Non-negative Matrix Factorization with Auxiliary Information on Overlapping Groups

Motoki Shiga and Hiroshi Mamitsuka, *Senior Member, IEEE*

Abstract—Matrix factorization is useful to extract the essential low-rank structure from a given matrix and has been paid increasing attention. A typical example is non-negative matrix factorization (NMF), which is one type of unsupervised learning, having been successfully applied to a variety of data including documents, images and gene expression, where their values are usually non-negative. We propose a new model of NMF which is trained by using auxiliary information of overlapping groups. This setting is very reasonable in many applications, a typical example being gene function estimation where functional gene groups are heavily overlapped with each other. To estimate true groups from given overlapping groups efficiently, our model incorporates latent matrices with the regularization term using a mixed norm. This regularization term allows group-wise sparsity on the optimized low-rank structure. The latent matrices and other parameters are efficiently estimated by a block coordinate gradient descent method. We empirically evaluated the performance of our proposed model and algorithm from a variety of viewpoints, comparing with four methods including MMF for auxiliary graph information, by using both synthetic and real world document and gene expression datasets.

Index Terms—Non-negative matrix factorization, auxiliary information, semi-supervised learning, sparse structured norm

1 Introduction

A general and prevalent data format in real world is a table or matrix, where columns are instances (or examples) and rows are their features, and vice versa. Matrix factorization is useful to extract essential low-rank structures from a given matrix and has been increasing interest in data mining and machine learning. A frequently-used matrix factorization is Principal Component Analysis (PCA), which is an eigenvalue decomposition that reduces the dimension of a feature space. Another well-used method is Singular Value Decomposition (SVD), by which informative low-rank structures of both instances and their features can be found. Indeed these methods are useful and easily implemented by basic linear algebra, but they are unable to be applied to several real-world applications, where matrix elements are all non-negatives. For example, documents, images and gene expression data have values of more than or equal to zero, and matrix factorization under non-negativity, which is called Non-negative Matrix Factorization (NMF), is useful for these types of data. In fact NMF has been considered in machine learning and data mining to detect informative and sparse low-rank structures of a given non-negative matrix, keeping the non-negativity in the low-rank matrices. Currently various approaches for NMF, such as scalable and fast algorithms for huge datasets, have been developed and successfully applied to a wide range of applications [1], [2], [3].

In many applications, available auxiliary information, such as groups of instances or network links over instances, can be given and in most cases improve the performance of clustering or classification [4], [5], [6]. A typical application is web page clustering, where page group information can be given and hyperlinks are also available. Such auxiliary information allows better clustering. This is a typical semi-supervised setting, and a variety of methods have been developed for semi-supervised learning in the past decade [7].

We address the issue of using overlapping groups as auxiliary information, where overlapping groups mean that instances can be included in more than one groups. This setting can be found in many applications. For example, documents can be assigned to more than one topics, such as news and sports. This is more pronounced when we use thesaurus or ontology, in which topics are hierarchical [8], [9], [3]. Another example of overlapping groups is gene functions, where many genes have more than one functions. In addition, gene function categories, such as MIPS functional categories [10], Gene Ontology (GO) [11], etc., are hierarchical, by which gene functions can be heavily overlapped. Thus overlapping groups are realistic and reasonable.

We propose a new NMF-based approach, which allows to consider auxiliary information on overlapping groups. The group information is, in the optimized cost function, a regularization term that is a mixed norm on a low rank matrix, consisting of ℓ_2 norms within groups and ℓ_1 norms between groups. A similar regularization term for overlapping

• *M. Shiga is with Informatics Course, Dept. of Electrical, Electronic and Computer Engineering, Faculty of Engineering, Gifu University, 1-1 Yanagido, Gifu, 501-1193, Japan. H. Mamitsuka is with Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, Japan.*
E-mail: shiga_m@gifu-u.ac.jp, mami@kuicr.kyoto-u.ac.jp

groups was already used for regression [12], in which parameters are vectors, while parameters in our NMF are matrices, implying that our regularization is a generalization of the regression case in [12].

We then present efficient algorithms for optimizing parameters, i.e. low-rank matrices, of our NMF model by using a block coordinate descent (BCD) method [13], which alternately optimizes blocks in the low-rank matrices, where we can consider two type of blocks: 1) vectors and 2) matrices. The optimization of each block is a convex problem, even with our regularization term. Thus we propose two ideas of using 1) latent matrices or 2) a dual problem without estimating latent matrices directly (we call this method a *direct* method). Latent matrices correspond to given auxiliary groups and at the same time are components of low-rank matrices to be optimized, and then the total cost is optimized through the latent matrices, allowing the optimization manner very smooth. The direct method does not use latent matrices explicitly and optimizes blocks directly in a dual problem manner, which requires a sequential optimization algorithm for a semi-definite problem. We thus develop totally four different algorithms by the combination of two block types and two ideas.

Empirically we first evaluated the efficiency of the four proposed algorithms by using well-organized synthetic datasets. Experimental results showed that the combination of latent matrices and of using vectors for blocks achieved the best performance. We then selected this combination for our proposed method hereafter. We further evaluated the effectiveness of our method on both synthetic and real datasets, comparing with three existing NMF-based methods and k-means. Experimental results showed that our method was clearly advantageous in performance against the competing methods. From these results we can conclude that our method favorably combines the input matrix with the auxiliary information on overlapping groups.

2 Semi-supervised NMF for Overlapping Groups

2.1 Notations

Let $\boldsymbol{X} \in \mathcal{R}_+^{M \times N}$ be a data matrix, where $\mathcal{R}_+$ is the set of all non-negative real numbers, M is the number of features (or row entities), and N is the number of instances (or column entities). The goal of NMF is a factorization of $\boldsymbol{X}$ into two low-rank matrices $\boldsymbol{U} \in \mathcal{R}_+^{M \times K}$ and $\boldsymbol{V} \in \mathcal{R}_+^{N \times K}$, where rank K satisfies $K < \min(M, N)$. Thus a basic factorization model of $\boldsymbol{X}$ is

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^T, \tag{1}$$

as shown in Fig. 1 (a). $\boldsymbol{U}$ and $\boldsymbol{V}$ are optimized by minimizing an approximation error such as the sum of squared errors or KL-divergence.

We have a set of groups[1] $\{\mathcal{G}_1, \ldots, \mathcal{G}_{G'}\}$ over instances. Assume that groups $\mathcal{G}_g$, where $g = 1, \ldots, G'$, are given auxiliary information, where each group $\mathcal{G}_g$ is a set of instances, *i.e.* $\mathcal{G}_g \subseteq \{1, 2, \ldots, N\}$. Fig. 1 (b) shows an illustrative example of instances with overlapping groups. Auxiliary groups can be overlapped, meaning that one instance belongs to more than one groups. We consider a semi-supervised setting, in which part of instances are unlabeled and not in any groups. That is, we set groups $\mathcal{G}_{G'+1}, \ldots, \mathcal{G}_G$, where each of these groups has an unique instance and is not overlapped with any other groups. In summary we have a set of groups $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$, in which the first G' groups $(\mathcal{G}_1, \ldots, \mathcal{G}_{G'})$ are given auxiliary groups and the rest $G - G'$ groups $(\mathcal{G}_{G'+1}, \ldots, \mathcal{G}_G)$ are with an unlabeled node.

Let $\boldsymbol{X}_{m\cdot}$ be the m-th row vector of $\boldsymbol{X}$, and $\boldsymbol{X}_{\cdot n}$ be the n-th column vector of $\boldsymbol{X}$. Let $\left[\boldsymbol{X}\right]_+$ be the matrix, which is obtained by replacing negative values in $\boldsymbol{X}$ with zeros. Let $\boldsymbol{X}^{(g)} \in \mathcal{R}_+^{M \times |\mathcal{G}_g|}$ be the sub-matrix of data matrix $\boldsymbol{X}$ which correspond to instances in $\mathcal{G}_g$. Similarly the sub-matrix of $\boldsymbol{V}$ is $\boldsymbol{V}^{(g)} \in \mathcal{R}_+^{|\mathcal{G}_g| \times K}$ and the sub-vector of $\boldsymbol{V}_{\cdot k}$ is $\boldsymbol{V}_{\cdot k}^{(g)} \in \mathcal{R}_+^{|\mathcal{G}_g| \times 1}$. $\overline{\mathcal{G}_g}$ is the complement set of $\mathcal{G}_g$, and $\boldsymbol{V}^{(\overline{g})}$ is the sub-matrix corresponding to instances in the complement set $\overline{\mathcal{G}_g}$. $\boldsymbol{V}^{(1:G')}$ is the sub-matrix corresponding to instances in the union set $\cup_{g=1}^{G'} \mathcal{G}_g$. Let $\|\boldsymbol{X}\|_F = \sqrt{\sum_i \sum_j \boldsymbol{X}_{ij}^2}$ be the Frobenius norm of matrix $\boldsymbol{X}$. For $q > 0$, let $\|\boldsymbol{x}\|_q$ be the ℓ_q norm of vector $\boldsymbol{x}$, *i.e.* $||\boldsymbol{x}||_q = (\sum_n x_n^q)^{\frac{1}{q}}$. Let $\|\boldsymbol{V}\|_{1,q} = \sum_{k=1}^K \|\boldsymbol{V}_{\cdot k}\|_q$ be $\ell_{1/q}$ mixed norm of matrix $\boldsymbol{V}$, which is ℓ_1 norm of the vector that consists of ℓ_q norm of columns, $\boldsymbol{V}_{\cdot k},\ k = 1, \ldots, K$.

2.2 Problem Setting

The basic NMF model (1) minimizes the following cost, i.e. the sum of squared errors:

$$\min_{\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0} \frac{1}{2} \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T \right\|_F^2. \tag{2}$$

This optimization has been conducted by various methods, including matrix multiplication [14], an active set method [15], and a block coordinate descent (BCD) method [13].

Our objective is to incorporate auxiliary information on overlapping groups $\mathcal{G}$ in the NMF framework to detect essential low-rank structures from given data matrix $\boldsymbol{X}$ more accurately. We then add two regularization terms (to consider group information $\mathcal{G}$) to the squared error, resulting in the following optimization problem:

$$\min_{\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0} \frac{1}{2} \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T \right\|_F^2 + \alpha \|\boldsymbol{U}\|_F^2 + \beta \cdot \Omega^{\mathcal{G}}(\boldsymbol{V}), \tag{3}$$

1. Both groups and clusters indicate sets of instances, while we use "group" for give information and "cluster" for estimates from given data.

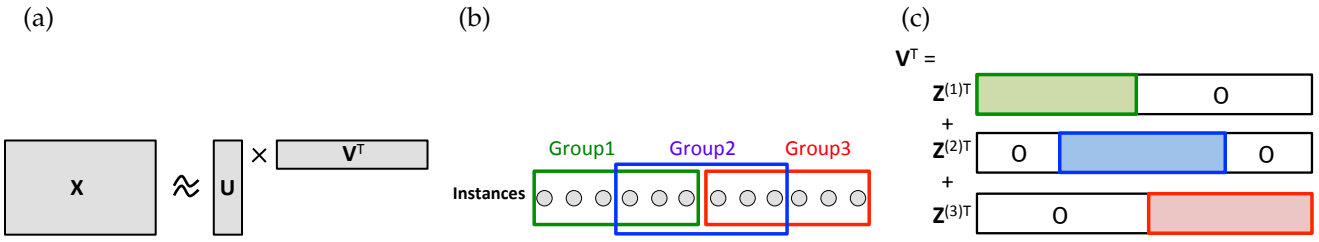

Fig. 1. Our NMF: (a) model, (b) auxiliary information on overlapping groups, (c) latent matrices for the overlapping groups in (b).

where α and β are weight parameters of the two regularization terms. The first regularization term is ℓ_2 regularization for $\boldsymbol{U}$, and the second term $\Omega^{\mathcal{G}}(\boldsymbol{V})$ is to incorporate auxiliary group information, which is our main result of this paper and will be described in the next section and later more. We need the regularization terms for both $\boldsymbol{U}$ and $\boldsymbol{V}$, because the regularization for one matrix only, say $\boldsymbol{U}$, still allows to keep the other matrix, say $\boldsymbol{V}$, flexible. The regularization on groups should cover all instances, while real applications have a semi-supervised setting, where group information is usually given to only part of all instances. Thus our setting has group g for $G' < g \leq G$, which has an unique instance.

