Group Sparsity in Nonnegative Matrix Factorization*

Jingu Kim[†] Renato D. C. Monteiro[‡] Haesun Park[†]

Abstract

A recent challenge in data analysis for science and engineering is that data are often represented in a structured way. In particular, many data mining tasks have to deal with group-structured prior information, where features or data items are organized into groups. In this paper, we develop group sparsity regularization methods for nonnegative matrix factorization (NMF). NMF is an effective data mining tool that has been widely adopted in text mining, bioinformatics, and clustering, but a principled approach to incorporating group information into NMF has been lacking in the literature. Motivated by an observation that features or data items within a group are expected to share the same sparsity pattern in their latent factor representation, we propose mixed-norm regularization to promote group sparsity in the factor matrices of NMF. Group sparsity improves the interpretation of latent factors. Efficient convex optimization methods for dealing with the mixed-norm term are presented along with computational comparisons between them. Application examples of the proposed method in factor recovery, semi-supervised clustering, and multilingual text analysis are demonstrated.

1 Introduction

Factorizations and low-rank approximations of matrices have been one of the most fundamental tools in machine learning and data mining. Singular value decomposition (SVD) and principal component analysis (PCA), for example, played a pivotal role in dimension reduction and noise removal. Constrained low-rank factorizations have also been widely used; among them, nonnegative matrix factorization (NMF) imposes nonnegativity constraints on the low-rank factor matrices, and the nonnegativity constraints enable natural interpretation of discovered latent factors [20]. Algorithms and applications of NMF have received much attention due to numerous successes in text mining, bioinformatics, blind source separation, and computer vision.

In this paper, we propose an extension of NMF that incorporates group structure as prior information. Many matrix-represented data sets are inherently structured as groups. For example, a set of documents that are labeled with the same topic forms a group. In drug discovery, various treatment methods are typically applied to a large number of subjects, and subjects that received the same treatment are naturally viewed as a group. In addition to these groups of data items, a set of features forms a group as well. In computer vision, different types of features such as pixel values, gradient features, and 3D pose features can be viewed as groups. Similarly in bioinformatics, features from microarray and metabolic profiling become different groups.

The motivation of our work is that there are similarities among data items or features belonging to the same group in that their low-rank representations share the same sparsity pattern. However, such similarities have not been previously utilized in NMF. In order to exploit the shared sparsity pattern, we propose to incorporate mixed-norm regularization in NMF. Our approach is based on $l_{1,q}$-norm regularization (See Section 3 for the definition of $l_{1,q}$-norm). Regularization by l_1-norm is well-known to promote a sparse representation [31]. When this approach is extended to groups of parameters, $l_{1,q}$-norm has been shown to induce a sparse representation at the level of groups [35]. By employing $l_{1,q}$-norm regularization, the latent factors obtained by NMF can be improved with an additional property of shared sparsity.

The adoption of mixed-norm regularization introduces a new challenge to the optimization algo-

[*] The work J. Kim and H. Park was supported in part by the National Science Foundation (NSF) grants CCF-0732318 and CCF-0808863. The work of R. D. C. Monteiro was partially supported by NSF grants CCF-0808863 and CMMI-0900094 and ONR grant ONR N00014-11-1-0062. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[†] School of Computational Science and Engineering, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332-0280, USA (email: {jingu,hpark}@cc.gatech.edu).

[‡] School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, Atlanta, GA 30332-0205, USA (email: monteiro@isye.gatech.edu).

rithm for NMF. Since the mixed-norm term is not a smooth function, conventional methods such as the steepest gradient descent cannot be applied. To address the difficulty, we present two algorithms based on recent developments in convex optimization. Both algorithms are developed using the block coordinate descent (BCD) method [4]. The first approach is a matrix-block BCD method, in which one of the two factor matrices is updated at each step fixing the other. The second approach is a vector-block BCD method, in which one column of a factor matrix is updated at each step fixing all other values. A strength of the two algorithms we propose is that they generally handle $l_{1,q}$-norm regularization for common cases: $q = \infty$ and $q = 2$. We also provide computational comparisons of the two methods.

