Ranking Preserving Nonnegative Matrix Factorization

Jing Wang[1,2], **Feng Tian**[2], **Weiwei Liu**[3], **Xiao Wang**[4,*], **Wenjie Zhang**[3] and **Kenji Yamanishi**[1]
[1] Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2] Faculty of Science and Technology, Bournemouth University, UK
[3] School of Computer Science and Engineering, The University of New South Wales, Australia
[4] School of Computer Science, Beijing University of Posts and Telecommunications, China
jing_wang@mist.i.u-tokyo.ac.jp, ftian@bournemouth.ac.uk, liuweiwei863@gmail.com,
wangxiao_cv@tju.edu.cn, wenjie.zhang@unsw.edu.au, yamanishi@mist.i.u-tokyo.ac.jp

Abstract

Nonnegative matrix factorization (NMF), a well-known technique to find parts-based representations of nonnegative data, has been widely studied. In reality, ordinal relations often exist among data, such as data *i* is more related to *j* than to *q*. Such relative order is naturally available, and more importantly, it truly reflects the latent data structure. Preserving the ordinal relations enables us to find structured representations of data that are faithful to the relative order, so that the learned representations become more discriminative. However, this cannot be achieved by current NMFs. In this paper, we make the first attempt towards incorporating the ordinal relations and propose a novel ranking preserving nonnegative matrix factorization (RPNMF) approach, which enforces the learned representations to be ranked according to the relations. We derive iterative updating rules to solve RPNMF's objective function with convergence guaranteed. Experimental results with several datasets for clustering and classification have demonstrated that RPNMF achieves greater performance against the state-of-the-arts, not only in terms of accuracy, but also interpretation of orderly data structure.

1 Introduction

As a well-known approach for finding parts-based representations of non-negative data, nonnegative matrix factorization (NMF) [Lee and Seung, 1999] has shown remarkable competitiveness in variety of applications, such as clustering [Wang *et al.*, 2017b] and classification [Zhang *et al.*, 2015].

Recent advances in NMF could be roughly categorized into the unsupervised and the semi-supervised. The former [Cai *et al.*, 2011; Wang *et al.*, 2017b] mainly focuses on the features of data. In addition to features, the later [Liu *et al.*, 2012; Zhang *et al.*, 2016] utilizes a small amount of supervision information to achieve more discriminative representation learning. Examples of such information are category labels or pairwise relations including must-link and cannot-link, which specify whether the data must be or cannot be in the same class. Semi-supervised NMFs have benefited from the supervision information. However, existing supervisions suffer two major limitations. One is that both labels and pairwise relations are absolute information, so they are non-trivial to obtain if there is no prior knowledge provided; the other is that these supervisions only characterize the relationships between data and class, yet data usually contain rich information beyond what they can describe. To this end, we are inspired to tackle these limitations and further achieve more accurate learning by exploring data information more deeply.

In reality, one may easily observe that ordinal relations are ubiquitous among data. Given three data *i*, *j* and *q*, an ordinal relation represents that *i* is more related to *j* than to *q*. For example, an image of "apple" is often more related to that of "banana" than to that of "ball"; a frame of a video sequence is often more related to its neighbouring frames than to those far way. Such comparative relation is naturally available and reliable [Liu *et al.*, 2016]. It reflects the relative order among data and therefore can also uncover latent data structure. Furthermore, recent studies in numerous research fields such as ordinal embedding [Le and Lauw, 2016], hashing [Liu *et al.*, 2016] and social networks [Song *et al.*, 2015], have demonstrated that the ordinal relation plays an important role in performance improvements. Unfortunately, the ordinal relation is largely ignored by existing NMFs.

In this paper, we propose a novel ranking preserving NMF (RPNMF) to learn orderly structured representations. Figure 1 outlines RPNMF with a two-class dataset with each class being marked in red or blue. The original data representation matrix $\mathbf{X}$ may not reflect the true data structure, since real data are complex and often contaminated by noises. However, given a set of ordinal relations such as data i is more related to j than to q, i.e., $R(i,j) > R(i,q)$, RPNMF can effectively enforce the distances among their corresponding representations to have the relation $D(\mathbf{h}_i, \mathbf{h}_j) < D(\mathbf{h}_i, \mathbf{h}_q)$. Moreover, from Figure 1, we can easily deduce $R(i,j) > R(i,q) >$

*Corresponding author

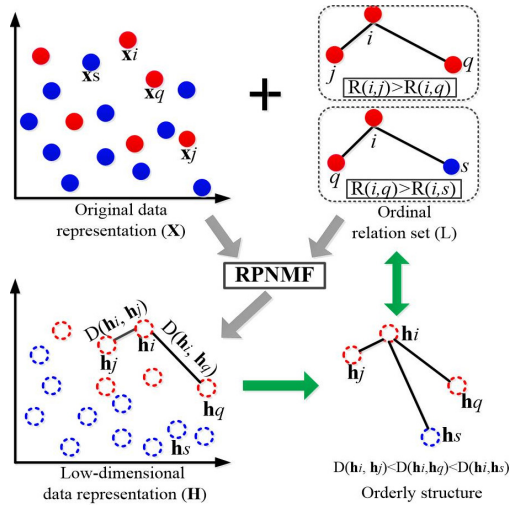


Figure 1: Outline of the proposed RPNMF.

$R(i, s)$. Accordingly, the representation of data i, i.e., $\mathbf{h}_i$, is enforced to be the closest to $\mathbf{h}_j$ and the furthest from $\mathbf{h}_s$, i.e., $D(\mathbf{h}_i, \mathbf{h}_j) < D(\mathbf{h}_i, \mathbf{h}_q) < D(\mathbf{h}_i, \mathbf{h}_s)$. As a result, a more discriminative representation matrix $\mathbf{H}$ is achieved with orderly structure rendered as a whole.