If given auxiliary information is disjoint groups, this minimization problem is already solved by [16], which however cannot be applied to the case that given groups are overlapped. On the other hand, our problem setting considers overlapping groups, and in this sense, our problem is a generalization of [16]. We emphasize that this generalization is essential for real applications, because overlapping groups can be found in many applications, typical examples being documents labeled by multiple topics and genes labeled by multiple functions.

2.3 Regularization with Overlapping Groups

For a penalty term, we introduce a mixed norm to incorporate auxiliary group information. We briefly explain the effect of the mixed norm using a simple sample, where we have three variables a_1, a_2 and a_3, for which auxiliary groups are defined as $\mathcal{G}_1 = \{1, 2\}$ and $\mathcal{G}_2 = \{3\}$. In this setting, the ℓ_1-penalty used by Lasso is defined by $|a_1| + |a_2| + |a_3|$, in which three coordinate directions are independently considered as three linear terms, leading to the sparsity in individual variables. On the other hand, the ℓ_2-penalty defined by $\sqrt{a_1^2 + a_2^2 + a_3^2}$, in which all three directions are equally considered, by which the sparsity is not encouraged. The mixed norm of the group lasso, $\sqrt{a_1^2 + a_2^2} + |a_3|$, considers directions of a_1 and a_2 equally, meaning that two directions a_1 and a_2 are equal in group $\mathcal{G}_1$, but the coordinate directions of $\mathcal{G}_1$ and $\mathcal{G}_2$ are differently considered, because of the convexity of the norm. This indicates that the mixed norm encourages the sparsity at the group level.

Extending this idea, we can introduce a regularization term for non-overlapping groups, *i.e.* $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$, for $i \neq j$, which is a mixed norm, *i.e.* a $\ell_{1/q}$ norm [16], as follows:

$$\Omega_{1,q}^{\mathcal{G},\text{non-overlap}}(\boldsymbol{V}) = \sum_{g=1}^{G} \left\| \boldsymbol{V}^{(g)} \right\|_{1,q}. \tag{4}$$

The above norm is ℓ_1 norm of vector $\boldsymbol{V}_{\cdot k}^{(g)}$ ($k = 1, \ldots, K,\ g = 1, \ldots, G$) which is the sub-vector of $\boldsymbol{V}$ in both column k and group $\mathcal{G}_g$. This regularization over matrices is a generalization of group information in a regression model where parameters are vectors [17]. The regularization term in (4) induces group-wise sparsity in each column of optimized matrix $\boldsymbol{V}$, because of the ℓ_1 norm. Similarly, for each group g, the regularization term in (4) induces selectivity over columns of $\boldsymbol{V}^{(g)}$, allowing only a few columns to have non-zero values, because of the ℓ_1 norm over columns of $\boldsymbol{V}^{(g)}$. In fact, this regularization can be applied directly to the case with overlapping groups, *i.e.* $\mathcal{G}_i \cap \mathcal{G}_j \neq \emptyset$ for $i \neq j$. However if we apply this norm to overlapping groups, this will cause a problem that if one group has instances which are not selected by this norm, this group and its all instances may not be selected even if these instances are in other groups [12], [18]. More formally, using the result in [12], we can easily prove that the support of the optimal column vector $\boldsymbol{V}_{\cdot k}$ is

$$\text{supp}(\boldsymbol{V}_{\cdot k}) = \left(\bigcup_{\mathcal{G} \in \mathcal{S}_0} \mathcal{G} \right)^C,$$

where $\mathcal{S}_0$ is the set of groups such that $\left\| \boldsymbol{V}_{\cdot k}^{(g)} \right\| = 0$ and $\mathcal{S}^C$ means the complement of set $\mathcal{S}$. That is, this regularization is likely to induce excessive sparsity that will eliminate even groups, which are closely related with the essential low-rank structure.

To avoid this excessive sparsity, we introduce latent matrices, $\boldsymbol{Z} = \{\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \ldots, \boldsymbol{Z}^{(G)}\}$ ($\boldsymbol{Z}^{(g)} \in \mathrm{R}^{N \times K}$), by which $\boldsymbol{V}$ can be defined as follows:

$$\boldsymbol{V} = \sum_{g=1}^{G} \boldsymbol{Z}^{(g)}, \quad \text{s.t.} \quad \overline{\boldsymbol{Z}^{(g)}} = 0,\ g = 1, \ldots, G. \tag{5}$$

Fig. 1 (c) shows an illustrative example of the latent matrix for the overlapping groups in Fig. 1 (b). This

figure shows that each matrix corresponds to a group, and all elements of the complement set of a group are fixed to zeros, which do not have to be optimized. We note that, for non-overlapping groups, latent matrices $\boldsymbol{Z}$ are clearly equivalent to $\boldsymbol{V}$.

We further assume a weight for each group, $\sqrt{|\mathcal{G}_g|}$, to balance among the regularization terms for groups, because the number of instances can be different by groups [19]. This weight is important, since our problem is a semi-supervised setting where one group may have a very small number of instances, say only one instance, which is definitely much smaller than the size of given groups. We finally define our regularization term on overlapping groups as follows:

$$\Omega^{\mathcal{G}}_{1,q}(\boldsymbol{Z}) \;=\; \sum_{g=1}^{G} \sqrt{|\mathcal{G}_g|}\, \left\| \boldsymbol{Z}^{(g)} \right\|_{1,q}. \tag{6}$$

From the result in [12], the support of norm (6) is

$$\mathrm{supp}(\boldsymbol{V}_{\cdot k}) = \bigcup_{\mathcal{G}\in\mathcal{S}_1} \mathcal{G},$$

where $\mathcal{S}_1$ is the set of groups such that $\left\| \hat{\boldsymbol{V}}^{(g)}_{\cdot k} \right\| > 0$. This result shows that the excessive sparsity problem is solved by our regularization term in (6).

Our optimization problem with overlapping group information is thus given as follows:

$$\min_{\boldsymbol{U}\geq 0, \boldsymbol{V}\geq 0, \boldsymbol{V}=\sum_{g=1}^{G} \boldsymbol{Z}^{(g)}} J(\boldsymbol{U},\boldsymbol{V}), \tag{7}$$

where

$$J(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{2}\left\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}^T\right\|_F^2 + \alpha\left\|\boldsymbol{U}\right\|_F^2 + \beta\Omega^{\mathcal{G}}_{1,q}(\boldsymbol{Z}). \tag{8}$$

While (8) is not convex for both $\boldsymbol{U}$ and $\boldsymbol{V}$, optimizing $\boldsymbol{V}$ with fixed $\boldsymbol{U}$ and optimizing $\boldsymbol{U}$ with fixed $\boldsymbol{V}$ are both convex. We present an efficient optimization algorithm which updates $\boldsymbol{U}$ and $\boldsymbol{V}$ alternately, under $q = 2$. We note that we can further develop an optimization algorithm for the general case of $q > 1$. Hereafter we call our NMF formulation (shown in (7)) *MFOG*, which stands for non-negative Matrix Factorization with Overlapping Groups.

3 Optimization Algorithm

3.1 Optimization for U

We first present optimization for $\boldsymbol{U}$, where auxiliary information is not considered. In this case, (8) can be easily transformed to optimize $\boldsymbol{U}$ as follows:

$$J_U(\boldsymbol{U};\alpha) = \frac{1}{2}\left\|\widehat{\boldsymbol{X}} - \boldsymbol{U}\widehat{\boldsymbol{V}}\right\|_2^2, \tag{9}$$

where

$$\widehat{\boldsymbol{V}} = \left(\boldsymbol{V}^T,\ \sqrt{2\alpha}\boldsymbol{I}_K\right), \qquad \widehat{\boldsymbol{X}} = \left(\boldsymbol{X},\ \boldsymbol{0}_{M\times K}\right).$$

(9) is a convex function for $\boldsymbol{U}$. The minimization for (9) can be solved by computing the solution of the following equation:

$$\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\boldsymbol{U}^T = \widehat{\boldsymbol{V}}\widehat{\boldsymbol{X}}^T, \tag{10}$$

and replacing negative values in the solution $\boldsymbol{U}^*$ with zeros [20], as follows:

$$\boldsymbol{U} \leftarrow \left[\boldsymbol{U}^*\right]_+. \tag{11}$$

The above algorithm, which is alternative least squares (ALS), is less sensitive for poor initialization than usual multiplicative algorithms [14], meaning that ALS avoids a path to a poor local minima [20].

3.2 Optimization for V (or Z)

This section presents optimization algorithms for $\boldsymbol{V}$, which is equal to the sum of $\boldsymbol{Z}^{(1)},\ldots,\boldsymbol{Z}^{(G)}$. This optimization is more difficult than that for $\boldsymbol{U}$, because of the mixed norm with overlapping groups. The optimization for a regression model with overlapping groups leads to two different ideas: 1) the first one is latent matrices. The approach by latent matrices is an extension of [12], where variables are duplicated, because instances can belong to multiple groups. This results in a slight increase in the space complexity. 2) the other idea is a dual problem approach, which directly optimizes $\boldsymbol{V}$ via the dual problem of the original optimization [21], skipping the optimization of latent matrices $\boldsymbol{Z}$, and so hereafter we call this approach a *direct* method, since we compute $\boldsymbol{V}$ directly. This method has to solve a semi-definite problem (SDP), which requires much more computational cost than the approach by latent matrices.

In the updating rule of the both optimization methods, the block unit can be two types: 1) a vector or 2) the entire matrix. We call the case of a vector *Vector-BCD* (or *Vec*), standing for Vector based Block Coordinate Descent, while we call the entire matrix case *Matrix-BCD* (or *Mat*), standing for Matrix based Block Coordinate Descent.

We combine these two types of blocks with two different ideas of optimization, finally resulting in four different algorithms. Fig. 2 summarizes these four combinations with related approaches, where four different algorithms are named as *Lat-Vec*, *Lat-Mat*, *Dir-Vec* and *Dir-Mat*, where *Lat* and *Dir* stand for latent matrices and a direct method based on a dual problem, respectively. We describe each of the four algorithms in the following sections.