We show the effectiveness of mixed-norm regularization for factor recovery using a synthetic data set. In addition, we demonstrate application examples in semi-supervised clustering and multilingual text mining. Our application examples are novel in that the use of group sparsity regularization for these applications has not been shown before. In the applications, the benefits of nonnegativity constraints and group sparsity regularization are successfully combined demonstrating that the mixed-norm regularized NMF can be effectively used for real-world data mining applications.

The rest of this paper is organized as follows. We begin with discussion on related work in Section 2. We then introduce the concept of group sparsity and lead to a problem formulation of NMF with mixed-norm regularization in Section 3. We describe optimization algorithms in Section 4. We provide the demonstration of recovery example, application examples, and computational comparisons in Section 5. We finalize the paper with discussion in Section 6.

Notations Notations used in this paper are as follows. A lowercase or an uppercase letter, such as x or X, denotes a scalar. Boldface lowercase and uppercase letters, such as $\mathbf{x}$ and $\mathbf{X}$, denote a vector and a matrix, respectively. Indices typically grow from 1 to an uppercase letter, e.g., $n \in \{1, \cdots, N\}$. Elements of a sequence are denoted by superscripts within parentheses, e.g., $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(N)}$, and the entire sequence is denoted by $\{\mathbf{X}^{(n)}\}$. For a matrix $\mathbf{X}$, $\mathbf{x}_{\cdot i}$ or $\mathbf{x}_i$ denotes its i^{th} column, $\mathbf{x}_{i\cdot}$ denotes its i^{th} row, and x_{ij} denotes its $(i, j)^{th}$ element. The set of nonnegative real numbers is denoted by $\mathbb{R}_+$, and $\mathbf{X} \geq 0$ indicates that the elements of $\mathbf{X}$ are nonnegative.

2 Related Work

Incorporating group information using mixed-norm regularization has been previously discussed in statistics and machine learning. Earlier, regularization for sparse representation was popularized with the l_1-norm penalized linear regression called Lasso [31]. L_1-norm penalization is known to promote a sparse solution and improve generalization. Techniques for promoting group sparsity using $l_{1,2}$-norm regularization have been investigated by Yuan and Lin and others [35, 19, 26] under the name of group Lasso. Approaches that adopt $l_{1,\infty}$-norm regularization have been subsequently proposed by Liu et al. and others [23, 29, 7] for multi-task learning problems. Regularization methods for more sophisticated structure have also been proposed recently [18, 24].

In matrix factorization, Bengio et al. [3] and Jenatton et al. [12] considered $l_{1,2}$-norm regularization in sparse coding and principal component analysis, respectively. Jenatton et al. [11] further considered hierarchical regularization with tree structure. Jia et al. [13] recently applied $l_{1,\infty}$-norm regularization to sparse coding with a focus on a computer vision application. Masaeli et al. [25] used the idea of $l_{1,\infty}$-norm regularization for feature selection in PCA. The group structure studied in our work is close to those of [3, 12, 13] since they also considered group sparsity shared across data items or features. On the other hand, the hierarchical regularization in [11] is different from ours because their regularization was imposed on parameters *within* each data item. In addition, we focus on *nonnegative* factorization in algorithm development as well as in applications whereas [3, 12, 13] focused on sparse coding or PCA.

In NMF literature, efforts to incorporate group structure have been fairly limited. Badea [2] presented a simultaneous factorization of two gene expression data sets by extending NMF with an offset vector, as in the affine NMF [9]. Li et al. [21] and Singh and Gordon [30] demonstrated how simultaneous factorization of multiple matrices can be used for knowledge transfer. Jenatton et al. [11] mentioned NMF as a special case in their work on sparse coding, but they only dealt with a particular example without further developments. In addition, the hierarchical structure considered in [11] is different from ours as explained in the previous paragraph. To our knowledge, algorithms and applications of applying group sparsity regularization to NMF have not been fully investigated before our work in this paper.