The main contributions of this paper are as follows:

- To our best knowledge, RPNMF is the first to incorporate the ordinal relation into NMF for representation learning. Due to the natural availability of the ordinal relation, RPNMF is also practical for real applications.
- With the ordinal relation, RPNMF effectively ranks the learned representations according to the true data structure of data, so that more discriminative representations are obtained.
- We derive an efficient linear iterative updating rule to solve the RPNMF's objective function, with its convergence guaranteed.
- Extensive experiments have demonstrated that RPNMF unifies the process of representation learning with orderly structure preserving, leading to more accurate clustering and classification against the state-of-the-arts.

2 Related Work

Several variants of NMF have been proposed to seek for more effective data representation in recent years. Under the unsupervised setting, [Cai *et al.*, 2011] proposed a graph regularized NMF (GNMF) to model the local manifold structure. Later, MCNMF [Wang *et al.*, 2017b] explores diverse information among multiple components (sub-features) of data. Under the semi-supervised setting, CNMF [Liu *et al.*, 2012] ensures data with the same label to have the same representations with the utilization of labels. CPSNMF [Wang *et al.*, 2016] propagates the pairwise relations from supervised data to unsupervised data to obtain the supervisions of the entire dataset. As one of the most representative pairwise NMFs, NMFCC [Zhang *et al.*, 2016] enforces the similarity/dissimilarity for data on a must-link/cannot-link. Although pairwise NMFs incorporate the relations of data as supervisions, our RPNMF works quite differently and is actually more advanced in three main aspects. Firstly, the must-link/cannot-link is limited to paired data which belong to the same class or two different ones, but the ordinal relation is a triplet information either within one class or across multiple ones. Secondly, with must-link/cannot-link, pairwise NMFs enforce representations to be close/far away, but RPNMF takes a step further by preserving relative order among data, even when they are from the same class. Thirdly, two pairwise relations may derive an order relation, such as if i is must-link to j but cannot-link to q, we can derive i is more related to j than to q. However, the reverse derivation may not hold. That is to say, pairwise relations could be utilized by RPNMF, but pairwise NMFs cannot effectively incorporate ordinal relations. An elaborate review of existing NMFs can be found in [Wang and Zhang, 2013]. However, none of them pay attention to the ordinal relations among data.

The ordinal relation was originally explored in ordinal embedding. The goal of ordinal embedding is to learn low-dimensional representations by utilizing ordinal relations only. Recently, a number of approaches have been proposed from different aspects. For example, LOE [Terada and Luxburg, 2014] simultaneously preserves the ordinal structure and the density structure of the data. MVTE [Amid and Ukkonen, 2015] uncovers multiple hidden attributes of data, since the data relations can be measured based on different attributes (e.g. colors, shapes). Thereafter, COE [Le and Lauw, 2016] preserves co-embedding cross multiple types of data. The most relevant work to ours is t-STE [Van Der Maaten and Weinberger, 2012], which collapses higher related data and repels data on a lower relation. However, t-STE and other approaches have not considered features of data.

3 Ranking Preserving NMF (RPNMF)

Given ordinal relations and a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, where each column is a data vector with m-dimensional features, we aim to learn a low-dimensional representation matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n] \in \mathbb{R}^{k \times n}$ of $\mathbf{X}$ with the ordinal relations preserved, where k (usually $k \ll \min\{m, n\}$) denotes the reduced dimension.

3.1 Preliminary

NMF seeks for a basis matrix $\mathbf{W}$ and a representation matrix $\mathbf{H}$, where the product of the two matrices can well approximate the original matrix $\mathbf{X}$, i.e., $\mathbf{X} \approx \mathbf{WH}$. Mathematically, NMF solves the following objective function:

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} \|\mathbf{X} - \mathbf{WH}\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ denotes Frobenius norm. Then the multiplicative algorithm is derived to infer $\mathbf{W}$ and $\mathbf{H}$ [Lee and Seung, 2001]. Obviously, the representation learning process of NMF depends on the features of data only. The learned $\mathbf{H}$ may not maintain ordinal structure when ordinal relations are available, as proven in the Proposition 1.

Proposition 1 Given an ordinal relation, i.e., data i is more related to j than to q, NMF cannot ensure the corresponding representation $\mathbf{h}_i$ be closer to $\mathbf{h}_j$ than to $\mathbf{h}_q$.

Proof. Here, we use the "proof by contradiction" with assuming that given an ordinal relation, i.e., data i is more related to j than to q. NMF can ensure the corresponding representation $\mathbf{h}_i$ be closer to $\mathbf{h}_j$ than to $\mathbf{h}_q$.

To prove this, we use the following counter example.