3.2.1 Lat-Vec:Optimization of Z by vector-BCD

We have theorems for updating rules as follows:

Theorem 1. *The optimal value of* $\boldsymbol{Z}^{(g)}_{\cdot k},\ g = 1,\ldots,G'$ *under fixed* $\boldsymbol{U}$ *and* $\boldsymbol{Z}^{(c)}_{\cdot j}\ (j\neq k,\ c\neq g)$ *is*

$$\boldsymbol{Z}^{(g)}_{\cdot k} \leftarrow \left[1 - \frac{\lambda^{(g)}_{LV}}{\left\|\boldsymbol{s}^{(g,k)}_{LV}\right\|_2}\right]_+ \cdot \boldsymbol{s}^{(g,k)}_{LV}, \tag{12}$$

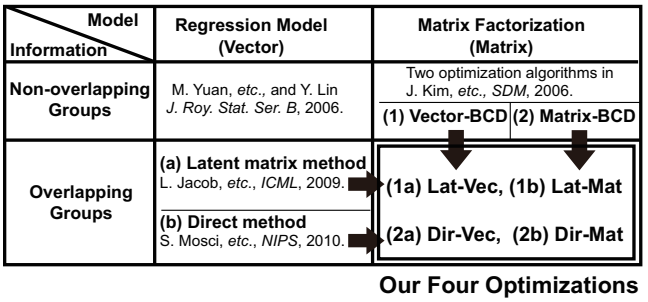

Fig. 2. Related optimization problem with auxiliary group information.

Input : data matrix $\boldsymbol{X}$, groups $\mathcal{G}$ and rank K
Output : low rank matrices $\boldsymbol{U}$ and $\boldsymbol{V}$

1: Initialize $\boldsymbol{U}$ and $\boldsymbol{Z}$.
2: **repeat**
3: Update $\boldsymbol{U}$ by Eq. (11)
4: **for** $k = 1 \ldots, K$ **do**
5: **for** $g = 1 \ldots, G'$ **do**
6: Update $\boldsymbol{Z}_{\cdot k}^{(g)}$ by Eq. (12)
7: **end for**
8: $\boldsymbol{V}_{\cdot k}^{(1:G')} \leftarrow \sum_{g=1}^{G'} \boldsymbol{Z}_{\cdot k}^{(g)}$
9: Update $\boldsymbol{V}_{\cdot k}^{(0)}$ by Eq. (14)
10: **end for**
11: **until** convergence

Fig. 3. Optimization of MFOG by Lat-Vec

where $\lambda_{LV}^{(g)} = \frac{\beta\sqrt{|\mathcal{G}_g|}}{\|\boldsymbol{U}_{\cdot k}\|_2^2}$ *and*

$$\boldsymbol{s}_{LV}^{(g,k)} = \left[\frac{\left(\boldsymbol{X}^{(g)T} - \boldsymbol{V}^{(g)}\boldsymbol{U}^T + \boldsymbol{Z}_{\cdot k}^{(g)}\boldsymbol{U}_{\cdot k}^T\right)\boldsymbol{U}_{\cdot k}}{\|\boldsymbol{U}_{\cdot k}\|_2^2}\right]_+ \quad (13)$$

Theorem 2. *For the joint group* $\mathcal{G}_0 = \cup_{g=G'+1}^{G}\mathcal{G}_g$, *the optimal* $\boldsymbol{Z}_{\cdot k}^{(0)}$ *is equal to* $\boldsymbol{V}_{\cdot k}^{(0)}$, *which is given by*

$$\boldsymbol{V}_{\cdot k}^{(0)} \leftarrow \left[1 - \frac{\lambda_{LV}^{(0)}}{\left\|\boldsymbol{s}_{LV}^{(0,k)}\right\|_2}\right]_+ \cdot \boldsymbol{s}_{LV}^{(0,k)}. \quad (14)$$

where $\lambda_{LV}^{(0)} = \frac{\beta}{\|\boldsymbol{U}_{\cdot k}\|_2^2}$.

Proofs of these theorems are described in Appendix. These theorems give us an optimization algorithm, Lat-Vec. Fig. 3 shows a pseudocode of Lat-Vec.

3.2.2 Lat-Mat:Optimization of Z by Matrix-BCD

Theorem 3. *Under fixed* $\boldsymbol{U}$ *and* $\boldsymbol{Z}^{(c)},\ c \neq g$, *the update rule for* $\boldsymbol{Z}^{(g)},\ g = 1, \ldots, G'$

$$\boldsymbol{Z}^{(g)} \leftarrow \left(1 - \frac{\lambda_{LM}^{(g)}}{\left\|\boldsymbol{S}_{LM}^{(g)}\right\|_2}\right)_+ \boldsymbol{S}_{LM}^{(g)}, \quad (15)$$

can reduce cost $J(\boldsymbol{U}, \boldsymbol{V})$ *defined in (8), where*

$$\boldsymbol{S}_{LM}^{(g)} = \left[\boldsymbol{Z}^{(g)} + \frac{1}{L_L}\left(\boldsymbol{X}^{(g)T} - \boldsymbol{V}^{(g)}\boldsymbol{U}^T\right)\boldsymbol{U}\right]_+, \quad (16)$$

Input : data matrix $\boldsymbol{X}$, groups $\mathcal{G}$ and rank K
Output : low rank matrix $\boldsymbol{U}$ and $\boldsymbol{V}$

1: Initialize $\boldsymbol{U}$ and $\boldsymbol{Z}$.
2: **repeat**
3: Update $\boldsymbol{U}$ by Eq. (11)
4: **for** $g = 1 \ldots, G'$ **do**
5: Update $\boldsymbol{Z}^{(g)}$ by Eq. (15)
6: **end for**
7: $\boldsymbol{V} \leftarrow \sum_{g=1}^{G} \boldsymbol{Z}^{(g)}$
8: Update $\boldsymbol{V}^{(0)}$ by Eq. (17)
9: **until** convergence

Fig. 4. Optimization of MFOG by Lat-Mat

$\lambda_{LM}^{(g)} = \frac{\beta\sqrt{|\mathcal{G}_g|}}{L_L}$ *and* L_L *is the Lipchitz constant which is obtained by multiplying* K *by the maximum eigen values of* $\boldsymbol{U}^T\boldsymbol{U}$.

Because of equation $\boldsymbol{V}^{(g)} = \boldsymbol{Z}^{(g)},\ g = G'+1, \ldots, G$, we can update $\boldsymbol{V}^{(0)}$ directly as follows:

Theorem 4. *Under fixed* $\boldsymbol{U}$ *and* $\boldsymbol{Z}^{(c)},\ c \neq g$, *the update rule for* $\boldsymbol{Z}^{(g)},\ g = G' + 1, \ldots, G$

$$\boldsymbol{V}^{(0)} \leftarrow \left(1 - \frac{\lambda_{LM}^{(0)}}{\left\|\boldsymbol{S}_{LM}^{(0)}\right\|_2}\right)_+ \boldsymbol{S}_{LM}^{(0)}. \quad (17)$$

can reduce cost $J(\boldsymbol{U}, \boldsymbol{V})$ *defined in (8), where* $\lambda_{LM}^{(0)} = \frac{\beta}{L_L}$.

The proofs of these theorems are given in Appendix, and these theorems give us algorithm Lat-Mat. Fig. 4 shows a pseudocode of Lat-Mat. We note that the iterative update of $\boldsymbol{Z}$ under fixed $\boldsymbol{U}$ (i.e. lines 4-5 of Fig. 4) can be accelerated by minor modification of using an idea called FISTA [22].

3.2.3 Dir-Vec:Direct optimization of V by vector-BCD

Theorem 5. *The updating rule of Dir-Vec can be formalized as follows:*

$$\boldsymbol{V}_{\cdot k} \leftarrow \boldsymbol{s}_{DV}^{(k)} - Proj_{\mathcal{K}_{\mu_k}^{\widehat{\mathcal{G}}}}\left(\boldsymbol{s}_{DV}^{(k)}\right), \quad (18)$$

where

$$\begin{aligned}
\boldsymbol{s}_{DV}^{(k)} &= \left[\frac{\left(\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T + \boldsymbol{U}_{\cdot k}\boldsymbol{V}_{\cdot k}^T\right)^T\boldsymbol{U}_{\cdot k}}{\|\boldsymbol{U}_{\cdot k}\|_2^2}\right]_+, \\
Proj_{\mathcal{K}_{\mu_k}^{\widehat{\mathcal{G}}}}\left(\boldsymbol{s}_{DV}^{(k)}\right) &= \arg\min_{\boldsymbol{a} \in \mathcal{K}_k^{\widehat{\mathcal{G}}}} \left\|\boldsymbol{a} - \boldsymbol{s}_{DV}^{(k)}\right\|_F^2, \\
\mathcal{K}_k^{\widehat{\mathcal{G}}} &= \left\{\boldsymbol{s};\ \left\|\left(\boldsymbol{s}_{DV}^{(k)}\right)^{(g)}\right\|_2 \leq \lambda_{DV}^{(k,g)},\ g \in \widehat{\mathcal{G}}\right\}, \\
\lambda_{DV}^{(k,g)} &= \frac{\beta\sqrt{|\mathcal{G}_g|}}{\|\boldsymbol{U}_{\cdot k}\|_2^2}, \\
\widehat{\mathcal{G}} &= \left\{\mathcal{G}_g;\ \left\|\left(\boldsymbol{s}_{DV}^{(k)}\right)^{(g)}\right\|_2 > \lambda_{DV}^{(k,g)}\right\} \subseteq \mathcal{G}.
\end{aligned}$$

Input : data matrix $\boldsymbol{X}$, groups $\mathcal{G}$ and rank K
Output : low rank matrix $\boldsymbol{U}$ and $\boldsymbol{V}$

1: Initialize $\boldsymbol{U}$ and $\boldsymbol{V}$.
2: **repeat**
3: Update $\boldsymbol{U}$ by Eq. (11)
4: **for** $k = 1 \ldots, K$ **do**
5: Update $\boldsymbol{V}_{\cdot k}$ by Eq. (18)
6: **end for**
7: **until** convergence

Fig. 5. Optimization for MFOG by Dir-Vec

The proof of this theorem is shown in Appendix. In (18), the projection operator cannot be solved analytically. Thus in order to obtain the projection, we use a projected newton method on the dual problem, which turns into a constrained semi-definite problem [21]. Fig. 5 shows a pseudocode of Dir-Vec.

3.2.4 Dir-Mat:Direct optimization of $\boldsymbol{V}$ by matrix-BCD

Theorem 6. *The updating rule of Dir-Mat can be formalized as follows:*

$$\boldsymbol{V} \leftarrow \boldsymbol{S}_{DM} - Proj_{\mathcal{H}^G}(\boldsymbol{S}_{DM}), \tag{19}$$

where

$$\begin{aligned}
\boldsymbol{S}_{DM} &= \left[\boldsymbol{V} + \frac{1}{L_D}\left(\boldsymbol{X}^T - \boldsymbol{V}\boldsymbol{U}^T\right)\boldsymbol{U}\right]_+, \\
Proj_{\mathcal{H}^G}(\boldsymbol{S}_{DM}) &= \arg\min_{\boldsymbol{S}\in\mathcal{H}^G} \left\|\boldsymbol{S} - \boldsymbol{S}_{DM}\right\|_2^2, \\
\mathcal{H}^G &= \Big\{\boldsymbol{S};\ \left\|\boldsymbol{S}_{\cdot k}^{(g)}\right\|_2 \le \lambda_{DM}^{(g)}, \\
&\qquad g = 1, \ldots, G,\ k = 1, \ldots, K\Big\}, \\
\lambda_{DM}^{(g)} &= \frac{\beta\sqrt{|\mathcal{G}_g|}}{L_D},
\end{aligned}$$

and L_D is the Lipchitz constant which is obtained by the maximum eigen values of $\boldsymbol{U}^T\boldsymbol{U}$.