Efficient optimization methods presented in this paper are built upon recent developments in convex optimization and NMF algorithms. The block coordinate descent (BCD) method forms the basis of our algorithms. In our first algorithm, which is a matrix-block BCD method, we adopt an efficient convex optimization method in [32]. The motivation of our second algorithm, which is a vector-block BCD method, is from the hierarchical alternating least squares (HALS) method [8] for standard NMF. Convex optimization theory, in particular the Fenchel duality [5, 1], plays an important role in both of the proposed algorithms.

3 Problem Statement

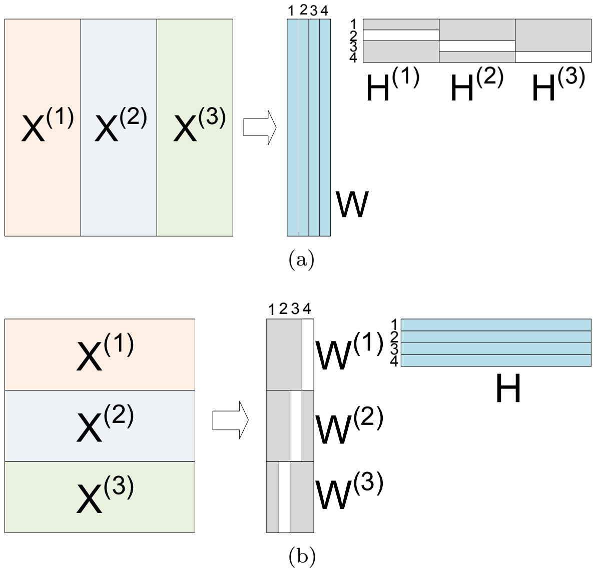
Let us begin our main discussion with a matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$. Without loss of generality, we assume that the rows of $\mathbf{X}$ represent features and the columns of $\mathbf{X}$ represent data items. In standard NMF, we are interested in discovering two low-rank factor matrices $\mathbf{W} \in \mathbb{R}_+^{m\times k}$ and $\mathbf{H} \in \mathbb{R}_+^{k\times n}$ such that $\mathbf{X} \approx \mathbf{WH}$. This is typically achieved by minimizing an objective function defined as

$$f(\mathbf{W},\mathbf{H}) = \frac{1}{2}\left\|\mathbf{X}-\mathbf{WH}\right\|_F^2 . \tag{3.1}$$

with constraints $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$. In this section, we show how we can take group structure into account by adding a mixed-norm regularization term into (3.1).

3.1 Group structure and group sparsity Let us first describe the group structure using motivating examples. Diagrams in Figure 1 show group structure considered in our work. In Figure 1-(a), the columns (that is, data items) are divided into three groups. This group structure is prevalent in clustered data, where data items belonging to each cluster define a group. In text mining, a group can represent documents having the same topic assignment. Another example can be seen from bioinformatics. For the purpose of drug discovery, one typically applies various treatment options to different groups of subjects. In this case, it is important to analyze the difference at the level of groups and discover fundamental understanding of the treatments. Subjects to which the same treatment is applied can be naturally viewed as a group.

On the other hand, groups can be formed from the rows (that is, features) as shown in Figure 1-(b). This structure can be seen from multi-view learning problems. In computer vision, as Jia et al. [13] discussed, a few different feature types such as pixel values, gradient-based features, and 3D pose features can be simultaneously used to build a recognition system. In text mining, a parallel multilingual corpus can be seen as a multi-view data set, where the term-document frequency features in each language form a group.



(a)

(b)

Figure 1: (a) Matrix with column groups and its factorization with group sparsity (b) Matrix with row groups and its factorization with group sparsity. The bright rows of $\mathbf{H}^{(i)}$ in (a) and the bright columns of $\mathbf{W}^{(i)}$ in (b) $(i = 1, 2, 3)$ represent zero subvectors.

Our motivation is that the feature or data instances that belong to a group are expected to share the same sparsity pattern in low-rank factors. In Figure 1-(a), the gray and white rows in $\mathbf{H}^{(1)}$, $\mathbf{H}^{(2)}$ and $\mathbf{H}^{(3)}$ represent nonzero and zero values, respectively. For example, columns in $\mathbf{H}^{(1)}$ share the same sparsity pattern that their second components are all zero. Such group sparsity improves the interpretation of the factorization model: For the reconstruction of data items in $\mathbf{X}^{(1)}$, only the first, the third, and the fourth latent components are used whereas the second component is irrelevant. Similar explanation holds for $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$ as well. That is, the association of latent components to data items can be understood at the level of groups instead of each data item.

