of NM2F can be processed in hours by utilizing data set sparseness. K was fixed to 200 based on preliminary experiments.

Results:

The computed average test log likelihood values of non-zero entries are presented in Table 2. It is noteworthy that these two data sets are heavily long-tailed; dynamic ranges of observed data values are extremely wide. Combined with the severe sparseness, predicting non-zero entries is a difficult task. As evidenced from these results, however, we confirmed that NM2F achieved better performance than NMF for both data sets. Auxiliary data and appropriate scaling parameters surely contributed to improving the validation scores. Though our primal objective is intuitive decomposition, these scores indicate the modeling capability of the proposed NM2F.

Figures 5, 6, and 7 present three bases estimated in each experiment. We show the top 10 highest valued artists $\boldsymbol{W}$ and tags $\boldsymbol{A}$, and users $\boldsymbol{W}$ and words $\boldsymbol{A}$ in the figures. The bars in the figures present the value of each basis.

In Figure 5, we present bases extracted from the Last.fm data set by NM2F. We can confirm that reasonably related artist names and tags appear in the same basis. For example, basis #1 includes popular women singers such as *Britney Spears* and *Lady Gaga*. In basis #2, classic rock musicians are included. Rap and hip-hop singers are presented in basis #3. For comparison, we present extracted based by NMF in

[1] http://ir.ii.uam.es/hetrec2011/

[2] http://togetter.com/

Figure 4: Example of a Togetter story shortened by authors.

Data Set	NMF	VBNMF	PMF	NM2F
Last.fm	-6.90 ± 0.03	N/A	N/A	$\mathbf{-6.17 \pm 0.03}$
Togetter	-27.27 ± 0.23	N/A	N/A	$\mathbf{-12.97 \pm 0.48}$

Table 2: Comparison of average test log likelihood of non-zero entries among different data sets. Scores are divided by $\times 10^{-3}$

Figure 6. All presented bases include highest valued artists of bases in Figure 5. The basis #1 is almost identical to that of #1 in Figure 5, which includes popular women singers. In the bases #2 and #3 of NMF, different kinds of musicians are included in a single NMF base. In contrast, the bases #2 and #3 of NM2F consist of musicians in similar genres. This result clearly illustrates how our NM2F make factors more interpretable and comprehensible by factorizing multiple matrices simultaneously.

Figure 7 shows examples of learned bases of Social Curation data set. Basis #1 consists mass media accounts (*Asahi_Shakai*, *nhk_tokuho*), government official (*pref_iwate*), military officials (*US7thFlt*, *JGSDF_pr*) and a Twitter original account. In basis #2, professional programmers and open source contributors are extracted. Finally, basis #3 includes various accounts (from amateurs to professionals) who post feeds about *the Arab Spring* headline. This result indicates that NM2F is able to extract meaningful bases from the highly sparse data.

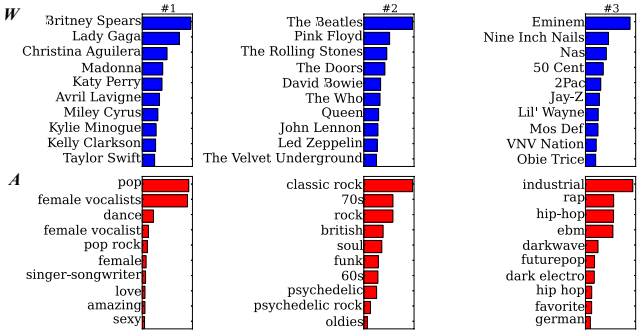

Figure 5: Artists and Tags with the highest values in bases of NM2F for Last.fm data.

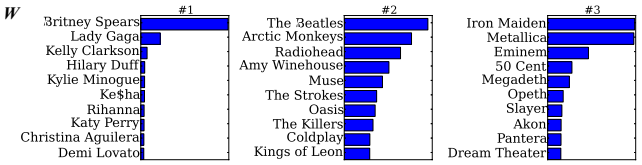

Figure 6: Artists with the highest values in bases of NMF for Last.fm data.