The proof of this theorem is shown in Appendix. As shown in Dir-Vec, we cannot solve the projection analytically, because of the mixed norm of overlapping groups. However, note that the mixed norm is computed by the sum of the norm over columns of $\boldsymbol{V}$, meaning that the convex set does not overlap with any column, resulting in that the projection for each column of $\boldsymbol{V}$ can be computed individually. Finally we can solve the projection by using the solver used in Dir-Vec. Fig. 6 shows a pseudocode of Dir-Mat.

4 Related Work

The most basic model of NMF, i.e. (1), which we hereafter call *MFBS*, has been solved by many different algorithms. The most classical algorithm is matrix multiplication [14], while the performance of this

Input : data matrix $\boldsymbol{X}$, groups $\mathcal{G}$ and rank K
Output : low rank matrix $\boldsymbol{U}$ and $\boldsymbol{V}$

1: Initialize $\boldsymbol{U}$ and $\boldsymbol{V}$.
2: **repeat**
3: Update $\boldsymbol{U}$ by Eq. (11)
4: Update $\boldsymbol{V}$ by Eq. (19)
5: **until** convergence

Fig. 6. Optimization for MFOG by Dir-Mat

method heavily depends on initial values. This disadvantage has been improved by many later methods, such as alternating least squares (ALS) [20], a block coordinate descent method (BCD) [13] and an active set method [15]. See the detail of these algorithms in review articles on NMF, such as [23].

MFBS has been extended in many different directions. One way was to incorporate auxiliary information, such as semi-supervised NMF under must- and cannot-link constraints [6], for which one solution is to learn matrices from constraints before low-rank matrices are optimized [24]. Auxiliary information with must-link and cannot-link can be simply a graph, in which nodes are instances. Graph Laplacian has thus been used to impose that nodes connected by edges should have the same label or similar values. There exist at least two different types of NMF which incorporate auxiliary graph information by using graph Laplacian which we call 1) *MFGL*, which stands for non-negative Matrix Factorization with Graph Laplacian [25], and 2) *MFGC*, which stands for non-negative Matrix Factorization with Graph Laplacian for Clustering [26]. In both of them, the regularization term has graph Laplacian, i.e. $\mathrm{Tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V})$, where $\mathrm{Tr}(\cdot)$ is matrix trace and $\boldsymbol{L}$ is graph Laplacian. For example, the objective function of MFGL is as follows:

$$\min_{\boldsymbol{U}\ge 0, \boldsymbol{V}\ge 0} \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T\right\|_F^2 + \beta\ \mathrm{Tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}), \tag{20}$$

The optimization can be performed based on matrix manipulation similar to that for the basic model [14]. MFGC is two-way clustering, where low rank matrix $\boldsymbol{V}$ is restricted to a binary cluster indicator matrix [26]. When we consider the setting of one-way clustering, the optimization problem can be given as follows:

$$\min_{\boldsymbol{U}\ge 0, \boldsymbol{V}\in\mathcal{CI}} \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T\right\|_F^2 + \beta\ \mathrm{Tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}), \tag{21}$$

where $\mathcal{CI}$ is the entire set of binary cluster indicator matrices. The optimization can be performed by a SVD-based fast algorithm, using the nature of binary cluster indicator matrices [26].

The difference between (20) (or (21)) and (8) is the regularization term for $\boldsymbol{V}$. The regularization term in MFGL and MFGC makes $\boldsymbol{V}_{i\cdot}$ and $\boldsymbol{V}_{j\cdot}$ of any instance pair (i, j) *close (or same) values* if (i, j) are connected

in the given graph, while the objective function in (8) of MFOG makes $\boldsymbol{V}_{i\cdot}$ and $\boldsymbol{V}_{j\cdot}$ of any instance pair (i, j) *a similar sparse pattern*, if the pair is in an auxiliary group. As a regularization term, graph Laplacian is more strict, because two values must be the same when two instances in the same group. This makes MFGL and MFGC perform better than MFOG if given auxiliary groups are totally correct and not overlapped. However, MFOG would outperform MFGL and MFGC, if auxiliary groups are noisy and overlapped, which is we note often the case with real-world applications. In fact the advantage of MFOG was confirmed by our experiments in the later experimental section.

Groups are also typical auxiliary information, and NMF for disjoint group information is already proposed [16], but the approach in [16] cannot be applied to the case that given groups are overlapped. Similarly group sparse coding [27] is also low-rank approximation with non-negative parameters under disjoint groups, while the problem setting of [27] is equivalent to [16]. We emphasize that the disjoint group setting is practically rare, and our setting of overlapping groups is more realistic. Our approach of using latent matrices and the mixed norm can handle overlapping groups appropriately and can be a generalization of [16]. Possible similar work on overlapping groups is structured principal component analysis [28], which however does not assume non-negatively in low-rank factorized matrices, resulting in a totally different optimization manner from ours.

5 EXPERIMENTS

We first compared our four optimization algorithms: Lat-Vec, Lat-Mat, Dir-Vec, and Dir-Mat, each other, by using synthetic datasets, and selected the best performance method in this experiment as our proposed method, MFOG. We then, by using both synthetic and real datasets, compared the clustering performance of MFOG with other methods, including three NMF-based methods, MFBS, MFGL[2] and MFGC and k-means (KM), under the setting that group information is given. Here MFGC and MFGL are rather competing methods, while MFBS and KM are baseline methods.

For all NMF-based methods, the cluster assignment was performed by $c_n = \arg\max_k V_{n,k}$ $(n = 1, \ldots, N)$, meaning single (or hard) cluster assignment. We used this setting, because of no standard methods for assigning multiple clusters to each instance from the output of NMF.

The performance was measured by normalized mutual information (NMI) [29], [3]. NMI between a set of predicted clusters $\mathcal{C}_P$ and a set of true clusters $\mathcal{C}_T$ is calculated as follows:

$$\mathrm{NMI} = \frac{\mathrm{MI}(\mathcal{C}_P, \mathcal{C}_T)}{\max\big(\mathrm{H}(\mathcal{C}_P), \mathrm{H}(\mathcal{C}_T)\big)},$$

where $\mathrm{MI}(\mathcal{C}_P, \mathcal{C}_T) = H(\mathcal{C}_P) + H(\mathcal{C}_T) - H(\mathcal{C}_P, \mathcal{C}_T)$, $\mathrm{H}(\mathcal{C}) = -P(\mathcal{C}) \log P(\mathcal{C})$, $\mathrm{H}(\mathcal{C}_P, \mathcal{C}_T) = -P(\mathcal{C}_P, \mathcal{C}_T) \log P(\mathcal{C}_P, \mathcal{C}_T)$, and $P(\mathcal{C})$ is the empirical distribution of cluster assignment $\mathcal{C}$.

All experiments were performed by using MacPro Early 2009 (Intel Xeon Quad-Core 2.66GHz, Memory 16GB) and Matlab 2013a. Throughout the experiments, weight α for $\boldsymbol{U}$ of MFOG was fixed at 0.01. Optimization was terminated when either of the following two conditions was satisfied: 1) $\delta = ||\boldsymbol{V}^{(t+1)} - \boldsymbol{V}^{(t)}||_F^2$, which is the squared difference between the two matrices obtained by two consecutive iterations, was less than or equal to ϵ, and 2) the number of iterations reached Itr_{max}. The values of ϵ and Itr_{max} were set to 10^{-8} and 1000, respectively, if any specific value is not shown.

5.1 Synthetic Datasets

We generated synthetic datasets by the following manner: We first set the size of data matrix $\boldsymbol{X}$ to be $M \times N$, and the rank of true low rank matrices $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ be K. We then generated true low rank matrices $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ as follows: $U^*_{m,k} = c_u$ for $\frac{(k-1)M}{K} + 1 \leq m \leq \frac{kM}{K}$ $(m = 1, \ldots, M,\ k = 1, \ldots, K)$; otherwise zero, where $c_u = \sqrt{K/M}$ is a constant value, which was set to normalize each column of $\boldsymbol{U}^*$. Similarly $V^*_{n,k} = c_v$ for $\frac{(k-1)N}{K} + 1 \leq n \leq \frac{kN}{K}$ $(n = 1, \ldots, N,\ k = 1, \ldots, K)$; otherwise zero, where $c_v = \sqrt{K/N}$ is also a constant value for normalizing each column of $\boldsymbol{V}^*$. We further assigned true cluster labels by $c_n = k$ for $\frac{(k-1)N}{K} + 1 \leq n \leq \frac{kN}{K}$ $(n = 1, \ldots, N)$, by using true low rank matrix $\boldsymbol{V}^*$. Finally we generated data matrix $\boldsymbol{X}$ by $\big[\boldsymbol{U}^*\boldsymbol{V}^{*T} + \boldsymbol{E}\big]_+$, where noise matrix $\boldsymbol{E} \in \mathcal{R}^{M \times N}$ was generated from Gaussian distribution $\mathcal{N}(0, \sigma^2)$. We generated auxiliary group information $\mathcal{G}_1 = \{1, \ldots, C + L\}$, ..., $\mathcal{G}_g = \{(g-1) \cdot C - L + 1, \ldots, g \cdot C + L\}$, ..., $\mathcal{G}_G = \{G \cdot C - L + 1, \ldots, N\}$, where C is the number of instances in an auxiliary group, G $(= \frac{N}{C})$ is the number of auxiliary groups and L is a parameter to adjust the overlap between given groups. In our experiments, we generated a semi-supervised learning setting, for which, we gave auxiliary group labels to only N_s instances, which were chosen randomly.

5.1.1 Comparing Optimization Algorithms for MFOG

We compared our four proposed optimization algorithms: Lat-Vec, Lat-Mat, Dir-Vec and Dir-Mat. The default parameters for synthetic datasets were set as follows: $N = 500$, $M = 50$, $K = 5$, $\sigma^2 = 10^{-2}$, $N_s = 500$, $L = 25$, $C = 100$ and $\beta = 0.01$. So these values were taken if any specific values are not shown.

2. Group information was transformed into a graph for MFGL and MFGC by connecting an edge between two nodes (instances) if these two instances are in the same group.

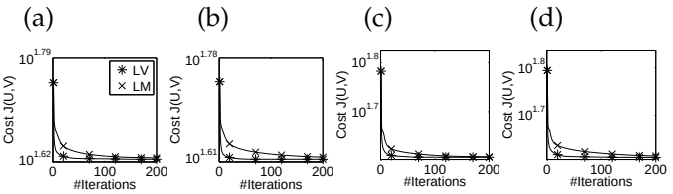

Fig. 7. Optimization convergence by Lat-Vec (LV) and Lat-Mat (LM) for (a) $L = 25,\ \beta = 10^{-2}$, (b) $L = 25,\ \beta = 10^{-3}$, (c) $L = 50,\ \beta = 10^{-2}$ and (d) $L = 50,\ \beta = 10^{-3}$.

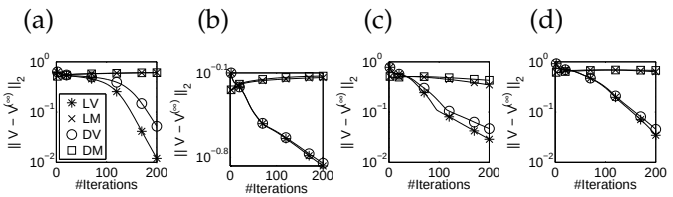

Fig. 8. Difference between $\boldsymbol{V}^{(t)}$ and $\boldsymbol{V}^{(\infty)}$ by Lat-Vec (LV), Lat-Mat (LM), Dir-Vec (DV) and Dir-Mat (DM) for (a) $L = 25,\ \beta = 10^{-2}$, (b) $L = 25,\ \beta = 10^{-3}$, (c) $L = 50,\ \beta = 10^{-2}$ and (d) $L = 50,\ \beta = 10^{-3}$

Comparison of our four optimization methods was conducted by varying L, G and β, while all other parameters were fixed.