In Figure 1-(b), group sparsity is shown for latent component matrices $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, and $\mathbf{W}^{(3)}$. A common interpretation of multi-view matrix factorization is that the i^{th} columns of $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, and $\mathbf{W}^{(3)}$ are associated with each other in a sense that they play the same role in explaining data. With group sparsity, missing associations can be discovered as follows. In Figure 1-(b), the second columns of $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are associated with each other, but there is no corresponding column in the third view since the second column of $\mathbf{W}^{(3)}$ appeared as zero.

Examples and interpretations provided here are not exhaustive, and we believe the group structure can be found in many other data mining problems. With these motivations in mind, now we proceed to discuss how group sparsity can be promoted by employing mixed-norm regularization.

3.2 Formulation with mixed-norm regularization

We discuss using the case of Figure 1-(a), where the columns are divided into groups. By considering the factorization of $\mathbf{X}^T$, however, all the formulations can be applied to the case of row groups.

Suppose the columns of $\mathbf{X} \in \mathbb{R}^{m \times n}$ are divided into B groups as $\mathbf{X} = \left(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(B)}\right)$, where $\mathbf{X}^{(b)} \in \mathbb{R}^{m \times n_b}$ and $\sum_{b=1}^{B} n_b = n$. Accordingly, the coefficient matrix is divided into B groups as $\mathbf{H} = \left(\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}\right)$ where $\mathbf{H}^{(b)} \in \mathbb{R}^{k \times n_b}$. The objective function in (3.1) is now written as a sum:

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \sum_{b=1}^{B} \left\| \mathbf{X}^{(b)} - \mathbf{W}\mathbf{H}^{(b)} \right\|_F^2 .$$

To promote group sparsity, we add a mixed-norm regularization term for coefficient matrices $\left\{\mathbf{H}^{(b)}\right\}$ using $l_{1,q}$-norm and consider the optimization problem

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} f(\mathbf{W}, \mathbf{H}) + \alpha \left\|\mathbf{W}\right\|_F^2 + \beta \sum_{b=1}^{B} \left\|\mathbf{H}^{(b)}\right\|_{1,q} . \tag{3.2}$$

We show the definition of $\|\cdot\|_{1,q}$ below. The Frobenius norm regularization on $\mathbf{W}$ is used to prevent the elements of $\mathbf{W}$ from growing arbitrarily large. Parameters α and β control the strength of each regularization term.

Now let us discuss the role of $l_{1,q}$-norm regularization. The $l_{1,q}$-norm of $\mathbf{Y} \in \mathbb{R}^{a \times c}$ is defined by

$$\|\mathbf{Y}\|_{1,q} = \sum_{j=1}^{a} \|\mathbf{y}_{j\cdot}\|_q = \|\mathbf{y}_{1\cdot}\|_q + \cdots + \|\mathbf{y}_{a\cdot}\|_q .$$

That is, the $l_{1,q}$-norm of a matrix is the sum of vector l_q-norms of its rows. Penalization with $l_{1,q}$-norm promotes as many number of zero rows as possible to appear in $\mathbf{Y}$. In (3.2), the penalty term on $\mathbf{H}^{(b)}$ promotes that coefficient matrices $\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}$ contain as many zero rows as possible, and the zero-rows correspond to group sparsity described in Section 3.1. Any scalar q, $1 < q \leq \infty$, can be potentially used, but for the development of algorithms, we focus on the two cases of $q = 2$ and $q = \infty$, which are common in related literature discussed in Section 2. In the following, we describe efficient optimization strategies for solving (3.2).

4 Optimization Algorithms

With mixed-norm regularization, the minimization problem (3.2) becomes more difficult than the standard NMF problem. We here propose two strategies based on the block coordinate descent (BCD) method in non-linear optimization [4]. The first method is a BCD method with matrix blocks; that is, a matrix variable is minimized at each step fixing all other entries. The second method is a BCD method with vector blocks; that is, a vector variable is minimized at each step fixing all other entries. In both algorithms, the $l_{1,q}$-norm term is handled via Fenchel duality [1, 5].