Given that

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0.9 & 1 \\ 1 & 1 & 1 & 3 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 4 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

of which, the first three columns $\mathbf{x}_i$, $\mathbf{x}_j$ or $\mathbf{x}_q$ are corresponding to data i, j and q, respectively. According to (1), we obtain a basis matrix $\mathbf{W}$ and a representation matrix $\mathbf{H}$ with

$$\mathbf{W} = \begin{bmatrix} 0.2661 & 1.5532 & 0.4821 \\ 0.9419 & 0.8432 & 1.3544 \\ 0.3476 & 1.6137 & 0.3875 \\ 1.0476 & 0.4657 & 2.0592 \\ 0.2785 & 1.6166 & 0.4675 \end{bmatrix}$$

and

$$\mathbf{H} = \begin{bmatrix} 0.4413 & 0.0642 & 0.3550 & 1.4064 \\ 0.5042 & 0.5211 & 0.4858 & 0.0233 \\ 0.1377 & 0.3449 & 0.1934 & 1.2220 \end{bmatrix}.$$

Hence the distance between i and j in original space is smaller than the distance between i and q, i.e., $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 < \|\mathbf{x}_i - \mathbf{x}_q\|_2^2$. Whereas the distance between i and j, as well as i and q in representation space are $\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 = 1.1854$ and $\|\mathbf{h}_i - \mathbf{h}_q\|_2^2 = 0.0109$, respectively. Therefore, it is easy to check that given data i is more related to j than to q, $\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 > \|\mathbf{h}_i - \mathbf{h}_q\|_2^2$, which means the assumption does not hold. Therefore, Proposition 1 is proven.

In the following, we introduce our ranking preserving NMF (RPNMF) to embed ordinal relations so that more discriminative representations could be achieved.

3.2 RPNMF-model

We use $R(i,j) > R(i,q)$ to denote an ordinal relation, i.e., data *i* is more related to *j* than to *q*. The objective of RPNMF is to enforce the representations of the data *i* and *j* to be closer than that of data *i* and *q* , i.e.,

$$R(i,j) > R(i,q) \Rightarrow D(\mathbf{h}_i, \mathbf{h}_j) < D(\mathbf{h}_i, \mathbf{h}_q), \quad (2)$$

where D denotes the distance measure. To achieve so, we minimize the following term:

$$I(R_+(i,j))I(D(\mathbf{h}_i, \mathbf{h}_j) \geq D(\mathbf{h}_i, \mathbf{h}_q))I(R_-(i,q)). \quad (3)$$

Here, $R_+(\cdot)/R_-(\cdot)$ indicates a higher/lower relative relation between two data. $I(\cdot)$ is an indicator which equals to "1" if the condition in the parenthesis is satisfied and "0" otherwise. It is clear to see that when $R(i,j) > R(i,q)$, minimizing (3) enforces $D(\mathbf{h}_i, \mathbf{h}_j) < D(\mathbf{h}_i, \mathbf{h}_q)$.

Use L to represent a set of ordinal relations, then (3) could be extended for $\forall(i,j,q) \in L$ as follows,

$$\sum_{(i,j,q)\in L} I(R_+(i,j))I(D(\mathbf{h}_i, \mathbf{h}_j) > D(\mathbf{h}_i, \mathbf{h}_q))I(R_-(i,q)). \quad (4)$$

Using Euclidean distance for D, we can rewrite (4) as

$$\sum_{(i,j,q)\in L} I(R_+(i,j))I(\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 > \|\mathbf{h}_i - \mathbf{h}_q\|_2^2)I(R_-(i,q)). \quad (5)$$

To ensure a distance between $\|\mathbf{h}_i - \mathbf{h}_j\|_2^2$ and $\|\mathbf{h}_i - \mathbf{h}_q\|_2^2$, a tunable threshold $\delta > 0$ is incorporated to regulate the distances in-between as

$$\sum_{(i,j,q)\in L} I(R_+(i,j))I(\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 > (\|\mathbf{h}_i - \mathbf{h}_q\|_2^2 - \delta))I(R_-(i,q)). \quad (6)$$

Because (6) is non-continuous, we use a ReLU loss function $f(t) = max(0,t)$ and obtain

$$\sum_{(i,j,q)\in L} I(R_+(i,j))max(0,t)I(R_-(i,q)), \quad (7)$$

where $t = \|\mathbf{h}_i - \mathbf{h}_j\|_2^2 - (\|\mathbf{h}_i - \mathbf{h}_q\|_2^2 - \delta)$. Clearly, when $R(i,j) > R(i,q)$, (7) penalizes t to encourage $\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 - \|\mathbf{h}_i - \mathbf{h}_q\|_2^2 \to -\delta$. However, since (7) is non-differentiable with respect to $\mathbf{h}_i$, $\mathbf{h}_j$ and $\mathbf{h}_q$, we use a Softplus loss function $f(t) = log(1 + exp(t))$ to approximate ReLU and get

$$\sum_{(i,j,q)\in L} I(R_+(i,j))log(1 + exp(t))I(R_-(i,q)). \quad (8)$$

By incorporating (8) into NMF, we obtain the objective function as follows,

$$\begin{aligned} \mathcal{F} = & \min_{\mathbf{W},\mathbf{H}\geq 0} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \\ & + \alpha \sum_{(i,j,q)\in L} I(R_+(i,j))log(1 + exp(t))I(R_-(i,q)), \end{aligned} \quad (9)$$

where the first term represents the errors between $\mathbf{X}$ and the product of $\mathbf{W}$ and $\mathbf{H}$. The second term is to maintain ordinal structure, and α is the trade-off parameter in between.

4 Optimization

In order to facilitate optimization, we rewrite (9) as

$$\begin{aligned} \mathcal{F} = & \min_{\mathbf{W},\mathbf{H}_\cup,\mathbf{h}_i,\mathbf{h}_j,\mathbf{h}_q\geq 0} \|\mathbf{X}_\cup - \mathbf{W}\mathbf{H}_\cup\|_F^2 + \sum_{i\in L} \|\mathbf{x}_i - \mathbf{W}\mathbf{h}_i\|_2^2 \\ & + \sum_{j\in L} \|\mathbf{x}_j - \mathbf{W}\mathbf{h}_j\|_2^2 + \sum_{q\in L} \|\mathbf{x}_q - \mathbf{W}\mathbf{h}_q\|_2^2 \\ & + \alpha \sum_{(i,j,q)\in L} I(R_+(i,j))log(1 + exp(t))I(R_-(i,q)), \end{aligned} \quad (10)$$

where $\mathbf{X}_\cup$ indicates the rest vectors without ordinal information and $\mathbf{H}_\cup$ is the corresponding representation matrix. The optimization problem in (10) is not convex, so it is non-trivial to find the global minimum. Here we divide (10) into several subproblems for alternately updating each variable with others fixed.