First, we compared the convergence of cost $J(\boldsymbol{U}, \boldsymbol{V})$ in (8) under four typical settings. Fig. 7 shows the learning curves, i.e. $J(\boldsymbol{U}, \boldsymbol{V})$ at each iteration, of Lat-Vec (LV) and Lat-Mat (LM), showing that LV always achieved lower errors under all settings.

Second, we computed $||\boldsymbol{V}^{(t)} - \boldsymbol{V}^{(\infty)}||_2$, i.e. the ℓ_2 norm of the difference between $\boldsymbol{V}^{(t)}$ and $\boldsymbol{V}^{(\infty)}$ (matrix obtained at a stationary point), where $\boldsymbol{V}^{(t)}$ and $\boldsymbol{V}^{(\infty)}$ were $\boldsymbol{V}$ after the t-th and 5000-th iterations. Fig. 8 shows the result of $||\boldsymbol{V}^{(t)} - \boldsymbol{V}^{(\infty)}||_2$. Similar to Fig. 7, Fig. 8 shows that the error obtained by vector blocks was clearly different from that by matrix blocks. For all four parameter settings, significantly smaller differences were obtained by Lat-Vec or Dir-Vec, i.e. vector blocks. So from Figs. 7 and 8, we can conclude that Lat-Vec achieved the best performance.

Third, we evaluated the average computational time and the average number of iterations until convergence ($\epsilon \leq 10^{-8}$ or $\text{Itr}_{max} = 5,000$) over 20 runs, by changing L, G and β. Tables 1 and 2 show the summary over this experiment. These tables show that, in all cases, the computational time of latent matrix-based approaches was much lower than that of direct approaches, while the number of iterations was comparable. The slowness of the direct approach is due to the projection operators, which require sequential optimization. Thus from a real computational complexity viewpoint, latent matrices-based methods were better than the direct methods. Another finding was that the number of iterations of vector blocks could be smaller than that of matrix blocks, probably because an exact analytic solution can be obtained at each iteration of the vector-based approaches. Overall we decided to select Lat-Vec as our proposed optimization algorithm for MFOG.

TABLE 1
Real computational time.

(A)

L	Lat-Vec	Lat-Mat	Dir-Vec	Dir-Mat
0	1.4 ± 0.8	6.4 ± 2.6	61.0 ± 39.1	256.8 ± 106.0
25	1.6 ± 1.2	4.1 ± 2.1	75.6 ± 66.5	279.5 ± 118.3
50	1.5 ± 0.9	4.6 ± 2.2	64.7 ± 44.1	310.8 ± 114.0

(B)

G	Lat-Vec	Lat-Mat	Dir-Vec	Dir-Mat
5	1.4 ± 0.8	5.2 ± 2.5	57.3 ± 35.7	245.1 ± 64.9
10	2.6 ± 1.5	6.4 ± 3.0	112.7 ± 71.8	449.4 ± 189.6
20	4.1 ± 2.0	7.8 ± 4.4	186.6 ± 128.8	895.3 ± 371.3

(C)

β	Lat-Vec	Lat-Mat	Dir-Vec	Dir-Mat
10^{-1}	1.1 ± 0.8	3.6 ± 1.5	11.5 ± 9.0	153.5 ± 78.0
10^{-2}	1.5 ± 0.8	4.4 ± 2.2	31.7 ± 19.4	415.6 ± 172.0
10^{-3}	1.2 ± 0.5	4.3 ± 1.9	51.9 ± 26.6	303.9 ± 104.3

TABLE 2
The number of updates until convergence.

(A)

L	Lat-Vec	Lat-Mat	Dir-Vec	Dir-Mat
0	473 ± 271	3389 ± 1358	473 ± 271	3389 ± 1358
25	495 ± 384	2165 ± 1134	577 ± 500	2665 ± 980
50	430 ± 249	2391 ± 1179	469 ± 310	2663 ± 959

(B)

G	Lat-Vec	Lat-Mat	Dir-Vec	Dir-Mat
5	419 ± 252	2752 ± 1309	433 ± 272	2387 ± 697
10	523 ± 315	2245 ± 1039	563 ± 380	3103 ± 1308
20	501 ± 244	1657 ± 932	602 ± 383	2605 ± 1002

(C)

β	Lat-Vec	Lat-Mat	Dir-Vec	Dir-Mat
10^{-1}	334 ± 234	1906 ± 781	344 ± 266	2045 ± 994
10^{-2}	472 ± 252	2341 ± 1162	517 ± 304	2618 ± 1058
10^{-3}	382 ± 170	2270 ± 1013	401 ± 208	2974 ± 1059

5.1.2 Comparing Performance of MFOG with Competing Methods

We first checked the values of cluster assignment matrix $\boldsymbol{V}$ of MFOG, comparing with competing methods, using a certain synthetic dataset ($N = 100$, $M = 20$, $K = 4$, $L = 10$, $\sigma^2, = 0.01$ and $N_s = 75$). Fig. 9 shows (a) given groups (auxiliary information on groups), (b) true low-rank matrices $\boldsymbol{V}^*$, and (c-e) matrices $\boldsymbol{V}$, which are optimized by MFOG, MFGL and MFBS. We note that given information on groups cover a larger number of instances in $\boldsymbol{V}^*$, meaning that input groups are overlapped with each other. We further note that $\boldsymbol{V}$ of MFGC is a binary matrix, by which MFGC could not be compared in this experiment. The shown results of MFOG, MFGL and MFBS are the best cases in terms of NMI. This figure indicates that $\boldsymbol{V}$ obtained by MFOG has the largest number of zero and the smallest number of non-zero, where the

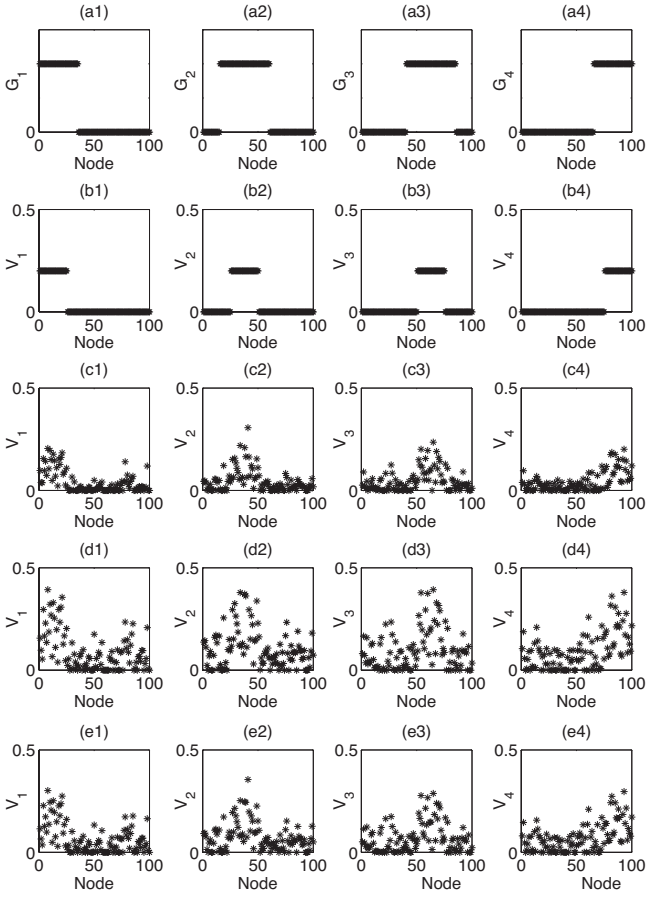

Fig. 9. (a1) – (a4) show given information on groups, which are overlapped with each other, (b1) – (b4) show the columns of true matrices $\boldsymbol{V}^*$, (c1) – (c4), (d1) – (d4) and (e1)–(e4) show the matrices optimized by MFOG, MFGL and MFBS, respectively.

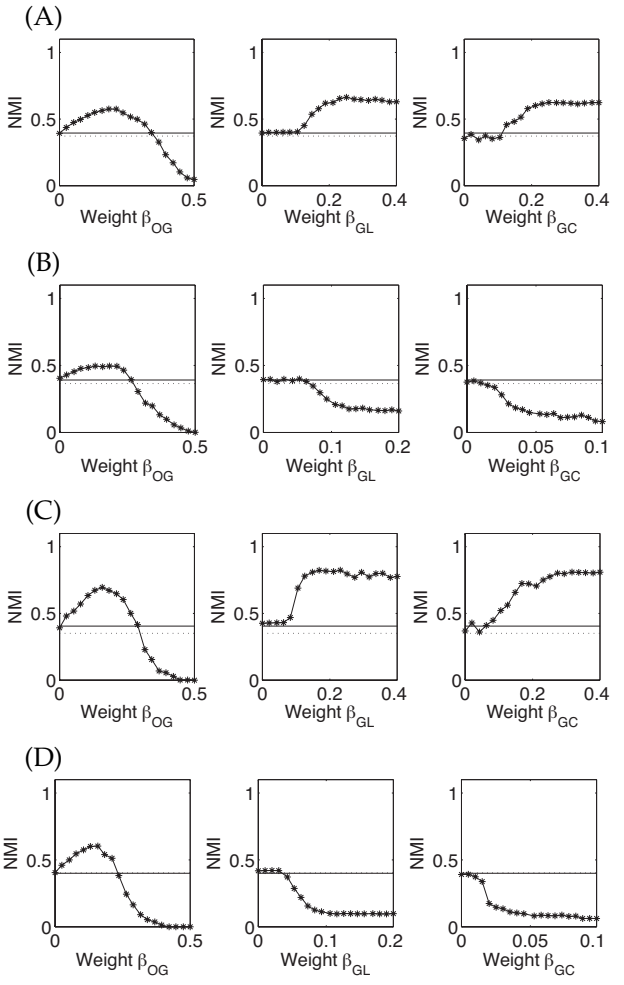

Fig. 10. Performance results on synthetic datasets under $\sigma^2 = 0.01$, and (A) $L = 0$, $N_s = 50$, (B) $L = 10$, $N_s = 50$, (C) $L = 0$, $N_s = 75$, (D) $L = 10$, $N_s = 75$. left: MFOG, middle: MFGL, right: MFGC, solid line: MFBS, and dotted line: KM

points with non-zero are well consistent with the true cluster structure. On the other hand, the values of the corresponding part of MFGL and MFBS were likely to be larger than zero, which blurs cluster estimation.

We then used four different types of datasets, obtained by changing L and N_s while keeping $\sigma^2 = 0.08$ to check NMI values. Fig. 10 shows NMI values of five competing methods, changing regularization parameter β. We note that the performance can be changed by varying the regularization parameter, and so the highest NMI in the range of all values of β should be checked. The first finding was that for all datasets, MFOG, MFGL, and MFGC outperformed MFBS and KM in the best NMI. Secondly, for disjoint groups ($L = 0$: A and C), MFOG, MFGL and MFGC achieved almost the same performance in terms of the highest value, while for overlapping groups ($L = 10$: B and D), MFOG clearly outperformed MFGL and MFGC.

