Uncoupled Nonnegative Matrix Factorization with Pairwise Comparison Data

Masahiro Kohjima
NTT Human Informatics Laboratories, NTT Corporation
Yokosuka, Japan
masahiro.kohjima.ev@hco.ntt.co.jp

ABSTRACT

In this paper, we propose a new method called uncoupled nonnegative matrix factorization (UNMF). UNMF enables us to analyze data that cannot be represented by a matrix, due to the lack of correspondence between the index and values of the matrix elements caused by e.g., data collection under the constraint of privacy protection. We derive the multiplicative update rules for parameter estimation and confirm the effectiveness of UNMF by numerical experiments.

CCS CONCEPTS

• **Computing methodologies** → **Non-negative matrix factorization**; • **Information systems** → *Recommender systems.*

KEYWORDS

uncoupled data, matrix factorization, Bregman divergence

ACM Reference Format:
Masahiro Kohjima. 2022. Uncoupled Nonnegative Matrix Factorization with Pairwise Comparison Data. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539813.3545149

1 INTRODUCTION

For the analysis of data represented by a nonnegative matrix such as document corpus, movie rating scores, purchase logs and survey questionnaires, nonnegative matrix factorization (NMF) [16, 17] is widely applied [6, 15, 23, 24]; NMF can extract latent patterns within the data and complete missing values by decomposing the input matrix into the product of two nonnegative factor matrices. By applying NMF to movie ratings, for instance, we can extract user's rating patterns such as Sci-fi and comedy lover and predict a user's rating scores to movies not yet watched by the user.

The motivation of this study is to extend NMF to cover the analysis of *uncoupled data*, data that cannot be represented by a matrix due to the lack of correspondence between the index and value of the matrix elements. This is necessary, for example, when analyzing survey data where the index-value relationship (e.g., correspondence between user and answer of the user to a questionnaire item) has been removed for privacy protection, or analyzing social data where the indexes and values are collected independently.

If contextual information about real entities corresponding to the index (e.g., user) such as user's sex, age and preferred movie genre are available, we can establish correspondence between the indexes and values by the approach called *matching* [18, 19] which allows NMF to be applied; however, this approach seems to require vast amounts of contextual information and failure of the recovery degrades NMF performance. Thus it is promising to build a method that does not involve correspondence recovery.

In this study, we propose a new method called uncoupled nonnegative matrix factorization (UNMF) that estimates factor matrices directly from uncoupled data without recovering the index-value correspondence. The key to our method construction is to use the approach of uncoupled regression (UR) [1, 5, 12, 20, 22], especially the approach of Xu et al. [22] which estimates regression models using uncoupled data and *pairwise comparison data* (PCD) containing ranking information that indicates that, given two indexes of matrix elements, which value of the element is larger than that of the other. This data can be collected by asking users "who is older than you?", "who has a higher income than you?" etc. Even if the question is about the sensitive matter such as age and income, it is easier for users to answer such indirect questions than direct questions since the value itself needs not to be disclosed.

Given both uncoupled data and PCD, we derive the objective function of UNMF by approximating the expected Bregman divergence and provide multiplicative update rules for estimating the factor matrices. This allows us to extract latent patterns underlying the uncoupled data and to predict values of the matrix elements. We conduct experiments on both synthetic and real data to confirm that UNMF matches the performance attained by *oracle* NMF on coupled data whose index-value relationships is known.

2 RELATED WORKS

NMF is regarded as a key unsupervised learning algorithm since its relation to many algorithms has been clarified e.g.,[9, 10]. For example, Ding et al. [10] proved that NMF is equivalent to probabilistic latent semantic indexing [11], which is the basis of latent Dirichlet allocation [3]. Thus, we develop our method on NMF.

Our problem is different from the missing value completion problem. Completion by NMF is done by first estimating factor matrices from observed coupled data where index-value relationships are known (see § 3). However, since the coupled data are not provided in our problem, NMF cannot be applied. So our UNMF is constructed.

Our derivation of the loss function for UNMF uses the approach of UR that conducts *supervised* learning from uncoupled data. Although various studies for UR are known [1, 5, 12, 20, 22], it has

---

ICTIR '22, July 11–12, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545149

Table 1: Bregman divergence with various convex functions

	domain	$\varphi(x)$	$\psi(x) = \nabla\varphi(x)$	$d_\varphi(y, x)$
SE	$\mathbb{R}$	$x^2/2$	x	$(y-x)^2/2$
KL	$\mathbb{R}_+$	$x\log(x)$	$\log(x)+1$	$y\log\frac{y}{x} - y + x$

been confirmed that the recovery of the correspondence between input and output is NP-hard in general [20]. So we adopt an approach that does not estimate the correspondence [1, 5, 22]. Among them, the model estimated needs to be monotone in [5] or linear in [1]. So we construct UNMF that conducts *unsupervised* learning from uncoupled data based on [22].