The optimizations of both **W-subproblem** and **$\mathbf{H}_\cup$-subproblem** lead to NMF formulation [Lee and Seung, 2001], so the updating rules for $\mathbf{W}$ and $\mathbf{H}_\cup$ are

$$\mathbf{W} \leftarrow \mathbf{W}\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}\mathbf{H}\mathbf{H}^T}, \quad (11)$$

$$\mathbf{H}_\cup \leftarrow \mathbf{H}_\cup\frac{\mathbf{W}^T\mathbf{X}_\cup}{\mathbf{W}^T\mathbf{W}\mathbf{H}_\cup}. \quad (12)$$

$\mathbf{h}_i$-subproblem: Updating $\mathbf{h}_i$ with other variables fixed leads to

$$\min_{\mathbf{h}_i \geq 0} \mathcal{F}(\mathbf{h}_i) = \|\mathbf{x}_i - \mathbf{W}\mathbf{h}_i\|_2^2 + \alpha I(R_+(i,j)) \sum_{(j,q)\in L} log(1+exp(t))I(R_-(i,q)). \tag{13}$$

By setting the derivative of $\mathcal{F}(\mathbf{h}_i)$ with respect to $\mathbf{h}_i$, we get the positive part ∇_+ and the negative part ∇_- as following,

$$\nabla_+ = \mathbf{W}^T\mathbf{x}_i + \alpha I(R_+(i,j)) \sum_{(j,q)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_q I(R_-(i,q)),$$
$$\nabla_- = \mathbf{W}^T\mathbf{W}\mathbf{h}_i + \alpha I(R_+(i,j)) \sum_{(j,q)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_q I(R_-(i,q)). \tag{14}$$

Based on the coordinate descent algorithm [Tan and Févotte, 2009], multiplying the ratio of ∇_+ to ∇_- in (14) with $\mathbf{h}_i$ leads to the following updating rule:

$$\mathbf{h}_i \leftarrow \mathbf{h}_i \frac{\mathbf{W}^T\mathbf{x}_i + \alpha I(R_+(i,j)) \sum_{(j,q)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_j I(R_-(i,q))}{\mathbf{W}^T\mathbf{W}\mathbf{h}_i + \alpha I(R_+(i,j)) \sum_{(j,q)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_q I(R_-(i,q))}. \tag{15}$$

Similarly, we obtain the updating rules for **$\mathbf{h}_j$-subproblem** and **$\mathbf{h}_q$-subproblem** as

$$\mathbf{h}_j \leftarrow \mathbf{h}_j \frac{\mathbf{W}^T\mathbf{x}_j + \alpha I(R_+(i,j)) \sum_{(i,q)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_i I(R_-(i,q))}{\mathbf{W}^T\mathbf{W}\mathbf{h}_j + \alpha I(R_+(i,j)) \sum_{(i,q)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_j I(R_-(i,q))}, \tag{16}$$

$$\mathbf{h}_q \leftarrow \mathbf{h}_q \frac{\mathbf{W}^T\mathbf{x}_q + \alpha I(R_+(i,j)) \sum_{(i,j)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_q I(R_-(i,q))}{\mathbf{W}^T\mathbf{W}\mathbf{h}_q + \alpha I(R_+(i,j)) \sum_{(i,j)\in L} \frac{exp(t)}{1+exp(t)} \mathbf{h}_i I(R_-(i,q))}. \tag{17}$$

4.1 Convergence Analysis

We have solved problem (10) by optimizing each subproblem alternately. Now, we analyze the convergence of the above and prove the value of the objective function is non-increasing under each updating rule in Proposition 2.

Proposition 2. The alternating optimization of (10) converges to local minimum.

Proof. Since the convergence of updating rules of $\mathbf{W}$ and $\mathbf{H}_\cup$ can be guaranteed by NMF [Lee and Seung, 2001], here we only study the convergence of (10) under the rules of $\mathbf{h}_i$, $\mathbf{h}_j$ and $\mathbf{h}_q$. Denoting the solution of the objective function $\mathcal{F}$ in (10) at the iteration round t as $(\mathbf{W}^{(t)}, \mathbf{H}_\cup^{(t)}, \mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)}, \mathbf{h}_q^{(t)})$. For the other variables being fixed with the value solved in the $(t-1)^{th}$ step, the minimization of (10) $w.r.t.$ $\mathbf{h}_i$ at the iteration round t is turned into (13), which is a standard convex loss function. Therefore, we can deduce that, at the iteration round t, the solution $\mathbf{h}_i^{(t)}$ satisfies $\mathcal{F}(\mathbf{W}^{(t)}, \mathbf{H}_\cup^{(t)}, \mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)}, \mathbf{h}_q^{(t)}) \leq \mathcal{F}(\mathbf{W}^{(t-1)}, \mathbf{H}_\cup^{(t-1)}, \mathbf{h}_i^{(t-1)}, \mathbf{h}_j^{(t-1)}, \mathbf{h}_q^{(t-1)})$, with $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)}$, $\mathbf{H}_\cup^{(t)} = \mathbf{H}_\cup^{(t-1)}$, $\mathbf{h}_j^{(t)} = \mathbf{h}_j^{(t-1)}$, $\mathbf{h}_q^{(t)} = \mathbf{h}_q^{(t-1)}$. Similarly, we can prove that the above situation can also be satisfied when minimizing (10) $w.r.t.$ $\mathbf{h}_j^{(t)}$ and $\mathbf{h}_q^{(t)}$ with other variables fixed. Therefore, the objective function $\mathcal{F}(\mathbf{W}, \mathbf{H}_\cup, \mathbf{h}_i, \mathbf{h}_j, \mathbf{h}_q)$ will monotonically decrease and the alternating iterations converge.