Figs. 11 and 12 show that the norm and variance, respectively, of one block (corresponding to a given auxiliary group) of $\boldsymbol{V}$, when β was changed, under two typical experimental settings. From these figures, for MFOG and larger β, the variance decreased to almost zero, and the norm was all reduced to zero except a few cases. This result indicates that group-wise sparse patterns in optimized $\boldsymbol{V}$ are generated by the regularization of MFOG. On the other hand, for MFGL and MFGC, the norm was always positive even if the variance was reduced to zero for larger β. This result indicates that, for larger β, elements in $\boldsymbol{V}$ were almost the same non-zero values in each block, implying no group-wise sparsity.

These difference on the performance and optimized $\boldsymbol{V}$ are caused by the difference in regularization terms. MFOG uses a group norm, which regularizes $\boldsymbol{V}$ rather loosely, keeping group-wise sparsity (elements in optimized $\boldsymbol{V}$ can be either zero or non-zero values). On the other hand, MFGL and MFGC provide the same non-zero values to the instances within a group in optimized $\boldsymbol{V}$. Then, if one instance is in multiple groups, the element value (in $\boldsymbol{V}$) of this instance is like an average value over the groups containing this instance, finally the optimized $\boldsymbol{V}$ being likely to be

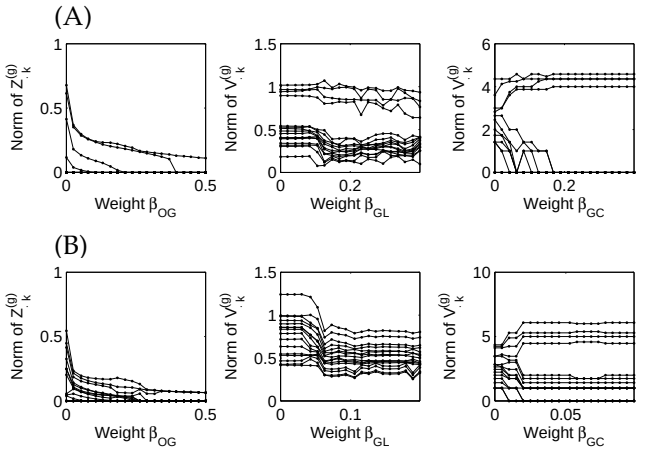

Fig. 11. The norm of each block of $\boldsymbol{V}$ corresponding to given groups under $\sigma^2 = 0.01$, and (A) $L = 0,\ N_s = 75$, (B) $L = 10,\ N_s = 75$. left: MFOG, middle: MFGL, right: MFGC

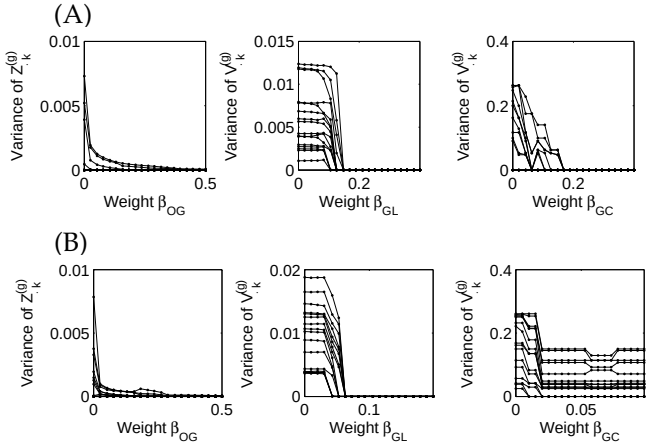

Fig. 12. The variance of each block of $\boldsymbol{V}$ corresponding to given groups under $\sigma^2 = 0.01$, and (A) $L = 0,\ N_s = 75$, (B) $L = 10,\ N_s = 75$. left: MFOG, middle: MFGL, right: MFGC

inconsistent with the true cluster structure. This difference makes MFOG detect essential cluster structures more precisely than MFGL and MFGC, resulting in that MFOG could outperform MFGL and MFGC.

5.2 Real Document Datasets

We examined the performance of MFOG by using three different real text datasets: 20NewsGroups[3], Reuters-21758[4], and TDT2[5], again comparing with MFGL, MFGC, MFBS and KM.

For the three datasets, we first discarded all documents that are in multiple categories, to clarify the true clusters and then randomly selected 1,000 documents uniformly, which can be in the four largest

3. Available from http://qwone.com/ jason/20Newsgroups/

4. Available from http://www.daviddlewis.com/resources/testcollections/reuters21578/

5. Nist Topic Detection and Tracking corpus at http://www.nist.gov/speech/tests/tdt/tdt98/index.htm

TABLE 3
The number of documents in the nine largest topics.

	20NewsGroups	Reuters-21758	TDT2
N	8932	7195	7289
$\lvert\mathcal{G}_1\rvert$	999	3713	1844
$\lvert\mathcal{G}_2\rvert$	997	2055	1828
$\lvert\mathcal{G}_3\rvert$	996	321	1222
$\lvert\mathcal{G}_4\rvert$	994	298	811
$\lvert\mathcal{G}_5\rvert$	991	245	441
$\lvert\mathcal{G}_6\rvert$	990	197	407
$\lvert\mathcal{G}_7\rvert$	990	142	272
$\lvert\mathcal{G}_8\rvert$	988	114	238
$\lvert\mathcal{G}_9\rvert$	987	110	226

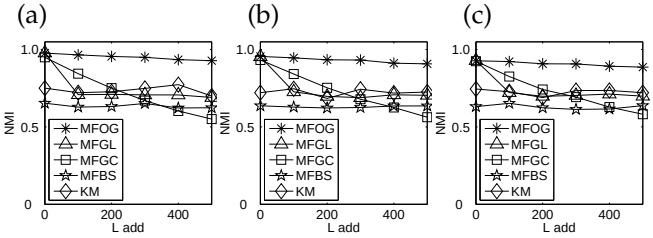

Fig. 13. NMI on **20Newsgroups**. The rate of labeled documents, R_s, was (a) 0.9, (b) 0.8 and (c) 0.7.

categories (corresponding to the four true clusters) and the 500 most frequent words, to generate the input document-term matrix. We computed TF-IDF values from the input matrix and normalized each column. By repeating the above manner, we generated 20 datasets for each of the above three datasets, meaning that the results were averaged over 20 runs. We generated overlapping groups by assigning group labels to L_{add} pairs of group–document randomly, meaning that for larger L_{add}, groups are overlapped more heavily. We tested $L_{add} = 0, 100, \ldots$ and 500. We finally generated unlabeled documents by discarding all group labels of randomly chosen documents, where the parameter we used was R_s, which was the rate of labeled documents in any groups, and $R_s = 0.9, 0.8$ and 0.7 were tested. Under each setting, we fixed $\alpha = 0.05$ while for β, twenty values between 10^{-3} and 10^3 were examined at an equal interval in the logarithmic scale to have the best average NMI. We initialized $\boldsymbol{V}$ using auxiliary group information as follows: for labeled instance n, $V_{n,k} = c_k$ if $n \in \mathcal{G}_k$, otherwise $V_{n,k} = \frac{c_k}{N}$, and for unlabeled instance $n \in \mathcal{G}_k$, $V_{n,k} = \frac{c_k}{K}$ $(k = 1, \ldots, K)$, where c_k is a constant value for normalizing each column of $\boldsymbol{V}$. We note that we need to initialize only $\boldsymbol{V}$, because for the t-th iteration, $\boldsymbol{U}$ is updated, depending on $\boldsymbol{V}^{\{t\}}$.

Figs. 13, 14 and 15 show the averaged NMI of five competing methods for 20NewsGroups, Reuters-21758 and TDT2, respectively. MFOG, MFGL, and MFGC outperformed MFBS and KM, for all cases, confirming the effectiveness of auxiliary group information. When $L_{add} = 0$, where groups do not overlap, MFOG, MFGL and MFGC were comparable

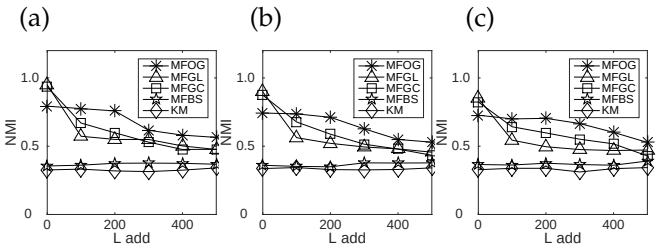

Fig. 14. NMI on **Reuters-21758**. The rate of labeled documents, R_s was (a) 0.9, (b) 0.8 and (c) 0.7.

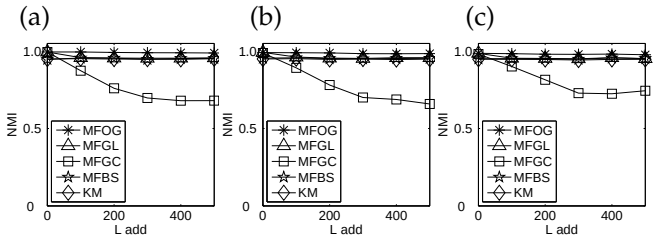

Fig. 15. NMI on **TDT2**. The rate of labeled documents, R_s was (a) 0.9, (b) 0.8 and (c) 0.7.

in performance with each other for 20Newsgroups and TDT2, while MFGC and MFGL were better than MFOG for Reuters-21758. On the other hand, when $L_{add} > 0$ where groups overlapped with each other, MFOG clearly outperformed MFGC and MFGL, particularly being significant for 20Newsgroups. In fact the performance of MFOG was kept very high for 20Newsgroups and TDT2, even for large L_{add}, i.e. a larger number of overlapping groups. This result indicates the robustness of MFOG against overlapping groups. The weakness of MFGL and MFGC regarding this point is caused by the strict regularization of graph Laplacian. This strong restriction of graph Laplacian causes hard to find the optimum weight β (in fact β is unstable), by which the curve of MFGC was fluctuating a lot.

We finally checked the NMI of five competing methods by changing the number of clusters under one typical setting. We note that this is a semi-supervised setting with heavily overlapped groups, which can be real world settings. Table 4 shows the averaged NMI over twenty runs for (a) 20NewsGroups, (b) Reuters-21758 and (c) TDT2. The largest NMI value for each column is indicated by boldface. This figure clearly shows that MFOG achieved the best performance for all 27 cases except seven cases, confirming the performance advantage of MFOG over the four competitive methods.

5.3 Real Gene Expression Datasets

We used two gene expression datasets: 1) Human tumor [30] and 2) Yeast cell cycle [31]. Human has expression values of 7129 genes from 42 human cells, and Yeast has those of 696 genes from 18 cells. We used Gene Ontology (GO) [11] to generate true clusters (each true set has three clusters) by the following procedure: We first chose the adequate size of GO terms: $50 \leq |\mathcal{G}| \leq 200$ for Human and $10 \leq |\mathcal{G}| \leq 50$ for Yeast and then discarded the GO terms with heavily overlapped with each other (Jaccard coefficient is over 0.75). We then computed the ratio of inter- and intra-cluster variance for all possible sets of three GO terms and chose the top 20 sets in terms of the largest ratios as the true cluster sets, removing the sets with genes, to which multiple true clusters are assigned. We then randomly generated additional groups, where the cluster size is the mean of the true clusters and randomly discarded group information of $(1 - R_s)N$ nodes for a semi-supervised setting.

Figs. 16 and 17 show the performance results on Human and Yeast, respectively, under the same parameter setting as the document clustering experiments. From these figures, the performance advantage of MFOG over the other methods is clear, particularly for a larger number of additional groups. This result indicates the high performance of MFOG for the setting of choosing correct clusters out of possible candidates.