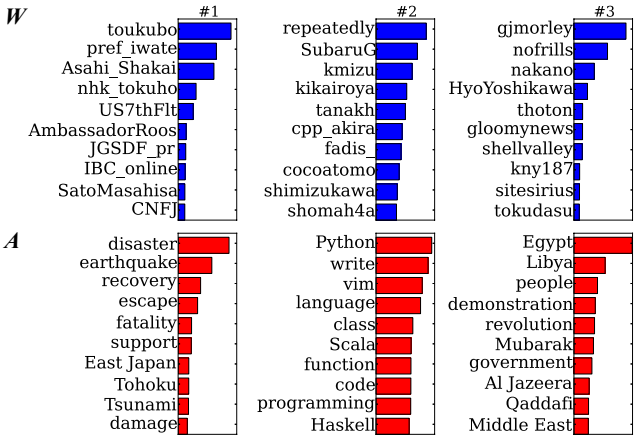

Figure 7: Users and Words with the highest values in bases of NM2F for social curation data.

5 Conclusion

In this paper, we proposed Non-negative Multiple Matrix Factorization (NM2F), which integrates auxiliary matrices in estimating bases and coefficients. Our key idea is to overcome the sparseness of the target data with complementary data. We derived a parameter estimation procedure as a multiplicative update rule. We presented NM2F with generalized Kullback-Leibler divergence as a probabilistic generative model with Poisson distributions. We also proved that NM2F is a special case of NMF with undefined regions. We evaluated NM2F experimentally using a synthetic data set and two real data sets, and confirmed that NM2F performs better than existing factorization methods in a quantitative measure. We also confirmed that the estimated bases of NM2F are reasonably intuitive and easy to understand. As a future work, we need to evaluate NM2F with other divergence choices. Also, finding a way to realize the automatic tuning of scaling parameters would be another interesting challenge.

A Parameter Update Rule Derivation

In this appendix, we provide details of the parameter estimation procedure mentioned in Section 3.4. The objective of the parameter estimation is to minimize the divergence in the Eq. (4).

$$\begin{aligned}
&\mathcal{D}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{A},\boldsymbol{B};\alpha,\beta)\\
&\simeq \sum_{i=1}^{I}\sum_{j=1}^{J}\Big(\hat{x}_{i,j} - x_{i,j}\log\sum_{k=1}^{K} w_{i,k}h_{k,j}\Big)\\
&+\alpha\sum_{n=1}^{N}\sum_{j=1}^{J}\Big(\hat{y}_{n,j} - y_{n,j}\log\sum_{k=1}^{K} a_{n,k}h_{k,j}\Big)\\
&+\beta\sum_{i=1}^{I}\sum_{m=1}^{M}\Big(\hat{z}_{i,m} - z_{i,m}\log\sum_{k=1}^{K} w_{i,k}b_{k,m}\Big). \qquad (18)
\end{aligned}$$

Let us introduce three auxiliary variables as $r_{i,j,k}$, $s_{n,j,k}$, $t_{i,m,k}$ ($\sum_k r_{i,j,k} = 1$, $\sum_k s_{n,j,k} = 1$, $\sum_k t_{i,m,k} = 1$) . By Jensen's inequality, an upper bound $\mathcal{F}$ of $\mathcal{D}$ is derived as below:

$$\begin{aligned}
&\mathcal{D}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{A},\boldsymbol{B};\alpha,\beta)\\
&\le \sum_{i=1}^{I}\sum_{j=1}^{J}\Big(\hat{x}_{i,j} - x_{i,j}\sum_{k=1}^{K} r_{i,j,k}\log\frac{w_{i,k}h_{k,j}}{r_{i,j,k}}\Big)\\
&+\alpha\sum_{n=1}^{N}\sum_{j=1}^{J}\Big(\hat{y}_{n,j} - y_{n,j}\sum_{k=1}^{K} s_{n,j,k}\log\frac{a_{n,k}h_{k,j}}{s_{n,j,k}}\Big)\\
&+\beta\sum_{i=1}^{I}\sum_{m=1}^{M}\Big(\hat{z}_{i,m} - z_{i,m}\sum_{k=1}^{K} t_{i,m,k}\log\frac{w_{i,k}b_{k,m}}{t_{i,m,k}}\Big)\\
&\triangleq \mathcal{F}. \qquad (19)
\end{aligned}$$