3 PRELIMINARY

Let n_I, n_J, n_R be positive integers. We define $\mathcal{I} = \{1, \cdots, n_I\}$ and $\mathcal{J} = \{1, \cdots, n_J\}$. In the standard setting of NMF, a set of pairs of index and value (i.e., coupled data) $\mathcal{D}_{comp} = \{(i_m, j_m, x_m)\}_{m=1}^{n_{comp}}$, which can be represented by nonnegative matrix $\boldsymbol{X} \in \mathcal{X}^{n_I \times n_J} \subset \mathbb{R}_+^{n_I \times n_J}$ whose (i_m, j_m)-th element is x_m, is given. n_{comp} is the number of pairs. The model parameter θ, which consists of two factor matrices $\boldsymbol{A} = \{\{a_{ir}\}_{r=1}^{n_R}\}_{i\in\mathcal{I}}$ and $\boldsymbol{B} = \{\{b_{rj}\}_{r=1}^{n_R}\}_{j\in\mathcal{J}}$, is estimated by minimizing a loss function that is defined by some divergence between $\boldsymbol{X}$ and the product of the factor matrices $\hat{\boldsymbol{X}}$ whose (i, j)-th element is $\hat{x}_{ij} = \sum_{r=1}^{n_R} a_{ir} b_{rj}$. It is known that various types of loss function can be expressed by Bregman Divergence (BD) D_φ [8, 21] which is defined as

$$D_\varphi(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \frac{1}{n_{comp}} \sum\nolimits_{(i_m, j_m, x_m)\in\mathcal{D}_{comp}} d_\varphi(x_m, \hat{x}_{i_m j_m}),$$
$$d_\varphi(x, y) = \varphi(x) - \varphi(y) - (x - y)\psi(y),$$

where φ is a convex function and ψ is its 1st derivative $\psi = \nabla\varphi$. BD generates various loss functions by varying φ [2, 4]. For example, BD is the squared error (SE) if $\varphi(x) = x^2/2$, and is the generalized KL-divergence (KL) if $\varphi(x) = x\log(x)$ (See Table 1). We also use BD for UNMF to handle various loss functions.

4 PROPOSED METHOD

4.1 Problem Formulation

Our proposed method, uncoupled nonnegative matrix factorization (UNMF), estimates model parameter $\theta = \{\boldsymbol{A}, \boldsymbol{B}\}$ without requiring coupled data $\mathcal{D}_{comp}$. Instead, the following three types of data $\mathcal{D}_U$, $\mathcal{D}_V$ and $\mathcal{D}_C$, made from uncoupled data and pairwise comparison data (PCD), are given in our problem.

To provide intuitive explanation of our problem, consider the following scenario of survey data collection. Each user j is asked to answer questionnaire items i (e.g., age, weight, annual income) and user's answer x is recorded without tying it to user's identifier j for privacy protection. Thus, even if many answers about item i are collected, we do not know which answer came from which user, i.e., the (user-) index and the values are uncoupled. We call the data obtained by this collection process uncoupled data.

Uncoupled Data $\mathcal{D}_U \cup \mathcal{D}_V$: Due to the lack of correspondence between the index and value, the uncoupled data cannot be represented by a matrix but can be represented by (i) a set of user-item pairs $\mathcal{D}_U = \{(i_m, j_m)\}_{m=1}^{n_U}$ that record user j_m answered item i_m, and (ii) a set of item-answer pairs $\mathcal{D}_V = \{(i_m, x_m)\}_{m=1}^{n_V}$ that record answers x_m for each item i_m. We call $\mathcal{D}_U$ *missing value data* (MVD) and $\mathcal{D}_V$ *missing index data* (MID). n_U and n_V is the number of pairs in MVD and MID, respectively. Note that $n_U \neq n_V$ in general since e.g., some invalid answers like "N/A" or "300 years old" may be removed from MID. We denote a subset of $\mathcal{D}_U$ such that $i_m = i$ ($j_m = j$) as $\mathcal{D}_{U_i}$ ($\mathcal{D}_{U_j}$). We also denote a subset of $\mathcal{D}_V$ such that $i_m = i$ as $\mathcal{D}_{V_i}$. The size of $\#\mathcal{D}_{V_i}$ is given by the symbol n_{V_i}.

Note that the scope of our study is not limited to this example. For example, user's viewing logs that record user j watches movie i and movie's rating logs that movie i is rated as score x are collected individually, they are represented by $\mathcal{D}_U$ and $\mathcal{D}_V$.

Pairwise Comparison Data $\mathcal{D}_C$: As stated in § 1, we use (iii) PCD which is defined by $\mathcal{D}_C = \{(i_m, j_m^+, j_m^-)\}_{m=1}^{n_C}$ where (i_m, j_m^+, j_m^-) indicates that the value of (i_m, j_m^+)-th element is larger than that of (i_m, j_m^-)-th element, and n_C is the number of data items. We denote a subset of $\mathcal{D}_C$ such that $i_m = i$ as $\mathcal{D}_{C_i}$. Similarly, we denote another subset such that $j_m^+ = j$ ($j_m^- = j$) as $\mathcal{D}_{C_j^+}$ ($\mathcal{D}_{C_j^-}$). We consider the setting where only a small amount of PCD is available and the order of all indices in uncoupled data cannot be determined.