4.1 Matrix-block BCD method

The matrix-block BCD method minimizes the objective function of (3.2) with one (sub)matrix at a time fixing all other variables. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Matrix-block BCD method for (3.2)

```
Input: $\mathbf{X} = \left(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(B)}\right) \in \mathbb{R}^{m \times n}$ ,$\alpha, \beta \in \mathbb{R}_+$
Output: $\mathbf{W} \in \mathbb{R}_+^{m \times k}$, $\mathbf{H} = \left(\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}\right) \in \mathbb{R}_+^{k \times n}$
Initialize $\mathbf{W}$ and $\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}$, e.g., with random entries.
repeat
    Update $\mathbf{W}$ as
        (4.3) $\mathbf{W} \leftarrow \arg\min_{\mathbf{W} \geq 0} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \alpha \|\mathbf{W}\|_F^2 .$
    For each $b = 1, \cdots, B$, update $\mathbf{H}^{(b)}$ as
        (4.4) $\mathbf{H}^{(b)} \leftarrow \arg\min_{\mathbf{H}^{(b)} \geq 0} \frac{1}{2} \left\|\mathbf{X}^{(b)} - \mathbf{W}\mathbf{H}^{(b)}\right\|_F^2 + \beta \left\|\mathbf{H}^{(b)}\right\|_{1,q} .$
until convergence
```

Subproblem (4.3) for $\mathbf{W}$ is easy to solve as it can be transformed to

$$\mathbf{W} \leftarrow \arg\min_{\mathbf{W} \geq 0} \frac{1}{2} \left\| \begin{pmatrix} \mathbf{H}^T \\ \sqrt{2\alpha}\mathbf{I}_k \end{pmatrix} \mathbf{W}^T - \begin{pmatrix} \mathbf{X}^T \\ \mathbf{0}_{k \times m} \end{pmatrix} \right\|_F^2 , \tag{4.5}$$

which is the nonnegativity-constrained least squares

Algorithm 2 A convex optimization method for (4.7)

```
Input: $\mathbf{B} \in \mathbb{R}^{p \times r}$, $\mathbf{C} \in \mathbb{R}^{p \times t}$, $\beta \in \mathbb{R}_+$
Output: $\mathbf{Z} \in \mathbb{R}_+^{r \times t}$
Choose $\mathbf{Z}^{(0)}$,$\tilde{\mathbf{Z}}^{(0)}$ and let $\tau^{(0)} = 1$ and $L = \sigma_{max}\left(\mathbf{B}^T\mathbf{B}\right)$.
for $k = 0, 1, 2, \cdots$, until convergence do
    $\mathbf{Y}^{(k)} \leftarrow \tau^{(k)}\mathbf{Z}^{(k)} + (1 - \tau^{(k)})\tilde{\mathbf{Z}}^{(k)}$
    Update
    (4.6)
    $\mathbf{Z}^{(k+1)} \leftarrow \arg\min_{\mathbf{Z} \geq 0} \left\|\mathbf{Z} - \mathbf{U}^{(k)}\right\|_F^2 + \frac{2\beta}{\tau^{(k)}L} \|\mathbf{Z}\|_{1,q}$,
    where $\mathbf{U}^{(k)} = \mathbf{Z}^{(k)} - \frac{1}{\tau^{(k)}L}\left(\mathbf{B}^T\mathbf{B}\mathbf{Y}^{(k)} - \mathbf{B}^T\mathbf{C}\right)$.
    $\tilde{\mathbf{Z}}^{(k+1)} \leftarrow \tau^{(k)}\mathbf{Z}^{(k+1)} + (1 - \tau^{(k)})\tilde{\mathbf{Z}}^{(k)}$
    Find $\tau^{(k+1)} > 0$ such that
        $\left(\tau^{(k+1)}\right)^{-2} - \left(\tau^{(k+1)}\right)^{-1} = \left(\tau^{(k)}\right)^{-2}$.
end for
Return $\tilde{\mathbf{Z}}^{(k)}$.
```

(NNLS) problem. An efficient algorithm for the NNLS problem, such as in [22, 14, 16, 17], can be used to solve (4.5). Solving subproblem (4.4) for $\mathbf{H}^{(b)}$ is a more involved task, and an algorithm for this problem is discussed in the following.