We have equality if and only if

$$\begin{aligned}
r_{i,j,k} &= \frac{w_{i,k}h_{k,j}}{\sum_{k=1}^{K} w_{i,k}h_{k,j}},\ s_{n,j,k} = \frac{a_{n,k}h_{k,j}}{\sum_{k=1}^{K} a_{n,k}h_{k,j}},\\
t_{i,m,k} &= \frac{w_{i,k}b_{k,m}}{\sum_{k=1}^{K} w_{i,k}b_{k,m}}. \qquad (20)
\end{aligned}$$

A partial differentiation of $\mathcal{F}$ for w_{ik} is

$$\frac{\partial\mathcal{F}}{\partial w_{ik}} = \sum_{j=1}^{J}\Big(h_{k,j} - x_{i,j}\frac{r_{i,j,k}}{w_{i,k}}\Big) + \beta\sum_{m=1}^{M}\Big(b_{k,j} - z_{i,m}\frac{t_{i,m,k}}{w_{i,k}}\Big). \qquad (21)$$

We set $\frac{\partial\mathcal{F}}{\partial w_{ik}} = 0$ then Eq. (21) could be rewritten as:

$$w_{i,k} = \frac{\sum_{j=1}^{J} x_{i,j}r_{i,j,k} + \beta\sum_{m=1}^{M} z_{k,m}t_{i,m,k}}{\sum_{j=1}^{J} h_{k,j} + \beta\sum_{m=1}^{M} b_{k,m}}. \qquad (22)$$

We obtain the multiplicative update rule of $w_{i,k}$:

$$w_{i,k}^{new} = w_{i,k}\frac{\Big(\sum_{j=1}^{J}\frac{x_{i,j}}{\hat{x}_{i,j}}h_{k,j} + \beta\sum_{m=1}^{M}\frac{z_{k,m}}{\hat{z}_{k,m}}b_{i,m}\Big)}{\sum_{j=1}^{J} h_{k,j} + \beta\sum_{m=1}^{M} b_{k,m}}. \qquad (23)$$

This update rule could be rewritten as below:

$$w_{i,k}^{new} = w_{i,k}\frac{\Big(\sum_{j\in\mathcal{J}_i}\frac{x_{i,j}}{\hat{x}_{i,j}}h_{k,j} + \beta\sum_{m\in\mathcal{M}_i}\frac{z_{k,m}}{\hat{z}_{k,m}}b_{i,m}\Big)}{\sum_{j} h_{k,j} + \beta\sum_{m} b_{k,m}}. \qquad (24)$$

Further, the second order partial differential of $\mathcal{F}$ for w_{ik} is written as

$$\frac{\partial^2\mathcal{F}}{\partial w_{i,k}^2} = \sum_{j=1}^{J}\Big(x_{i,j}\frac{r_{i,j,k}}{w_{i,k}^2}\Big) + \beta\sum_{m=1}^{M}\Big(z_{i,m}\frac{t_{i,m,k}}{w_{i,k}^2}\Big) \ge 0. \qquad (25)$$

Eq. (25) indicates that $\mathcal{F}$ is biconvex with $w_{i,k}$. Therefore, the monotonous convergence at the local optimum is guaranteed.

References

[Abdallah and Plumbley, 2004] S.A. Abdallah and M.D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. In *Proc. ISMIR*, 2004.

[Aharon *et al.*, 2006] M. Aharon, M. Elad, and A.M. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1):48–67, 2006.