TABLE 4
Performance results with $R_s = 0.7$ and $L_{add} = 500$ on (a) 20Newsgroups, (b) Reuters-21758, (c) TDT2.

(a)

G	MFOG	MFGL	MFGC	MFBS	KM
2	0.9457	0.9122	**0.9488**	0.9115	0.9246
3	**0.9393**	0.8773	0.7353	0.8747	0.8690
4	**0.8925**	0.7020	0.5829	0.6221	0.7570
5	**0.8773**	0.7157	0.6055	0.6783	0.6789
6	**0.8732**	0.6814	0.6136	0.6449	0.6410
7	**0.8374**	0.6506	0.6106	0.6090	0.6006
8	**0.8507**	0.6475	0.6416	0.6027	0.5956
9	**0.8012**	0.6002	0.6450	0.5405	0.6156

(b)

G	MFOG	MFGL	MFGC	MFBS	KM
2	0.3588	**0.4972**	0.2399	0.3477	0.3377
3	0.4172	**0.4219**	0.3881	0.3810	0.3615
4	**0.5180**	0.4526	0.4232	0.3607	0.3479
5	**0.5423**	0.4554	0.5260	0.3793	0.3570
6	0.5956	0.4590	**0.6113**	0.3453	0.3824
7	**0.5855**	0.4514	0.5678	0.3480	0.3958
8	**0.6028**	0.4465	0.5750	0.3283	0.4011
9	**0.6225**	0.4503	0.6180	0.3542	0.3833

(c)

G	MFOG	MFGL	MFGC	MFBS	KM
2	0.3588	**0.4972**	0.2399	0.3477	0.3377
3	0.4172	**0.4219**	0.3881	0.3810	0.3615
4	**0.5180**	0.4526	0.4232	0.3607	0.3479
5	**0.5423**	0.4554	0.5260	0.3793	0.3570
6	0.5956	0.4590	**0.6113**	0.3453	0.3824
7	**0.5855**	0.4514	0.5678	0.3480	0.3958
8	**0.6028**	0.4465	0.5750	0.3283	0.4011
9	**0.6225**	0.4503	0.6180	0.3542	0.3833

6 Concluding Remarks

We have proposed a new model of non-negative matrix factorization, MFOG, and efficient algorithms

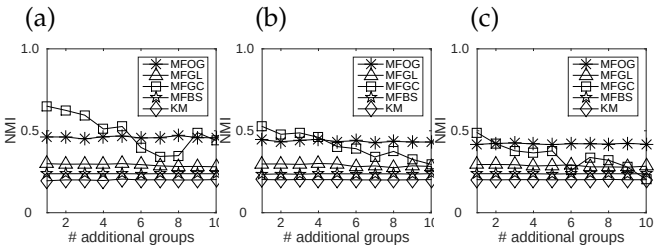

Fig. 16. NMI on Human. The rate of labeled genes R_s was (a) 0.9, (b) 0.8 and (c) 0.7.

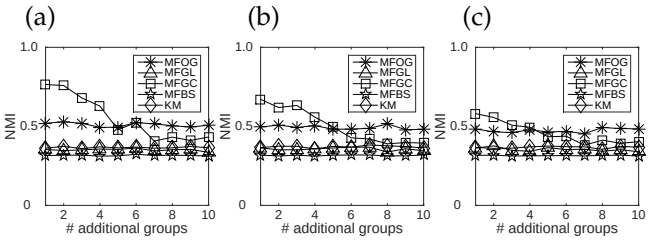

Fig. 17. NMI on Yeast. The rate of labeled genes R_s was(a) 1.0, (b) 0.9, (c) 0.8 and (d) 0.7.

for learning the model, to use auxiliary information on overlapping groups. Two key features of MFOG are a mixed norm with latent matrices, which allow to detect the true clusters, relaxing the excessive sparsity, and the efficient factorization algorithm based on a vector block coordinate descent method (vector-BCD). MFOG combines these two points cooperatively in terms that latent matrices allow to use vector-BCD. Experimental results with both synthetic and real dataset clearly showed the performance advantage of MFOG over four competing methods.

One issue is to find the optimal β (or best β), in terms of the highest performance. We checked the distribution of the best β obtained for all settings in experiments of each real dataset. Table 5 shows the variance of the best $\log(\beta)$ of three methods (MFOG, MFGL and MFGC), indicating that the variance of MFOG is the smallest and the best β was obtained most stably by MFOG. This result reveals that MFOG is easiest in terms of choosing β among the three competing methods.

Possible future work is to optimize the number of clusters, which might be possible by using another regularization term, such as [32]. Missing value estimation (or matrix completion) [33] would be also possible future work. We think that our framework with auxiliary information would be useful under these two issues.

TABLE 5
The variance of best $\log(\beta)$.

	MFOG	MFGL	MFGC
News20	**0.0705**	1.7495	0.7043
Reuters	**0.0593**	1.2211	1.1187
TDT2	**0.0802**	1.5348	2.2973
Human	**0.0519**	0.9256	0.8909
Yeast	**0.1009**	2.5736	1.0980

APPENDIX

Proof of Theorem 1

By some algebra, optimization of $\boldsymbol{Z}_{\cdot k}^{(g)}$ $(g = 1,\ldots,G')$ under fixed $\boldsymbol{U}$ and $\boldsymbol{Z}_{\cdot j}^{(c)}$ $(j \neq k,\ c \neq g)$ can be transformed from (7) to

$$\min_{\boldsymbol{Z}_{\cdot k}^{(g)} \geq 0} \frac{1}{2} \left\| \boldsymbol{Z}_{\cdot k}^{(g)} - \boldsymbol{s}_{LV}^{(g,k)} \right\|_2^2 + \lambda_{LV}^{(g)} \left\| \boldsymbol{Z}_{\cdot k}^{(g)} \right\|_2. \tag{22}$$

Taking Fenchel duality with dual variable $\boldsymbol{s} \geq 0$, the dual problem of (22) is given as follows:

$$\max_{\|\boldsymbol{s}\|_2 \leq \lambda_{LV}^{(g)}} \left\{ - \sup_{Z_{\cdot k}^{(g)} \geq 0} Z_{\cdot k}^{(g)\ T}(-\boldsymbol{s}) - \frac{1}{2} \left\| Z_{\cdot k}^{(g)} - \boldsymbol{s}_{LV}^{(g,k)} \right\|_2^2 \right\}$$

The minimum for $Z_{\cdot k}^{(g)}$ is given by the first derivative, and then we have $Z_{\cdot k}^{(g)} = \boldsymbol{s}_{LV}^{(g,k)} - \boldsymbol{s}$. Using this equation, the dual problem can be transformed to

$$\min_{\|\boldsymbol{s}\|_2 \leq \lambda_{LV}^{(g)}} \frac{1}{2} \left\| \boldsymbol{s} - \boldsymbol{s}_{LV}^{(g,k)} \right\|_2^2. \tag{23}$$

Using the solution of (23), which is given by

$$\boldsymbol{s}^* = \left(1 - \left[1 - \frac{\lambda_{LV}^{(g)}}{\|\boldsymbol{s}_{LV}^{(g,k)}\|_2} \right]_+ \right) \boldsymbol{s}_{LV}^{(g,k)} \tag{24}$$

and $Z_{\cdot k}^{(g)} = \boldsymbol{s}_{LV}^{(g,k)} - \boldsymbol{s}^*$, then we have (12).

Proof of Theorem 2

Each group $\mathcal{G}_g$ $(g = G'+1,\ldots,G)$ has only one instance and does not overlap with another group, obviously meaning $\boldsymbol{V}^{(g)} = \boldsymbol{Z}^{(g)}$ $(g = G'+1,\ldots,G)$. Thus we can directly optimize $\boldsymbol{V}^{(0)}$, which corresponds to $\mathcal{G}_0 = \cup_{g=G'+1}^{G} \mathcal{G}_g$. By the same manner as the proof of Theorem 1, we can have (14).

Proof of Theorem 3

Here we define a matrix for all latent matrices as

$$\boldsymbol{Z} = \left(\boldsymbol{Z}^{(1)T},\ \boldsymbol{Z}^{(2)T},\ldots,\ \boldsymbol{Z}^{(G)T} \right)^T.$$

Fixing $\boldsymbol{U}$, the optimization for $\boldsymbol{Z}$ is as follows:

$$\min_{\boldsymbol{Z} \geq 0} f_Z(\boldsymbol{Z}) + \beta\ \Omega_{1,2}^{\mathcal{G}}(\boldsymbol{Z}),$$

where

$$f_Z(\boldsymbol{Z}) \quad = \quad \frac{1}{2} \left\| \boldsymbol{X} - \boldsymbol{U} \left(\sum_{g=1}^{G} \boldsymbol{Z}^{(g)} \right)^T \right\|_F^2.$$

By a proximal gradient approach, updating $\boldsymbol{Z}$ from the current matrix after t-th update $\boldsymbol{Z}^{\{t\}}$ is given by

$$\boldsymbol{Z}^{\{t+1\}} \leftarrow \min_{\boldsymbol{Z} \geq 0} f_Z(\boldsymbol{Z}^{\{t\}}) + \nabla f_Z(\boldsymbol{Z}^{\{t\}})(\boldsymbol{Z} - \boldsymbol{Z}^{\{t\}}) + \beta \cdot \Omega_{1,2}^{\mathcal{G}}(\boldsymbol{Z}) + \frac{L_L}{2} \|\boldsymbol{Z} - \boldsymbol{Z}^{\{t\}}\|_F^2,$$

where L_L is the Lipchitz constant which is obtained by multiplying K by the maximum eigen values of $\boldsymbol{U}^T\boldsymbol{U}$. Calculating $f_Z(\boldsymbol{Z}^{\{t\}})$ and $\nabla f_Z(\boldsymbol{Z}^{\{t\}})$, the right side can be transformed as follows:

$$\min_{\boldsymbol{Z}\geq 0} \frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{S}_{LM}\|_F^2 + \frac{\beta}{L_L}\Omega_{1,2}^{\mathcal{G}}(\boldsymbol{Z}),$$

where the sub-matrix of $\boldsymbol{S}_{LM}$, related to group g ($g = 1,\ldots,G$), is described as $\boldsymbol{S}_{LM}^{(g)} \in \mathcal{R}^{|\mathcal{G}_g|\times K}$. We analytically derive the update via the dual problem. Here we define a dual matrix $\boldsymbol{S}$ where the size of $\boldsymbol{S}$ is equal to that of $\boldsymbol{Z}$. Then the dual problem is given as follows:

$$\max_{\boldsymbol{S}} \left\{ -\sup_{\boldsymbol{Z}} \left[\mathrm{Tr}\big(\boldsymbol{Z}^T(-\boldsymbol{S})\big) - \frac{1}{2}\Big\|\boldsymbol{Z}-\boldsymbol{S}_{LM}\Big\|_F^2 \right] \right\} \quad \text{such that } \ \Omega^*(\boldsymbol{S}) \leq \lambda_{LM}^{(g)}, \ \ g=1,\ldots,G,$$

where $\Omega^*(\cdot)$ is the dual norm of $\Omega_{1,2}^{\mathcal{G}}(\cdot)$. When $q = 2$, then $\Omega^*(\cdot)$ is the ℓ_2 norm. We note that function $\mathrm{Tr}\big(\boldsymbol{Z}^T(-\boldsymbol{S})\big) - \frac{1}{2}\big\|\boldsymbol{Z}-\boldsymbol{S}_{LM}\big\|_F^2$ is convex, by which the solution satisfies that the derivation of the above cost respect to $\boldsymbol{Z}$ is zero. Thus we have $\boldsymbol{Z} = \boldsymbol{S}_{LM} - \boldsymbol{S}$ as the solution. By substituting $\boldsymbol{Z}$ into the equation of the dual problem, the dual problem can be transformed to

$$\min_{\boldsymbol{S}} \frac{1}{2}\big\|\boldsymbol{S}-\boldsymbol{S}_{LM}\big\|_F^2 \ \text{ s.t. } \ \Omega^*(\boldsymbol{S}) \leq \lambda_{LM}^{(g)}, \ \ g=1,\ldots,G.$$

Then the optimal $\boldsymbol{S}^{(g)}$ related to each group $g = 1,\ldots,G$ can be given as follows:

$$\boldsymbol{S}^{*(g)} = \left(1 - \left[1 - \frac{\lambda_{LM}^{(g)}}{\big\|\boldsymbol{S}_{LM}^{(g)}\big\|_2}\right]_+\right)\boldsymbol{S}_{LM}^{(g)}.$$

Substituting the above equation into $\boldsymbol{Z} = \boldsymbol{S}_{LM} - \boldsymbol{S}^*$, we have (15).