Subproblem (4.4) can be written as the following general form. Given two matrices $\mathbf{B} \in \mathbb{R}_+^{p \times r}$ and $\mathbf{C} \in \mathbb{R}_+^{p \times t}$, we would like to solve

$$\min_{\mathbf{Z} \geq 0} \frac{1}{2} \|\mathbf{BZ} - \mathbf{C}\|_F^2 + \beta \|\mathbf{Z}\|_{1,q}. \tag{4.7}$$

Observe that the objective function of (4.7) is composed of two terms: $g(\mathbf{Z}) = \frac{1}{2}\|\mathbf{BZ} - \mathbf{C}\|_F^2$ and $h(\mathbf{Z}) = \beta \|\mathbf{Z}\|_{1,q}$. Both $g(\mathbf{Z})$ and $h(\mathbf{Z})$ are convex functions, the first term $g(\mathbf{Z})$ is differentiable, and $\nabla g(\mathbf{Z})$ is Lipschitz continuous. Hence, an efficient convex optimization method can be adopted.

Algorithm 2 presents a variant of Nesterov's first order method, suitable for solving (4.7). The Nesterov's method and its variants have been widely used due to its simplicity, theoretical strength, and empirical efficiency (See, for example, [27, 32]). An important requirement in Algorithm 2 is the ability to efficiently solve subproblem (4.6). Observe that the problem can be separated with respect to each row of $\mathbf{Z}$. Focusing on the i^{th} row of $\mathbf{Z}$, it suffices to solve a problem in the following form:

$$\min_{\mathbf{z} \geq 0} \frac{1}{2} \|\mathbf{z} - \mathbf{v}\|_2^2 + \eta \|\mathbf{z}\|_q \tag{4.8}$$

where $\mathbf{z}$, $\mathbf{v}$, and η replace $\mathbf{z}_{i\cdot}$, $\left(\mathbf{U}^{(k)}\right)_{i\cdot}$, and $\frac{\beta}{\tau^{(k)}L}$, respectively.

It is important to observe that this problem can be handled without the nonnegativity constraints. The following proposition summarizes this observation. Jenatton et al. [11] briefly mentioned the statement but did not provide the proof. Let $[\cdot]_+$ denote the element-wise projection operator to nonnegative numbers.

Proposition 1. *Consider minimization problem* (4.8) *and the following minimization problem:*

$$\min_{\mathbf{z}} \frac{1}{2} \left\|\mathbf{z} - [\mathbf{v}]_+\right\|_2^2 + \eta \|\mathbf{z}\|_q. \tag{4.9}$$

If $\mathbf{z}^*$ *is the minimizer of* (4.9)*, then* $\mathbf{z}^*$ *is element-wise nonnegative, and it also attains the global minimum of* (4.8)*.*

Proof. The nonnegativity of $\mathbf{z}^*$ can be seen by the fact that any negative element can be set as zero decreasing the objective function of (4.9). The remaining relationship can be seen by considering an intermediate problem

$$\min_{\mathbf{z} \geq 0} \frac{1}{2} \left\|\mathbf{z} - [\mathbf{v}]_+\right\|_2^2 + \eta \|\mathbf{z}\|_q. \tag{4.10}$$

Comparing (4.9) and (4.10), since the minimizer $\mathbf{z}^*$ of the unconstrained problem in (4.9) satisfies nonnegativity, it is clearly a minimizer of the constrained problem in (4.10). Now, let the minimizer of (4.8) be $\tilde{\mathbf{z}}^*$, and consider the set of indices $\mathcal{N} = \{i : v_i \leq 0\}$. Then, it is easy to check $\left(\tilde{\mathbf{z}}^*\right)_i = \left(\mathbf{z}^*\right)_i = 0$ for all $i \in \mathcal{N}$. Moreover, ignoring the variables corresponding to $\mathcal{N}$, problems (4.8) and (4.10) are equivalent. Therefore, $\mathbf{z}^*$ is the minimizer of (4.8). □