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Cai *et al.*, 2011] D. Cai, X. He, and J. Han. Graph regularized non-negative matrix factorization for data representation. *IEEE Trans.*, 33:1548–1560, 2011.

[Cao *et al.*, 2007] B. Cao, D. Shen, J.T. Sun, X. Wang, Q. Yang, and Z. Chen. Detect and track latent factors with online nonnegative matrix factorization. In *Proc. IJCAI*, 2007.

[Cemgil, 2009] A.T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

[Cichocki *et al.*, 2007] A. Cichocki, R. Zdunek, and S. Amari. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In *Proc. ICA*, 2007.

[Cichocki *et al.*, 2009] A. Cichocki, A. H. Phan R. Zdunek, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis, John Wiley*. Wiley, 2009.

[Dikmen and Févotte, 2012] O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Trans.*, 2012.

[Ding *et al.*, 2006] C. Ding, T. Li, and W. Peng. NMF and PLSI: Equivalence and a hybrid algorithm. In *Proc. SIGIR*, 2006.

[Duh *et al.*, 2012] K. Duh, T. Hirao, A. Kimura, K. Ishiguro, T. Iwata, and C. A. Yeung. Creating stories: Social curation of Twitter messages. In *Proc. ICWSM*, 2012.

[Hoffman *et al.*, 2011] M. D. Hoffman, D. M. Blei, and P. R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proc. ICML*, 2011.

[Hofmann, 1999] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, 1999.

[Hoyer, 2004] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[Ishiguro *et al.*, 2012] K. Ishiguro, A. Kimura, and K. Takeuchi. Towards automatic image understanding and mining via social curation. In *Proc. ICDM*, 2012.

[Koren *et al.*, 2009] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE*, 2009.

[Lee and Seung, 1999] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, October 1999.

[Lee *et al.*, 2000] D. D. Lee, Daniel D., and S. H. Sebastian. Algorithms for non-negative matrix factorization. In *Proc. NIPS*, 2000.

[Lin *et al.*, 2011] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proc. SIGKDD*, 2011.

[Liu *et al.*, 2010] C. Liu, H. Yang, J. Fan, L. He, and Y. Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on MapReduce. In *Proc. WWW*, 2010.

[Liu *et al.*, 2012] H. Liu, Z. Yang, Z. Wu, and X. Li. A-optimal non-negative projection for image representation. In *Proc. CVPR*, 2012.

[Ma *et al.*, 2008] H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: Social recommendation using probabilistic matrix factorization. In *Proc. CIKM*, 2008.

[Mei *et al.*, 2008] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *Proc. WWW*, 2008.

[Nakano *et al.*, 2011] M. Nakano, J. L. Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama. Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model. In *Proc. WASPAA*, 2011.

[Noel *et al.*, 2012] J. Noel, S. Sanner, K. Tran, P. Christen, L. Xie, E. V. Bonilla, E. Abbasnejad, and N. D. Penna. New objective functions for social collaborative filtering. In *Proc. WWW*, 2012.

[Purushotham *et al.*, 2012] S. Purushotham, Y. Liu, and C. C. J. Kuo. Collaborative topic regression with social matrix factorization for recommendation systems. In *Proc. ICML*, 2012.

[Salakhutdinov and Mnih, 2008] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Proc. NIPS*, 2008.

[Schmidt *et al.*, 2009] M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *Proc. ICASSP*, 2009.

[Smaragdis and Brown, 2003] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. WASPA*, 2003.

[Steyvers *et al.*, 2004] M. Steyvers, P. Smyth, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proc. SIGKDD*, 2004.

[Wang and Blei, 2011] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proc. SIGKDD*, 2011.

[Wang *et al.*, 2011] F. Wang, C. Tan, A.C. König, and P. Li. Efficient document clustering via online nonnegative matrix factorizations. In *Proc. SIAM*, 2011.

[Xu *et al.*, 2003] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. SIGIR*, 2003.