Proof of Theorem 4

Because of equation $\boldsymbol{V}^{(g)} = \boldsymbol{Z}^{(g)}$ ($g = G'+1,\ldots,G$), we can update $\boldsymbol{V}^{(0)}$ simultaneously. Using the same manner of the proof of Theorem 3, we have (17).

Proof of Theorem 5

From (7), the optimization for $\boldsymbol{V}_{\cdot k}$ can be transformed into the following equation:

$$\min_{\boldsymbol{V}_{\cdot k}\geq 0,\ \boldsymbol{V}_{\cdot k}=\sum_g \boldsymbol{Z}_{\cdot k}^{(g)}} \frac{1}{2}\Big\|\boldsymbol{V}_{\cdot k}-\boldsymbol{s}_{DV}^{(k)}\Big\|_F^2 + \sum_{g=1}^{G}\lambda_{DV}^{(k,g)}\Big\|\boldsymbol{Z}_{\cdot k}^{g}\Big\|_2.$$

The dual problem of the above optimization is given as follows:

$$\max_{\boldsymbol{s}\in\mathcal{K}_k^G} \left\{ -\sup_{\boldsymbol{V}_{\cdot k}} \left[\boldsymbol{V}_{\cdot k}^T(-\boldsymbol{s}) - \frac{1}{2}\Big\|\boldsymbol{V}_{\cdot k}-\boldsymbol{s}_{DV}^{(k)}\Big\|_F^2 \right] \right\}$$

where $\boldsymbol{s}$ is a dual vector, which has the same size as $\boldsymbol{V}_{\cdot k}$., and $\mathcal{K}_k^G$ is the dual norm of $\sum_{g=1}^{G}\lambda_{DV}^{(k,g)}\big\|\boldsymbol{Z}_{\cdot k}^{g}\big\|_2$, which is the intersection of convex sets (or cylinders specifically) $\mathcal{K}_k^G$ By some algebra, the dual problem can be transformed to the projection as follows:

$$\mathrm{Proj}_{\mathcal{K}_k^G}(\boldsymbol{s}_{DV}^{(k)}) = \arg\min_{\boldsymbol{s}\in\mathcal{K}_k^G}\Big\|\boldsymbol{s}-\boldsymbol{s}_{DV}^{(k)}\Big\|_F^2.$$

By using Moreau's decomposition [34], the optimization for $\boldsymbol{V}_{\cdot k}$ can be performed by the following updating rule:

$$\boldsymbol{V}_{\cdot k} \leftarrow \boldsymbol{s}_{DV}^{(k)} - \mathrm{Proj}_{\mathcal{K}_k^G}\Big(\boldsymbol{s}_{DV}^{(k)}\Big).$$

We note that the projection can be performed more easily by checking only active groups $\widehat{\mathcal{G}}$ because of the following equation

$$\mathrm{Proj}_{\mathcal{K}_{\mu_k}^{\widehat{G}}}\Big(\boldsymbol{s}_{DV}^{(k)}\Big) = \mathrm{Proj}_{\mathcal{K}_{\mu_k}^{G}}\Big(\boldsymbol{s}_{DV}^{(k)}\Big).$$

Thus we have (18).

Proof of Theorem 6

In (8), the cost related to $\boldsymbol{V}$ is as follows:

$$f_V(\boldsymbol{V}) + \beta\cdot\Omega_{1,2}^{\mathcal{G}}(\boldsymbol{Z}), \quad \text{s.t. } \boldsymbol{V} = \sum_{g=1}^{G}\boldsymbol{Z}^{(g)},$$

where

$$f_V(\boldsymbol{V}) \ = \ \frac{1}{2}\Big\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}^T\Big\|_F^2.$$

By using a proximal gradient approach, updating $\boldsymbol{V}$ from the current matrix $\boldsymbol{V}^{\{t\}}$ is given by

$$\boldsymbol{V}^{\{t+1\}} \leftarrow \min_{\boldsymbol{V}\geq 0} f_V(\boldsymbol{V}^{\{t\}}) + \nabla f_V(\boldsymbol{V}^{\{t\}})(\boldsymbol{V}-\boldsymbol{V}^{\{t\}}) + \beta\cdot\Omega_{1,2}^{\mathcal{G}}(\boldsymbol{Z}) + \frac{L_D}{2}\big\|\boldsymbol{V}-\boldsymbol{V}^{\{t\}}\big\|_2^2.$$

where L_D is the Lipchitz constant which is obtained by the maximum eigen values of $\boldsymbol{U}^T\boldsymbol{U}$. Dividing this by constant L_D, the above optimization can be transformed to

$$\min_{\boldsymbol{V}\geq 0}\frac{1}{2}\big\|\boldsymbol{V}-\boldsymbol{S}_{DM}\big\|_2^2 + \frac{\beta}{L_D}\Omega_{1,2}^{\mathcal{G}}(\boldsymbol{Z}).$$

Similar to the derivation of Dir-Vec, the dual problem of the above optimization with dual matrix $\boldsymbol{S}$ is given as follows:

$$\mathrm{Proj}_{\mathcal{K}^G}(\boldsymbol{S}_{DM}) \ = \ \arg\min_{\boldsymbol{S}\in\mathcal{H}^G}\big\|\boldsymbol{S}-\boldsymbol{S}_{DM}\big\|_2^2$$

where $\mathcal{H}^G$ is the dual norm of norm $\frac{\beta}{L_D}\Omega_{1,2}^{\mathcal{G}}(\boldsymbol{Z})$. By using Moreau's decomposition [34], we have update rule (19).

ACKNOWLEDGMENTS

This work is partially supported by Okawa Foundation, JSPS KAKENHI 25870322 and 24300054, and Collaborative Research Program of Institute for Chemical Research, Kyoto University (grant #2012-24, #2013-19 and #2014-26).

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[2] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 12, pp. 4164–4169, March 2004.

[3] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 6, pp. 902–913, Jun. 2011.

[4] Y. Wang and M. Kitsuregawa, "Evaluating contents-link coupled web page clustering for web search results," in *CIKM*. New York, NY, USA: ACM, 2002, pp. 499–506.

[5] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *KDD*. New York, NY, USA: ACM, 2004, pp. 59–68.

[6] Y. Chen and L. Wang, "Non-negative matrix factorization for semisupervised heterogeneous data coclustering," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1459–1474, October 2010.

[7] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[8] R. Rak, L. A. Kurgan, and M. Reformat, "Multilabel associative classification categorization of medline articles into mesh keywords," *IEEE Eng. Med. Biol. Mag.*, vol. 26, no. 2, pp. 47–55, 2007.

[9] I. Sato and H. Nakagawa, "Knowledge discovery of multiple-topic document using parametric mixture model with dirichlet prior," in *KDD*. New York, NY, USA: ACM, 2007, pp. 590–598.

[10] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Güldener, G. Mannhaupt, M. Münsterkötter, and H. W. Mewes, "The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes," *Nucleic Acids Res.*, vol. 32, no. 18, pp. 5539–45, 2004.

[11] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[12] L. Jacob, G. Obozinski, and J. P. Vert, "Group lasso with overlap and graph lasso," in *ICML*. New York, NY, USA: ACM, June 2009, pp. 433–440.

[13] A. Cichocki, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE Transactions on Fund. Elec. Comm. and Comp.*, vol. 92, no. 3, pp. 708–721, March 2009.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, vol. 13. Denver, CO: MIT Press, November 2001, pp. 556–562.

[15] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM J. Sci. Comput.*, vol. 33, no. 6, pp. 3261–3281, Nov. 2011.

[16] J. Kim, R. Monteiro, and H. Park, "Group sparsity in nonnegative matrix factorization," in *SDM*. Anaheim, California: SIAM, April 2012, pp. 851–862.

[17] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. B*, vol. 68, no. 1, pp. 49–67, February 2006.

[18] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, Nov. 2011.

[19] N. Simon and R. Tibshirani, "Standardization and the group lasso penalty," *Statistica Sinica*, vol. 22, no. 3, pp. 983–1001, 2012.

[20] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Comp. Stat. Data Anal.*, vol. 52, no. 1, pp. 155–173, September 2007.

[21] S. Mosci, S. Villa, A. Verri, and L. Rosasco, "A primal-dual algorithm for group sparse regularization with overlapping groups," in *NIPS*, vol. 23. Vancouver, British Columbia, Canada: Curran Associates, Inc., December 2010, pp. 2604–2612.

[22] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

[23] Y.-X. Wang and Y.-J. Zhang, "Non-negative matrix factorization: a comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 99, no. PrePrints, 2012.

[24] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," in *NIPS*. Vancouver, British Columbia, Canada: MIT Press, December 2002, pp. 505–512.

[25] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized non-negative matrix factorization for data representation," *IEEE Trans. Pat. Anal. Mach. Intel.*, vol. 33, no. 8, pp. 1548–1560, Dec. 2011.

[26] H. Wang, F. Nie, H. Huang, and F. Makedon, "Fast nonnegative matrix tri-factorization for large-scale data co-clustering," in *IJCAI*. Barcelona, Catalonia, Spain: AAAI Press, 2011, pp. 1553–1558.

[27] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *NIPS*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 2009, pp. 82–89.

[28] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *AISTATS*, Chia Laguna Resort, Sardinia, Italy, May 2010, pp. 366–373.

[29] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *SIGIR*. New York, NY, USA: ACM, July 2003, pp. 267–273.

[30] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau *et al.*, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[31] P. Lu, A. Nakorchevskiy, and E. M. Marcotte, "Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 18, pp. 10 370–10 375, 2003.

[32] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1592–1605, 2013.

[33] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[34] R. Tomioka, T. Suzuki, and M. Sugiyama, "Super-linear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation," *J. Mach. Learn. Res.*, vol. 12, pp. 1537–1586, Jul. 2011.

Motoki Shiga received his B.E., M.E. and Ph.D. in Information Science from Gifu University in 2001, 2003 and 2006, respectively. He has been working on data mining, machine learning and their applications to bioinformatics.

Hiroshi Mamitsuka received B.S. in Biophysics and Biochemistry, M.E. in Information Engineering and PhD in Information Sciences all from the University of Tokyo in 1988, 1991 and 1999, respectively. He has been working in machine learning, data mining and bioinformatics. His current research interests are in mining from graphs and networks in biology and chemistry.