4.2 Complexity Analysis

The computations of updating (11) and (12) are $\mathcal{O}(mkn)$ and $\mathcal{O}(mkn_l)$, respectively, where n_l represents the number of columns of $\mathbf{X}_\cup$. As to (15), we use g_i to represent the number of relations associated with $\mathbf{h}_i$. Similarly, g_j and g_q are used for (16) and (17). The cost for updating (15), (16) and (17) are $\mathcal{O}(k(m+g_i))$, $\mathcal{O}(k(m+g_j))$ and $\mathcal{O}(k(m+g_q))$. We also use n_i, n_j and n_q to denote the number of non-repreated data i, j and q in the set L, respectively. So the total costs for $\mathbf{h}_i$, $\mathbf{h}_j$ and $\mathbf{h}_q$ are $\mathcal{O}(k(m+g_i)n_i)$, $\mathcal{O}(k(m+g_j)n_j)$ and $\mathcal{O}(k(m+g_q)n_q)$, respectively. Since $\max\{n_l, n_i, n_j, n_q\} < n$, the overall computation cost of RPNMF is $\mathcal{O}(k(mn + g_i n_i + g_j n_j + g_q n_q))$. Therefore, RPNMF has linear complexity $w.r.t$ the number of data n.

5 Experiments

To demonstrate the effectiveness of RPNMF, we conducted both clustering and classification experiments not only on image benchmarks, i.e., Yale, ORL, Coil20 and NHill, but also on sequential datasets, including a cartoon video sequence and Hdm motion capture.

5.1 Datasets

The **Yale** [Liu *et al.*, 2012] contains 11 face images for each of 15 subjects. Each subject's face images are in different facial expressions or configurations. The **ORL** [Liu *et al.*, 2012] consists of 400 face images of 40 different subjects. Similar to the Yale, the images were taken with various lighting and facial expressions. The **Coil20** [Wang *et al.*, 2017b]is composed of 1440 images for 20 objects. The 72 images of each object were captured by a fixed camera at a pose intervals of 5 degree. The **NHill** [Wang *et al.*, 2017b] is a face dataset sampled from the movie "Notting Hill". The faces of 5 main casts were used, including 4660 faces in 76 tracks. The **Cartoon** [Wang *et al.*, 2017a] is a video sequence extracted from a short animation available online, which has 282 frames of three scenes. The **Hdm05** is a motion capture dataset. As in [Wang *et al.*, 2017a], we chose the scene 1-1 which contains 9842 frames and 14 activities.

5.2 Experiment Setup

We compared RPNMF with several state-of-the-arts, including unsupervised NMFs: NMF [Lee and Seung, 1999], RNMF [Kong *et al.*, 2011], GNMF [Cai *et al.*, 2011], ORNMF [Wang *et al.*, 2017a], MCNMF [Wang *et al.*, 2017b], a semi-supervised NMF: NMFCC [Zhang *et al.*, 2016]) and an ordinal embedding approach: t-STE [Van Der Maaten and Weinberger, 2012]. Worth to mention that, unlike existing semi-supervised NMFs which utilize absolute information as priors, RPNMF makes the first attempt to introduce comparative relations. Thus, it is infeasible to compare with existing semi-supervised NMFs due to different supervisions. Nevertheless, to demonstrate that the ordinal relation cannot be effectively utilized by existing pairwise NMFs, we compared

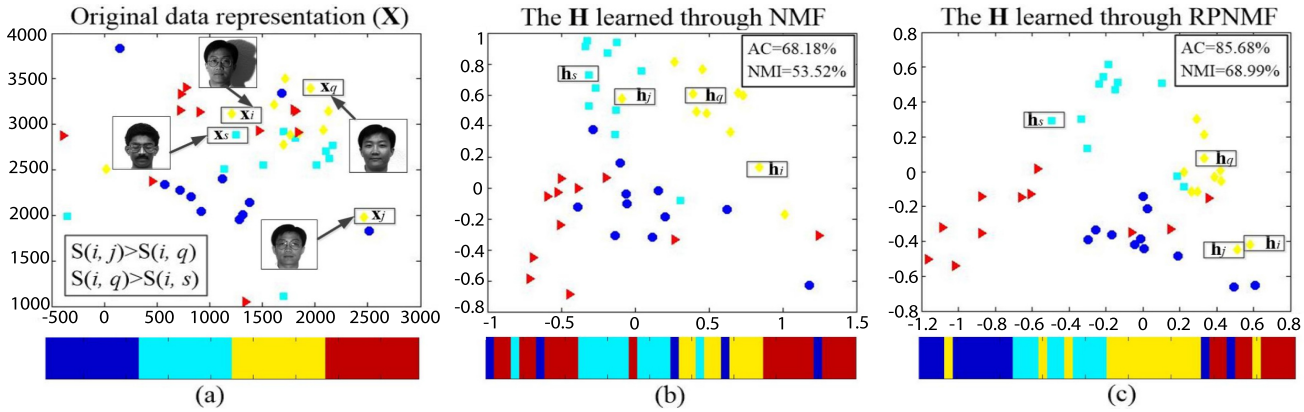


Figure 2: An example of the learned representations and clustering results of NMF and RPNMF. (a) The top figure represents original representations of face images from the dataset Yale and the bottom figure is the ground truth. The four colors represent four groups of images. Two examples of true relations among images, i.e., $R(i,j) > R(i,q)$ and $R(i,q) > R(i,s)$, were given by observing these images in terms of different subjects and with/without glass, respectively. For (b) and (c), the top figures are the learned $\mathbf{H}$ of NMF and RPNMF, and the corresponding bottom figures are clustering results based on $\mathbf{H}$. All representations displayed after the dimensionalities of their features are reduced to 2-D by PCA.