4.2 Loss Function and Approximation

To derive the loss function for UNMF, we define random variables and probability distributions. Let I, J and X be random variables on $\mathcal{I}$, $\mathcal{J}$ and $\mathcal{X}$, respectively. We assume these random variables follow some distribution $P_{I,J,X}$. The marginal distribution on $\mathcal{I}$ (marginalized over J, X) is denoted as P_I. We also denote the conditional distribution and its probability density function (PDF) given $I = i$ as $P_{J,X|i}$ and $f_{J,X|i}$, respectively. Its marginal distribution on $\mathcal{J}$ ($\mathcal{X}$) is denoted as $P_{J|i}$ ($P_{X|i}$). Using these notations, we define the loss function for UNMF by the following expected BD:

$$\mathcal{R}(\theta) = \mathbb{E}_{I,J,X}[d_\varphi(X, \hat{x}_{IJ})] = \mathbb{E}_I[\mathcal{R}_i(\theta)], \tag{1}$$

where $\mathbb{E}_{I,J,X}$ and $\mathbb{E}_I$ is the expectation over $P_{I,J,X}$ and that over P_I, respectively. $\mathcal{R}_i$ is defined as follows.

$$\mathcal{R}_i(\theta) = \mathfrak{C}_i - \mathbb{E}_{J|i}[\varphi(\hat{x}_{iJ}) - \hat{x}_{iJ}\psi(\hat{x}_{iJ})] - \mathbb{E}_{J,X|i}[X\psi(\hat{x}_{iJ})], \tag{2}$$

where $\mathfrak{C}_i$ is a constant term and $\mathbb{E}_{J,X|i}$ ($\mathbb{E}_{J|i}$) is the expectation over $P_{J,X|i}$ ($P_{J|i}$). Note that if I, J follows a uniform distribution on $\mathcal{I}, \mathcal{J}$ and X is replaced by sample realization, the loss $\mathcal{R}$ is equivalent to BD D_φ used in standard NMF (See § 3).

The difficulty of evaluating $\mathcal{R}$ comes from the final term in Eq. (2) which involves the expectation over $P_{J,X|i}$, $\mathbb{E}_{J,X|i}[X\psi(\hat{x}_{iJ})]$. This term cannot be evaluated even if taking sample approximation since the user j and the answer x are not observed simultaneously.

To (approximately) evaluate the above problematic term, we introduce a pair of random variables on $\mathcal{J}$, (J^+, J^-); it indicates the value of the (i, J^+)-th element is larger than that of the (i, J^-)-th element. Formal definition is given by

$$J^+ = \begin{cases} J & (X \geq X') \\ J' & (X < X') \end{cases}, \quad J^- = \begin{cases} J' & (X \geq X') \\ J & (X < X') \end{cases},$$

where (J, X) and (J', X') are two independent random variables following $P_{J,X|i}$. We denote the conditional distribution and its PDF of J^+ (J^-) given $I = i$ as $P_{J^+|i}$ ($P_{J^-|i}$) and $f_{J^+|i}$ ($f_{J^-|i}$), respectively. The expectation over $P_{J^+|i}$ ($P_{J^-|i}$) is written as $\mathbb{E}_{J^+|i}$ ($\mathbb{E}_{J^-|i}$). Later, we will use the fact that sample (i_m, j^+_m, j^-_m) in PCD is regarded as a realization of (J^+, J^-) given $I = i_m$. The use of (J^+, J^-) connects the terms using expectation over $P_{J^+|i}$, $P_{J^-|i}$ and $P_{X,J|i}$:

Lemma 4.1. *Let $f_{X|i}$ and $F_{X|i}$ be PDF and cumulative density function (CDF) of probability distribution $P_{X|i}$, respectively. Then, $\mathbb{E}_{J^+|i}[\psi(\hat{x}_{iJ^+})] = 2\mathbb{E}_{J,X|i}[F_{X|i}(X)\psi(\hat{x}_{iJ})]$ and $\mathbb{E}_{J^-|i}[\psi(\hat{x}_{iJ^-})] = 2\mathbb{E}_{J,X|i}[\{1 - F_{X|i}(X)\}\psi(\hat{x}_{iJ})]$.*

Proof. From the definition of J^+, $f_{J^+|i}$ is written as

$$
\begin{aligned}
f_{J^+|i}(j) &= \frac{1}{Z}\iint \sum\nolimits_{j'\in\mathcal{J}} f_{J,X|i}(j,x) f_{J,X|i}(j',x')\mathbb{I}(x{>}x')dxdx' \\
&= \frac{1}{Z}\int f_{J,X|i}(j,x)\Big[\int f_{X|i}(x')\mathbb{I}(x > x')dx'\Big]dx \\
&= \frac{1}{Z}\int f_{J,X|i}(j,x)F_{X|i}(x)dx,
\end{aligned}
$$

where Z is a normalizing factor and $Z = 1/2$ is obtained by integration by parts. Then, we get $\mathbb{E}_{J^+|i}[\psi(\hat{x}_{iJ^+})] = \int \psi(\hat{x}_{ij})f_{J^+|i}(j)dj = 2\mathbb{E}_{J,X|i}[F_{X|i}(X)\ \psi(\hat{x}_{iJ})]$. The equation for $\mathbb{E}_{J^-|i}[\psi(\hat{x}_{iJ^-})]$ is obtained in an analogous manner. □

The term $\mathbb{E}_{J,X|i}[F_{X|i}(X)\psi(\hat{x}_{iJ})]$ in Lemma 4.1 is different from the problematic term $\mathbb{E}_{J,X|i}[X\psi(\hat{x}_{iJ})]$ in Eq. (2) but can be used as the approximation. This leads the approximated loss function $\tilde{\mathcal{R}}$:

Theorem 4.2. *Suppose that there exists a constant M such that $\psi(x) < M$ for all $x \in \mathcal{X}$. The loss $\mathcal{R}(\theta)$ is approximated by $\tilde{\mathcal{R}}(\theta; w, \lambda) = \mathbb{E}_I[\tilde{\mathcal{R}}_i(\theta; w_i, \lambda_i)]$ and its approximation error $\mathbb{E}_I[|\mathcal{R}_i(\theta) - \tilde{\mathcal{R}}_i(\theta)|]$ is bounded by $\mathbb{E}_I[\mathrm{Err}_i(w_i)]$ where*

$$
\begin{aligned}
\tilde{\mathcal{R}}_i(\theta; w_i, \lambda_i) &= \mathfrak{C}_i - \mathbb{E}_{J|i}[\varphi(\hat{x}_{iJ}) - (\hat{x}_{iJ} - \lambda_i)\psi(\hat{x}_{iJ})] \\
&\qquad -\big(w_{i1} - \frac{\lambda_i}{2}\big)\mathbb{E}_{J^+|i}[\psi(\hat{x}_{iJ^+})] - \big(w_{i2} - \frac{\lambda_i}{2}\big)\mathbb{E}_{J^-|i}[\psi(\hat{x}_{iJ^-})], \\
\mathrm{Err}_i(w_i) &= \int f_{X|i}(x)\big|h_i(x; w_i)\big|dx, \qquad (3) \\
h_i(x; w_i) &= x - 2w_{i1}F_{X|i}(x) - 2w_{i2}\{1 - F_{X|i}(x)\}.
\end{aligned}
$$

Proof. Taking the sum of the equations in Lemma 1, we get $\mathbb{E}_{J,X|i}[\psi(\hat{x}_{iJ})] = \frac{1}{2}\mathbb{E}_{J^+|i}[\psi(\hat{x}_{iJ^+})] + \frac{1}{2}\mathbb{E}_{J^-|i}[\psi(\hat{x}_{iJ^-})]$. Then, as the value of $\tilde{\mathcal{R}}_i$ does not depend on λ_i, we focus on $\tilde{\mathcal{R}}_i(\theta; w_i, 0)$. Since

$$
\begin{aligned}
&|\mathcal{R}_i(\theta) - \tilde{\mathcal{R}}_i(\theta; w_i, 0)| \\
&= \big|\mathbb{E}_{J,X|i}[X\psi(\hat{x}_{iJ})] - w_{i1}\mathbb{E}_{J^+|i}[\psi(\hat{x}_{iJ^+})] - w_{i2}\mathbb{E}_{J^-|i}[\psi(\hat{x}_{iJ^-})]\big| \\
&= \Big|\sum\nolimits_{j\in\mathcal{J}}\int f_{J,X|i}(j,x)\psi(\hat{x}_{ij})h_i(x; w_i)dx\Big| \\
&\leq \sum\nolimits_{j\in\mathcal{J}}\int f_{J,X|i}(j,x)\big|\psi(\hat{x}_{ij})\big|\big|h_i(x; w_i)\big|dx \leq M\mathrm{Err}_i(w_i),
\end{aligned}
$$

which yields the bound. □

The approximation is exact if $F_{X|i}$ is a uniform distribution, i.e., $F_{X|i}(x) = (x - a_i)/(b_i - a_i)$ for all $x \in [a_i, b_i]$, since $h_i(x; b_i/2, a_i/2) = 0$. For general non-uniform distributions we can optimize w_i by minimizing the bound (explained later).

Since $\tilde{\mathcal{R}}$ does not involve the term with expectation over $P_{J,X|i}$, we can evaluate it using MVD $\mathcal{D}_U$ and PCD $\mathcal{D}_C$. By removing the constant terms and approximating the expectation by sample averaging, we get the following empirical loss function $\hat{\mathcal{R}}$:

$$
\begin{aligned}
&\hat{\mathcal{R}}(\theta; w, \lambda) \qquad (4) \\
&= -\frac{1}{n_U}\sum\nolimits_{(i_m,j_m)\in\mathcal{D}_U}\{\varphi(\hat{x}_{i_m j_m}) - (\hat{x}_{i_m j_m} - \lambda_{i_m})\psi(\hat{x}_{i_m j_m})\} \\
&\quad - \frac{1}{n_C}\sum\nolimits_{(i_m,j^+_m,j^-_m)\in\mathcal{D}_C}\{(w_{i_m 1} - \lambda_{i_m}/2)\psi(\hat{x}_{i_m j^+_m}) \\
&\qquad\qquad + (w_{i_m 2} - \lambda_{i_m}/2)\psi(\hat{x}_{i_m j^-_m})\} \\
&= \sum\nolimits_{i\in\mathcal{I}}\sum\nolimits_{j\in\mathcal{J}}\big[-\gamma^U_{ij}\{\varphi(\hat{x}_{ij}) - \hat{x}_{ij}\psi(\hat{x}_{ij})\} - z_{ij}\psi(\hat{x}_{ij})\big],
\end{aligned}
$$