Proposition 1 transforms (4.8) into (4.9), where the nonnegativity constraints are dropped. This transformation is important since (4.9) can now be solved via Fenchel duality as follows. According to Fenchel duality [5, 1], the following problem is dual to (4.9):

$$\min_{\mathbf{s}} \frac{1}{2} \left\|\mathbf{s} - [\mathbf{v}]_+\right\|_2^2 \text{ such that } \|\mathbf{s}\|_{q^*} \leq \eta, \tag{4.11}$$

where $\|\cdot\|_{q^*}$ is the dual norm of $\|\cdot\|_q$. Problem (4.11) is a projection problem to a l_{q^*}-norm ball of size η. We refer readers to [1, 5] and references therein for more details of dual norm. In our discussion, it suffices to note that the dual norm of $\|\cdot\|_2$ is itself, and the dual norm of $\|\cdot\|_\infty$ is $\|\cdot\|_1$. Therefore, problem (4.11) with $q = 2$ becomes

$$\min_{\mathbf{s}} \frac{1}{2} \left\|\mathbf{s} - [\mathbf{v}]_+\right\|_2^2 \text{ such that } \|\mathbf{s}\|_2 \leq \eta, \tag{4.12}$$

which can be solved simply by normalization. With $q = \infty$, (4.11) is written as

$$\min_{\mathbf{s}} \frac{1}{2} \left\| \mathbf{s} - [\mathbf{v}]_+ \right\|_2^2 \text{ such that } \|\mathbf{s}\|_1 \leq \eta, \tag{4.13}$$

which can be solved as described in [28, 10]. Once the minimizer $\mathbf{s}^*$ of (4.11) is computed, the optimal solution for (4.8) is found as $\mathbf{z}^* = [\mathbf{v}]_+ - \mathbf{s}^*$.

4.2 Vector-block BCD Method

The matrix-block BCD algorithm has been shown to be quite successful for NMF and its variations. However, recent observations [17] indicate that the vector-block BCD method [8] is also very efficient, often outperforming the matrix-block BCD method. Accordingly, we develop the vector-block BCD method for (3.2) as follows.

In the vector-block BCD method, optimal solutions to subproblems with respect to each column of $\mathbf{W}$ and each rows of $\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(b)}$ are sought. The overall procedure is shown in Algorithm 3.

Algorithm 3 Vector-block BCD method for (3.2)

```
Input: $\mathbf{X} = (\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(B)}) \in \mathbb{R}^{m\times n}$ ,$\alpha, \beta \in \mathbb{R}_+$
Output: $\mathbf{W} \in \mathbb{R}+^{m\times k}$, $\mathbf{H} = (\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}) \in \mathbb{R}_+^{k\times n}$
Initialize $\mathbf{W}$ and $\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}$, e.g., with random entries.
repeat
    For each $i = 1, \cdots, k$, update $\mathbf{w}_i (\in \mathbb{R}^{m\times 1})$ as
    (4.14)
        $\mathbf{w}_i \leftarrow \arg\min_{\mathbf{w}\geq 0} \frac{1}{2} \|\mathbf{R}_i - \mathbf{w}\mathbf{h}_{i\cdot}\|_F^2 + \alpha \|\mathbf{w}\|_2^2 ,$
    where $\mathbf{R}_i = \mathbf{X} - \sum_{j=1, j\neq i}^{k} \mathbf{w}_j \mathbf{h}_{j\cdot}$.
    For each $b = 1, \cdots, B$ and then for each $i = 1, \cdots, k$, update $\mathbf{h}_{i\cdot}^{(b)} (\in \mathbb{R}^{1\times n_b})$ as
    (4.15)
        $\mathbf{h}_{i\cdot}^{(b)} \leftarrow \arg\min_{\mathbf{h}\geq 0} \frac{1}{2} \left\| \mathbf{R}_i^{(b)} - \mathbf{w}_i \mathbf{h} \right\|_F^2 + \beta \|\mathbf{h}\|_q ,$
    where $\mathbf{R}_i^{(b)} = \mathbf{X}^{(b)} - \sum_{j=1, j\neq i}^{k} \mathbf{w}_j \mathbf{h}_{j\cdot}^{(b)}$.
until convergence
```