Metrics	Methods	Yale	ORL	Coil20	NHill	Cartoon	Hdm05
AC	t-STE	41.82	54.00	62.99	75.06	82.27	67.88
	NMF	41.55	54.90	62.49	77.04	77.78	60.72
	RNMF	38.55	54.20	59.57	74.08	77.57	58.21
	GNMF	41.58	59.60	68.39	75.88	74.46	61.14
	ORNMF	42.06	50.50	67.92	81.29	79.08	71.00
	MCNMF	42.42	58.28	65.14	77.54	75.89	67.49
	NMFCC	41.82	60.00	69.31	81.42	76.24	67.45
	RPNMF	**43.60**	**63.52**	**69.56**	**92.64**	**90.78**	**74.17**
NMI	t-STE	44.96	65.13	60.83	67.76	49.84	69.29
	NMF	46.52	76.22	74.35	65.27	66.65	68.78
	RNMF	43.98	75.33	73.24	64.74	65.33	65.16
	GNMF	46.30	77.80	77.30	62.97	63.48	71.93
	ORNMF	46.46	67.33	74.81	69.88	69.43	74.15
	MCNMF	46.83	69.71	74.31	66.63	57.16	70.65
	NMFCC	46.30	78.39	75.66	71.18	54.39	68.94
	RPNMF	**48.30**	**79.29**	**78.63**	**86.55**	**73.20**	**74.84**

Table 1: Clustering Results (%)

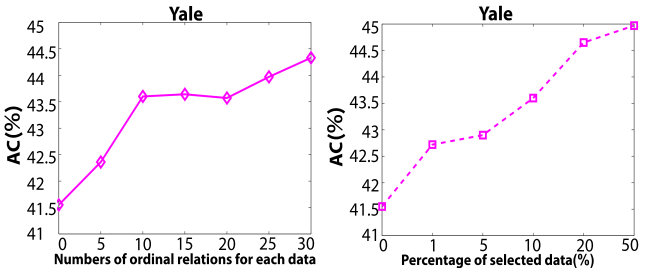
Metrics	Methods	Yale	ORL	Coil20	NHill	Cartoon	Hdm05
AC	t-STE	54.55	81.25	90.97	94.78	82.72	80.64
	NMF	57.58	95.00	93.40	91.09	81.93	84.98
	RNMF	48.48	88.75	94.44	79.61	82.14	71.39
	GNMF	42.42	92.50	81.60	91.31	82.14	82.88
	ORNMF	63.64	91.25	94.79	79.83	82.86	76.42
	MCNMF	64.70	95.00	86.46	92.49	83.54	76.88
	NMFCC	57.58	90.00	93.06	90.34	84.21	85.92
	RPNMF	**66.67**	**96.25**	**94.79**	**96.24**	**85.26**	**88.11**
F-score	t-STE	28.57	35.38	55.17	91.42	74.71	18.85
	NMF	30.00	60.00	59.57	53.41	74.30	18.68
	RNMF	19.05	40.00	68.00	64.02	72.22	11.62
	GNMF	24.00	50.00	39.08	73.02	72.22	17.60
	ORNMF	25.00	46.15	66.67	64.26	74.65	14.39
	MCNMF	28.57	60.00	40.00	85.23	75.43	13.33
	NMFCC	30.00	60.00	61.54	81.71	75.47	20.63
	RPNMF	**35.29**	**66.67**	**68.09**	**92.54**	**76.32**	**25.00**

Table 2: Classification Results (%)

RPNMF with NMFCC by splitting each ordinal relation into a pair of must-link and cannot-link. The parameters for each compared method were set according to the parameter settings in original papers. For RPNMF, we varied the regularization parameter α and δ within $\{0.0001, 0.001, 0.01, 0.1, 1\}$ and $\{0.001, 0.01, 0.1, 1, 10, 100\}$, respectively. To construct ordinal relations for t-STE and RPNMF, we first randomly selected 10% data for each dataset, and then constructed 30 ordinal relations for each selected data as in [Chang *et al.*, 2014]. Without losing generality, the ordinal relations were constructed from two aspects. For each image dataset, we use the data matrix $\mathbf{X}$ and labels because they can be directly used for constructing ordinal relations, although the relations can also be formed by observing images. In particular, one half were constructed with labels as data with the same label are more related than those with different ones. The other half were constructed from p-nearest neighbouring graph since data are usually more related to their nearest neighbours than those far away and p was set as 5 according to [Gong *et al.*, 2017]. For the sequential datasets, we chose the first 15 relations with each containing two frames from the same scene/activity and one from another. Each of the second 15 relations is formed by choosing from a scene/activity two neighbouring frames and one farther away. All the experiments were done using Matlab 2014 in an Intel Core 3.50GHZ desktop.