where $z_{ij} = \gamma^{C^+}_{ij} w_{i1} + \gamma^{C^-}_{ij} w_{i2} + \big(\gamma^U_{ij} - \frac{\gamma^{C^+}_{ij} + \gamma^{C^-}_{ij}}{2}\big)\lambda_i$, and $\gamma^U_{ij}, \gamma^{C^+}_{ij}, \gamma^{C^-}_{ij}$ are defined as $\gamma^U_{ij} = n_U^{-1}\sum_{(i_m,j_m)\in\mathcal{D}_U}\mathbb{I}(i_m = i, j_m = j)$, $\gamma^{C^+}_{ij} = n_C^{-1}\sum_{(i_m,j^+_m,j^-_m)\in\mathcal{D}_C}\mathbb{I}(i_m = i, j^+_m = j)$, $\gamma^{C^-}_{ij} = n_C^{-1}\sum_{(i_m,j^+_m,j^-_m)\in\mathcal{D}_C}\mathbb{I}(i_m = i, j^-_m = j)$. $\mathbb{I}(\cdot)$ is the indicator function. Note that λ_i can take an arbitrary value (see the proof of Theorem 4.2). That is, even if j^+_m or j^-_m is missing, we can compute the loss by setting $\lambda_i = 2w_{i1}$ or $2w_{i2}$ in Eq. (4). In the experiment, we set $\lambda_i = 0$ or $\lambda_i = (w_{i1} + w_{i2})/2$, similar to [22].

Algorithm 1 Uncoupled Nonnegative Matrix Factorization

```
Input: $\mathcal{D}_U, \mathcal{D}_V, \mathcal{D}_C, n_R$     Output: $\theta = \{\boldsymbol{A}, \boldsymbol{B}\}$
Estimate $w_i$ by minimizing $\widehat{\mathrm{Err}}_i(w_i)$ for each $i \in \mathcal{I}$.
Initialize$\boldsymbol{A}$ and $\boldsymbol{B}$.
repeat
    Update $\boldsymbol{A}$ by Eq. (5) for SE or Eq. (6) for KL
    Update $\boldsymbol{B}$ by Eq. (5) for SE or Eq. (6) for KL
until a stopping condition is met
```

4.3 Algorithm

We construct the estimation algorithm using $\hat{\mathcal{R}}$ (Eq. (4)) and the error bound Err_i (Eq. (3)). The algorithm consists of two steps: (i) the estimation of w_i by minimizing the error bound and (ii) the estimation of θ by solving the optimization problem $\min_\theta \hat{\mathcal{R}}$ subject to the non-negativity of $\boldsymbol{A}$ and $\boldsymbol{B}$. Pseudo code of the algorithm is shown in Alg. 1. The details are explained below.

Estimation of $\theta = \{\boldsymbol{A}, \boldsymbol{B}\}$: As shown in the next subsection, the following "multiplicative" update rules can be used for estimation.

$$
(SE)\ a_{ir} \leftarrow a_{ir}\frac{\sum_{j\in\mathcal{J}} z_{ij}b_{rj}}{\sum_{j\in\mathcal{J}} \gamma^U_{ij}\hat{x}_{ij}b_{rj}},\ b_{jr} \leftarrow b_{jr}\frac{\sum_{i\in\mathcal{I}} z_{ij}a_{ir}}{\sum_{i\in\mathcal{I}} \gamma^U_{ij}\hat{x}_{ij}a_{ir}}, \qquad (5)
$$

$$
(KL)\ a_{ir} \leftarrow a_{ir}\frac{\sum_{j\in\mathcal{J}} \frac{z_{ij}}{\hat{x}_{ij}}b_{rj}}{\sum_{j\in\mathcal{J}} \gamma^U_{ij}b_{rj}},\ b_{jr} \leftarrow b_{jr}\frac{\sum_{i\in\mathcal{I}} \frac{z_{ij}}{\hat{x}_{ij}}a_{ir}}{\sum_{i\in\mathcal{I}} \gamma^U_{ij}a_{ir}}. \qquad (6)
$$

Equation (5) and (6) correspond to the update rules for the settings where BD used in (1) is SE ($\varphi(x) = x^2/2$) and KL($\varphi(x) = x\log(x)$), respectively. If z_{ij} is non-negative (we can ensure this by appropriately setting w_i and λ_i), we can easily confirm that the right hand side of the update rule is (I) always non-negative and (II) unchanged if $\hat{x}_{ij} = z_{ij}/\gamma^U_{ij}$. Randomly setting initial (nonnegative) values of $\boldsymbol{A}, \boldsymbol{B}$ and iteratively updating the matrices yields the factorization results. Later, we show that the objective function is monotonically decreasing and converges to (local) minima.

Remark 1: These update rules are equivalent to the one for (standard) NMF [17] if $\gamma^U_{ij} = 1/n_U$ and $n_U z_{ij}$ is regarded as the observed element x_{ij}. Matrix $\tilde{Z}$ which is defined so that its (i, j)-th element $\tilde{z}_{ij} = n_U z_{ij}$ can be seen as the *Pseudo observation matrix.*

Remark 2: In the computation of update rules, we can skip the indexes (i, j) which are not included in $\mathcal{D}_U$ and $\mathcal{D}_C$ since $\gamma^U_{ij} = z_{ij} = 0$. So the computational cost of the each step of this algorithm is $O(LR)$ where L is the total number of indexes that appears at least once in $\mathcal{D}_U$ or $\mathcal{D}_C$.

Estimation of $w{=}\{w_i\}_{i\in\mathcal{I}}$**:** The upper bound Err_i can be evaluated by sample approximation using MID $\mathcal{D}_V$ as follows:

$$\widehat{\mathrm{Err}}_i(w_i){=}\frac{1}{n_{V_i}}\sum\nolimits_{x_m\in\mathcal{D}_{V_i}}\left|x_m{-}2w_{i1}\hat{F}_{X|i}(x_m){-}2w_{i2}\{1{-}\hat{F}_{X|i}(x_m)\}\right|,$$

where $\hat{F}_{X|i}$ is the empirical $F_{X|i}$, $\hat{F}_{X|i}(x) = (1/n_{V_i})\sum_{x_m\in\mathcal{D}_{V_i}}\mathbb{I}(x_m \leq x)$. Adding some constraints for keeping the nonnegativity of z_{ij} [1], w_i is estimated by minimizing $\widehat{\mathrm{Err}}_i$ using, e.g., L-BFGS method.