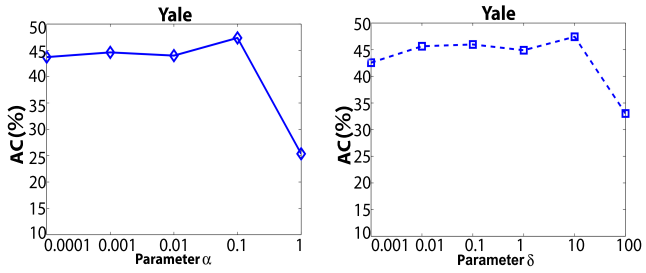
5.3 Results and Analysis

Clustering. We applied k-means to the learned representation matrix $\mathbf{H}$ and adopted two widely used metrics, accuracy (AC) [Liu *et al.*, 2017b] and normalized mutual information (NMI) [Liu *et al.*, 2012], to assess the quality of clustering results. Since k-means is sensitive to initial values, we repeated the clustering 50 times, each with a new set of initial



(a) (b)

Figure 3: Clustering AC of RPNMF $w.r.t$ ordinal relations.



(a) (b)

Figure 4: Clustering AC of RPNMF $w.r.t$ parameters α and δ.

centroid. Moreover, since all the compared methods converge to local minimum, we ran each method 10 times to avoid randomness. The average results are reported in Table 1. It can be seen that RPNMF achieves the best results on all datasets, which demonstrates the effectiveness of incorporating ordinal relations. Notably, NMFCC which incorporates pairwise relations by splitting the ordinal relations does not perform as well as RPNMF, proving that the ordinal relations cannot be directly or fully utilized by pairwise NMF approach.

Classification. We used $\mathbf{H}$ to classify data into a set of labels. For each dataset, 80% data from each class was randomly selected as training dataset and the rest as testing dataset. Similar to [Liu and Tsang, 2017; Liu *et al.*, 2017a], the LIBLINEAR package [Fan *et al.*, 2008] was used to train the classifiers. Same as clustering, we repeated each method 10 times and report the average AC and F-score [Pan *et al.*, 2016], shown in Table 2. As we can see, RPNMF consistently outperforms the other methods on all cases, which demonstrates the effectiveness of RPNMF on classification. This could be due to a fact that the ordinal relation represents the relative order among data, thus brings more discriminations.

Ordinal relations analysis. We closely examined $\mathbf{H}$ for NMF and RPNMF to analyze the effect of the ordinal relations. Due to page restriction, we took a subset of Yale as an example for the following analyses. The Figure 2(a) shows that $\mathbf{x}_i$ is closer to $\mathbf{x}_q$ than to $\mathbf{x}_j$, although there exists $R(i,j) > R(i,q)$. Seen from Figure 2(b), NMF still gives $D(\mathbf{h}_i, \mathbf{h}_j) > D(\mathbf{h}_i, \mathbf{h}_q)$, which experimentally proves that

NMF cannot maintain the ordinal structure and validates the Proposition 1. Consequently, it leads $\mathbf{h}_j$ to be close to $\mathbf{h_s}$ which belongs to a different group, so that an unsatisfied clustering result (AC $= 68.18\%$) occurs. In contrast, RPNMF effectively enforces $D(\mathbf{h}_i, \mathbf{h}_j) < D(\mathbf{h}_i, \mathbf{h}_q) < D(\mathbf{h}_i, \mathbf{h}_s)$ with $R(i,j) > R(i,q) > R(i,s)$ as shown in Figure 2(c). Apparently, both $\mathbf{h}_i$ and $\mathbf{h}_j$ represent the faces of the same subject with glass, which are definitely of highest relativity. Since $\mathbf{h}_s$ represents a face of a different subject, it is the least related to $\mathbf{h}_i$. Therefore, the $\mathbf{H}$ learned through RPNMF demonstrates a clearer structure of data and a more accurate clustering result (AC $= 85.86\%$) is achieved.

Since the number of ordinal relations may influence the performance of RPNMF, we also analyzed this in Figure 3 from two aspects. Specifically, we first selected 10% data from the Yale and varied the relations associated with each selected data from 0 to 30 with 5 interval as in Figure 3(a). We then varied percentages of data within $\{0\%, 1\%, 5\%, 10\%, 20\%, 50\%\}$ and fixed 30 ordinal relations for each selected data in Figure 3(b). Both figures show that, overall, AC increases with more supervisions which prove further the effectiveness of incorporating ordinal relations.

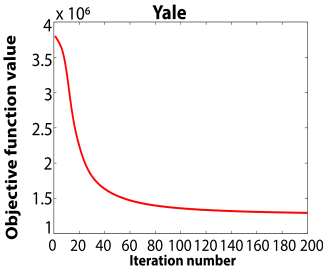


Figure 5: Convergence curve.

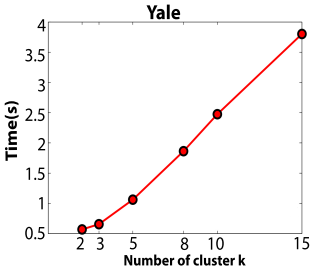


Figure 6: Execution time.

Parameter analysis. We tested the effect of parameter α and δ of RPNMF on the Yale. First, we fixed $\delta = 10$ to test α with varying from 0.0001 to 1, and then fixed $\alpha = 0.1$ to test δ varying from 0.001 to 100. Figure 4 shows that both α and δ perform with a similar trend. For example, AC is relatively stable when α increases from 0.0001 to 0.1 then drop sharply when $\alpha > 0.1$. This well demonstrates the robustness and effectiveness of RPNMF when both α and δ are chosen within a suitable range.