4.4 Algorithm Derivation and Analysis

Here we derived the update rules for SE (Eq. (5)) by the majorization minimization (MM) [7, 13] and show its theoretical property [2]. The loss $\hat{\mathcal{R}}$ (Eq. (4)) with $\varphi(x) = \frac{x^2}{2}$, $\hat{\mathcal{R}}_{SE}$, is expanded as

$$\hat{\mathcal{R}}_{SE}(\theta; w, \lambda) = \sum\nolimits_{i\in\mathcal{I}}\sum\nolimits_{j\in\mathcal{J}}\Big[\frac{\gamma^U_{ij}}{2}\Big(\sum\nolimits_{r=1}^{n_R} a_{ir}b_{rj}\Big)^2 - z_{ij}\hat{x}_{ij}\Big].$$

By applying Jensen's inequality, we can derive the majorizing function $\mathcal{F}$ using auxiliary variable $\boldsymbol{S} = \{\{s_{ijr}\}_{r=1}^{n_R}\}_{i\in\mathcal{I},j\in\mathcal{J}}$ that satisfy $s_{ijr} \geq 0$ $(\forall(i, j, r))$ and $\sum_r s_{ijr} = 1$ $(\forall(i, j))$:

$$\mathcal{F}(\theta, \boldsymbol{S}; w, \lambda) = \sum\nolimits_{i\in\mathcal{I}}\sum\nolimits_{j\in\mathcal{J}}\Big[\frac{\gamma^U_{ij}}{2}\sum\nolimits_{r=1}^{n_R}\frac{(a_{ir}b_{rj})^2}{s_{ijr}} - z_{ij}\hat{x}_{ij}\Big].$$

It can be verified that this majorizing function has two properties: (i) $\hat{\mathcal{R}}_{SE}(\theta; w, \lambda) \leq \mathcal{F}(\theta, \boldsymbol{S}; w, \lambda)$, (ii) $\hat{\mathcal{R}}_{SE}(\theta; w, \lambda) = \min_{\boldsymbol{S}}\mathcal{F}(\theta, \boldsymbol{S}; w, \lambda)$, where the equality holds if and only if $s_{ijr} = a_{ir}b_{jr}/(\sum_{r'} a_{ir'}b_{jr'})$. In the scheme of MM, minimization of function $\hat{\mathcal{R}}_{SE}$ is indirectly conducted by the following two steps:

(MM-step 1) Minimize $\mathcal{F}$ w.r.t. $\boldsymbol{A}$ or $\boldsymbol{B}$,

(MM-step 2) Minimize $\mathcal{F}$ w.r.t. $\boldsymbol{S}$ (to safisty $\mathcal{F} = \hat{\mathcal{R}}_{SE}$).

The necessary condition for a local minimum of $\mathcal{F}$ w.r.t. $\boldsymbol{A}$, which is the partial derivative $\frac{\partial\mathcal{F}}{\partial a_{ir}} = 0$, is simplified to $a_{ir} = (\sum_{j\in\mathcal{J}} z_{ij}b_{rj})/\{\sum_{j\in\mathcal{J}}\gamma^U_{ij}b^2_{rj}/s_{ijr}\}$. Substituting the equality condition of $\boldsymbol{S}$ into this, we get the update rule for $\boldsymbol{A}$ shown in Eq. (5). The update rule for $\boldsymbol{B}$ is derived in an analogous manner. The following theorem shows that a local minimum is obtained by iteratively updating $\boldsymbol{A}$ and $\boldsymbol{B}$.

Theorem 4.3. *The loss function $\hat{\mathcal{R}}_{SE}$ is monotonically decreasing under the update by Eq.* (5)*. The loss function is invariant if and only if $\boldsymbol{A}, \boldsymbol{B}$ are at a stationary point.*

Proof. Assume that $\theta = \{\boldsymbol{A}, \boldsymbol{B}\}, \boldsymbol{S}$ satisfy $\hat{\mathcal{R}}_{SE}(\theta) = \mathcal{F}(\theta, \boldsymbol{S})$ [3]. We denote the value of $\boldsymbol{A}$ after MM-step 1 as $\boldsymbol{A}^{new}$ and that of $\boldsymbol{S}$ after MM-step 2 as $\boldsymbol{S}^{new}$. Since $\mathcal{F}$ is convex w.r.t $\boldsymbol{A}$, $\mathcal{F}(\theta, \boldsymbol{S}) \geq \mathcal{F}(\{\boldsymbol{A}_{new}, \boldsymbol{B}\}, \boldsymbol{S})$ hold. The property of majorizing function $\mathcal{F}$ also yields $\mathcal{F}(\{\boldsymbol{A}_{new}, \boldsymbol{B}\}, \boldsymbol{S}) \geq \mathcal{F}(\{\boldsymbol{A}_{new}, \boldsymbol{B}\}, \boldsymbol{S}_{new}) = \hat{\mathcal{R}}_{SE}(\{\boldsymbol{A}_{new}, \boldsymbol{B}\})$. Then, we get $\hat{\mathcal{R}}_{SE}(\theta) \geq \hat{\mathcal{R}}_{SE}(\{\boldsymbol{A}_{new}, \boldsymbol{B}\})$. Since the proof for the update of $\boldsymbol{B}$ can be obtained in an analogous manner, we complete the proof. □

5 EXPERIMENTS

We conducted experiments on synthetic and real datasets to confirm the effectiveness of UNMF. Since UNMF is the first factorization method that can deal with uncoupled data, we investigate how closely the performance of UNMF can approach to that of *oracle* NMF using coupled data by increasing the size of PCD.

Synthetic data (SYNTH): We generated (true) matrix $\boldsymbol{X}^*$ whose sizes are $n_I{=}n_J{=}10$ by adding Gaussian noise with mean of 0.0 and s.t.d of 0.6 to the matrix $\tilde{X}$ whose (i, j)-th element is $2.0{+}\frac{j+1}{2}\mathbb{I}(i, j \leq 4) + \frac{j-4}{2}\mathbb{I}(i, j \geq 5)$. We prepared five data sets by dividing the elements of $\boldsymbol{X}^*$ into five, using 80% of the data as a training set and the remaining 20% as a test set. Removing the index-value relationship from the training data, we made MVD $\mathcal{D}_U$ and MID $\mathcal{D}_V$. PCD $\mathcal{D}_C$ were also made by randomly extracting $n_{per}{=}\{6, 8, 10, \cdots, 28\}$ pairs of columns from each row in the training data.

Real data (ML and SUSHI): We used MovieLens (ML) [4] and sushi preference data (SUSHI) [14] [5]. ML includes users' review scores of movies ranging from 1.0 (min) to 5.0 (max). SUSHI also includes users' preference scores of sushis ranging from 0.0 (min) to 4.0 (max). By taking the average of the rating score of each user for each movie-genre/sushi-minor-group, we constructed user×genre /group rating matrix $\boldsymbol{X}^*$ whose size is $n_I = 943, n_J = 18$ for ML and $n_I = 5000, n_J = 11$ for SUSHI. Similar to the synthetic data, we prepared five data sets by dividing the data and using 80% of the data as a training set and 20% as a test set, and made MVD MID, and PCD [6]. PCD $\mathcal{D}_C$ were made by randomly extracting n_{per} pairs of columns from each row in the training data. Note that when the values of the chosen indexes are equivalent, e.g., $x_{ij} = x_{ij'}$, we created two pairs $(i_m = i, j^+_m = j, j^-_m = j')$ and $(i_{m'}, j^+_{m'} = j', j^-_{m'} = j)$ [7].

Evaluation metric: As the performance metric, we adopted test mean absolute error (Test MAE) defined as $\frac{1}{|\mathcal{D}_{test}|}\sum_{(i_m,j_m,x_m)\in\mathcal{T}}|x_m - \hat{x}_{i_m j_m}|$, where $\mathcal{T}$ is the set of element indexes in the test set and $|\cdot|$ indicates the number of elements in the set. We used UNMF with $\varphi(x) = x^2$ (i.e., squared error) and $\lambda_i = 0$ or $\lambda_i = (w_{i1} + w_{i2})/2$.

Baselines: The performance of UNMF is compared with that of NMF using coupled data made from training data (ORACLE NMF); PCD were not used for NMF. The number of factors, n_R, for UNMF and ORACLE-NMF was set to 2 for SYNTH and to 3 for ML and SUSHI in common. The performance of random prediction (RANDOM), which takes the value on $[0.5, 6.5]$ (SYNTH), $[1.0, 5.0]$ (ML) or $[0.0, 4.0]$ (SUSHI) uniformly, is also shown as a benchmark.

---

[1] In experiment, we added $w_{i1}, w_{i2} \geq 0$ which is a sufficient condition for the nonnegativity of z_{ij} when $\lambda_i = 0$. We also use this when $\lambda_i = (w_{i1} + w_{i2})/2$ for fair comparison since the nonnegativity was confirmed experimentally.

[2] The derivation for KL (Eq. (6)) and the analysis follow in an analogous manner.

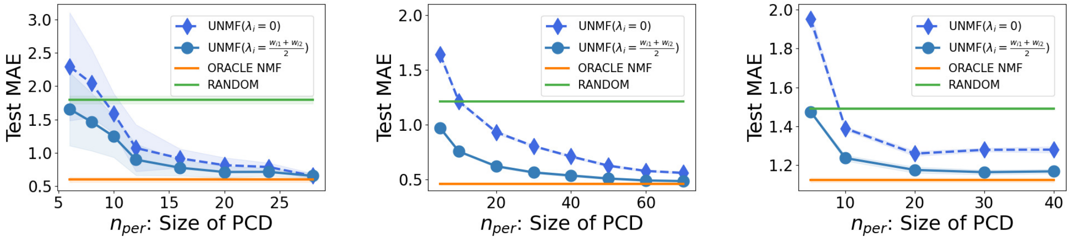
[3] Note that we omit the notation of w and λ.