Computational speed analysis. Having proven the convergence of the updating rules of RPNMF in the section 4.1, here we experimentally demonstrated its convergence on the Yale in Figure 5. It can be seen that the objective function values are non-increasing and drop sharply within 200 iterations, which empirically validates the Proposition 2. As discussed in the section 4.2, RPNMF has linear complexity against the number of data n. To verify this claim, we varied the number of clusters k (each cluster contains 11 data) within {2, 3, 5, 8, 10, 15} and report the average execution time in Figure 6. Clearly, RPNMF is computationally linear.

6 Conclusion

In this paper, we have explored a new stream of semi-supervised NMF and proposed a novel ranking preserving nonnegative matrix factorization (RPNMF). Unlike existing approaches which utilize labels or pairwise relations as supervisions, RPNMF is the first to incorporate the relative or-

der among data, i.e., ordinal relation. By unifying the orderly structure preservation and representation learning, RPNMF explicitly ranks the representations according to the relations. Extensive experiments on both image and sequential datasets have demonstrated that RPNMF can not only uncover the true data structure which beyond what existing NMFs can offer, but also achieves more accurate clustering and classification against the state-of-the-arts. With the natural availability of the ordinal relation, RPNMF is also practical in real applications.

Acknowledgements

This work was supported by JST CREST (No. JPMJCR1304), National Natural Science Foundation of China (No. 61702296) and EU H2020 project (No. 691215).

References

[Amid and Ukkonen, 2015] Ehsan Amid and Antti Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *International Conference on Machine Learning*, pages 1472–1480, 2015.

[Cai *et al.*, 2011] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1548–1560, 2011.

[Chang *et al.*, 2014] Shiyu Chang, Guo-Jun Qi, Charu C Aggarwal, Jiayu Zhou, Meng Wang, and Thomas S Huang. Factorized similarity learning in networks. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 60–69. IEEE, 2014.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[Gong *et al.*, 2017] Chen Gong, Dacheng Tao, Wei Liu, Liu Liu, and Jie Yang. Label propagation via teaching-to-learn and learning-to-teach. *IEEE transactions on neural networks and learning systems*, 28(6):1452–1465, 2017.

[Kong *et al.*, 2011] Deguang Kong, Chris Ding, and Heng Huang. Robust nonnegative matrix factorization using l21-norm. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 673–682. ACM, 2011.

[Le and Lauw, 2016] Dung D Le and Hady W Lauw. Euclidean co-embedding of ordinal data for multi-type visualization. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 396–404. SIAM, 2016.

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

[Lee and Seung, 2001] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[Liu and Tsang, 2017] Weiwei Liu and Ivor W Tsang. Making decision trees feasible in ultrahigh feature and label dimensions. *The Journal of Machine Learning Research*, 18(1):2814–2849, 2017.

[Liu *et al.*, 2012] Haifeng Liu, Zhaohui Wu, Xuelong Li, Deng Cai, and Thomas S Huang. Constrained nonnegative matrix factorization for image representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1299–1311, 2012.

[Liu *et al.*, 2016] Hong Liu, Rongrong Ji, Yongjian Wu, and Wei Liu. Towards optimal binary code learning via ordinal embedding. In *AAAI*, pages 1258–1265, 2016.

[Liu *et al.*, 2017a] Weiwei Liu, Ivor W Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *The Journal of Machine Learning Research*, 18(1):3300–3337, 2017.

[Liu *et al.*, 2017b] Xinwang Liu, Miaomiao Li, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel k-means with incomplete kernels. In *AAAI*, pages 2259–2265, 2017.

[Pan *et al.*, 2016] Shirui Pan, Jia Wu, Xingquan Zhu, Chengqi Zhang, and Yang Wang. Tri-party deep network representation. *Network*, 11(9):12, 2016.

[Song *et al.*, 2015] Dongjin Song, David A Meyer, and Dacheng Tao. Top-k link recommendation in social networks. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 389–398. IEEE, 2015.

[Tan and Févotte, 2009] Vincent YF Tan and Cédric Févotte. Automatic relevance determination in nonnegative matrix factorization. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[Terada and Luxburg, 2014] Yoshikazu Terada and Ulrike Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855, 2014.

[Van Der Maaten and Weinberger, 2012] Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[Wang and Zhang, 2013] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1336–1353, 2013.

[Wang *et al.*, 2016] Di Wang, Xinbo Gao, and Xiumei Wang. Semisupervised nonnegative matrix factorization via constraint propagation. *IEEE transactions on cybernetics*, 46(1):233–244, 2016.

[Wang *et al.*, 2017a] Jing Wang, Feng Tian, Chang Hong Liu, Hongchuan Yu, Xiao Wang, and Xianchao Tang. Robust nonnegative matrix factorization with ordered structure constraints. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 478–485. IEEE, 2017.

[Wang *et al.*, 2017b] Jing Wang, Feng Tian, Xiao Wang, HC Yu, Changhong Liu, and Liang Yang. Multi-component nonnegative matrix factorization. International Joint Conferences on Artificial Intelligence, 2017.

[Zhang *et al.*, 2015] Xiang Zhang, Naiyang Guan, Zhilong Jia, Xiaogang Qiu, and Zhigang Luo. Semi-supervised projective nonnegative matrix factorization for cancer classification. *PloS one*, 10(9):e0138814, 2015.

[Zhang *et al.*, 2016] Xianchao Zhang, Linlin Zong, Xinyue Liu, and Jiebo Luo. Constrained clustering with nonnegative matrix factorization. *IEEE transactions on neural networks and learning systems*, 27(7):1514–1526, 2016.