[4] https://grouplens.org/datasets/movielens/100k/

[5] https://www.kamishima.net/sushi/

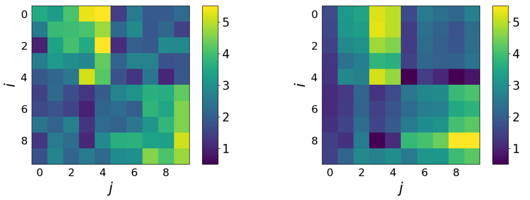
[6] Precisely speaking, we randomly extracted $\min(n_{per}, \binom{\ell_i}{2})$ pairs where ℓ_i is the number of training data whose row-index is i. The total number of PCD in ML is roughly $n_C \approx n_I n_{per}$.

[7] The reason why we use genre/group-level matrix is to avoid that many pairs with equal values are generated when creating PCD; the use of the item-level rating scores, which are discrete values, can generate many pairs with equal values.



(a) SYNTH (b) ML (c) SUSHI

Figure 1: Result of test MAE performance for (a) synthetic data (SYNTH), (b) MovieLens (ML), and (c) Sushi preference data (SUSHI) Average and standard deviation are shown. Lower values are better.



(a) True matrix $\boldsymbol{X}^*$ (b) Estimated matrix $\hat{\boldsymbol{X}}(=\boldsymbol{AB})$

Figure 2: Visualization of (a) true matrix X^* and (b) estimated matrix $\hat{X}$ for synthetic data experiment when $n_{per} = 12$.

Results: Figure 1 shows the results of the three experiments. In the experiments, the performance of UNMF improves with the number of PCD, and approaches the performance of ORACLE NMF. This is an amazing result since UNMF does not use coupled data. Moreover, we can confirm that UNMF with $\lambda_i = \frac{w_{i1}+w_{i2}}{2}$ shows stable performance even when only small amount of PCD is available. Figure 2 also shows that UNMF well recovered the true matrix $\boldsymbol{X}^*$. These results imply that UNMF is effective in estimating factor matrices from MVD, MID and PCD.

6 CONCLUSION

In this paper, we proposed UNMF in order to analyze uncoupled data that cannot be represented by a matrix. We derived the loss function that can be evaluated using uncoupled data and PCD from the expected Bregman divergence, and provided the multiplicative update algorithm with the theoretical support. The effectiveness of the proposal was confirmed by experiments on both synthetic and real data. Future work for this research includes examining other divergences and a Bayesian extension.

REFERENCES

[1] A. Abid, A. Poon, and J. Zou. 2017. Linear regression with shuffled labels. *arXiv preprint arXiv:1705.01342* (2017).
[2] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty. 2005. Clustering with Bregman divergences. *Journal of machine learning research* 6, 10 (2005).
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[4] L. M. Bregman. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics* 7, 3 (1967), 200–217.
[5] A. Carpentier and T. Schlüter. 2016. Learning relationships between data obtained independently. In *Artificial Intelligence and Statistics*. 658–666.
[6] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. 2009. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons.
[7] J. De Leeuw. 1994. Block-relaxation algorithms in statistics. In *Information systems and data analysis*. Springer, 308–324.
[8] I. S. Dhillon and S. Sra. 2005. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in Neural Information Processing Systems*. 283–290.
[9] C. Ding, X. He, and H. D. Simon. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM international conference on data mining*. 606–610.
[10] C. Ding, T. Li, and W. Peng. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis* 52, 8 (2008), 3913–3927.
[11] T. Hofmann. 1999. Probabilistic latent semantic indexing. In *International ACM SIGIR conference on Research and development in information retrieval*. 50–57.
[12] D. Hsu, K. Shi, and X. Sun. 2017. Linear regression without correspondence. In *Advances in Neural Information Processing Systems*. 1530–1539.
[13] D. R. Hunter and K. Lange. 2004. A tutorial on MM algorithms. *The American Statistician* 58, 1 (2004), 30–37.
[14] T. Kamishima. 2003. Nantonac collaborative filtering: recommendation based on order responses. In *ACM SIGKDD international conference on Knowledge discovery and data mining*. 583–588.
[15] M. Kohjima, T. Matsubayashi, and H. Sawada. 2015. Probabilistic Non-negative Inconsistent-resolution Matrices Factorization. In *ACM international conference on Information and knowledge management*. 1855–1858.
[16] D. D. Lee and H. S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.
[17] D. D. Lee and H. S. Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*. 556–562.
[18] A. McCallum, K. Nigam, and L. H. Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *ACM SIGKDD international conference on knowledge discovery and data mining*. 169–178.
[19] A. E. Monge and C. Elkan. 1996. The field matching problem: algorithms and applications.. In *International conference on knowledge discovery and data mining*, Vol. 2. 267–270.
[20] A. Pananjady, M. J. Wainwright, and T. A. Courtade. 2017. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory* 64, 5 (2017), 3286–3300.
[21] S. Sra and I. S. Dhillon. 2006. Nonnegative matrix approximation: Algorithms and applications. In *Technical Report # TR-06-27*. Computer Science Department, University of Texas at Austin.
[22] L. Xu, G. Niu, J. Honda, and M. Sugiyama. 2019. Uncoupled regression from pairwise comparison data. In *Advances in Neural Information Processing Systems*. 3992–4002.
[23] W. Xu, X. Liu, and Y. Gong. 2003. Document clustering based on non-negative matrix factorization. In *International ACM SIGIR conference on Research and development in information retrieval*. 267–273.
[24] S. Zhang, W. Wang, J. Ford, and F. Makedon. 2006. Learning from incomplete ratings using non-negative matrix factorization. In *SIAM international conference on data mining*. 549–553.