The solution of (4.14) is given as a closed form:

$$\mathbf{w}_i \leftarrow \left[\frac{\mathbf{R}_i \mathbf{h}_{i\cdot}^T}{2\alpha + \|\mathbf{h}_{i\cdot}\|^2} \right]_+ .$$

Subproblem (4.15) is easily seen to be equivalent to

$$\min_{\mathbf{h}\geq 0} \frac{1}{2} \left\| \mathbf{h} - \frac{\left(\mathbf{R}_i^{(b)}\right)^T \mathbf{w}_i}{\|\mathbf{w}_i\|_2^2} \right\|_2^2 + \frac{\beta}{\|\mathbf{w}_i\|^2} \|\mathbf{h}\|_q , \tag{4.16}$$

which is a special case of (4.8). Therefore, (4.16) can be solved via Proposition 1 and the dual problem (4.11).

Remark. It is worth emphasizing the characteristics of the matrix-block and the vector-block BCD methods. The optimization variables of (3.2) are $(\mathbf{W}, \mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)})$, and the BCD method divides these variables into a set of blocks. The matrix-block BCD method in Algorithm 1 divides the variables into $(B+1)$ blocks represented by $\mathbf{W}, \mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}$. The vector-block BCD method in Algorithm 3 divides the variables into $k(B+1)$ blocks represented by the columns of $\mathbf{W}, \mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}$. Both methods eventually rely on Proposition 1 and the dual problem (4.11). Although both methods share the same convergence property that every limit point is stationary [4], but their actual efficiency may be different as we show in Section 5.4.

5 Implementation Results

Our implementation section is composed of four subsections. We first demonstrate the effectiveness of group sparsity regularization with a synthetically generated example. We then show an application of the column grouping (Figure (1)-(a)) in semi-supervised clustering and an application of the row grouping (Figure 1-(b)) in multilingual text analysis. Finally, we present computational comparisons of the matrix-block and the vector-block BCD methods.

5.1 Factor recovery

Our first demonstration is the comparison of several regularization methods using a synthetically created data set. Figure 2 shows the original data and recovery results. The five original images in the top of Figure 2-(a) are of 32 × 32 pixels, and each of them are vectorized to construct a $1,024 \times 5$ latent component matrix $\mathbf{W}$. Five coefficient matrices $\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(5)}$ of size 5×30 each are constructed by setting the i^{th} row of $\mathbf{H}^{(i)}$ as zero for $i = 1, 2, 3, 4, 5$ and then filling all other entries by taking random numbers from the uniform distribution on $[0, 1]$. The top image of Figure 2-(b) shows the zero and nonzero pattern of $\mathbf{H} = \left(\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(5)}\right) \in 5 \times 120$, where dark entries represent nonzeros and bright entries represent zeros.

The zero rows of each block are clearly shown as bright rows.

We multiplied $\mathbf{W}$ with $\mathbf{H}$ to generate a matrix with five blocks and added Gaussian noise so that the signal-to-noise ratio is 0.3. Under this high noise condition, we tested the ability of various regularization methods in terms of recovering the group structure of the original matrices. Strong noise is common in applications such as video surveillance or Electroencephalography (EEG) analysis in neuroscience. Two alternative regularization methods are considered as competitors:

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W},\mathbf{H}) + \alpha \left\|\mathbf{W}\right\|_F^2 + \beta \left\|\mathbf{H}\right\|_F^2 , \tag{5.17}$$

and

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W},\mathbf{H}) + \alpha \left\|\mathbf{W}\right\|_F^2 + \beta \sum_{j=1}^{n} \left\|\mathbf{h}_{\cdot j}\right\|_1^2 . \tag{5.18}$$

Problems (5.17) and (5.18) impose the Frobenius norm and l_1-norm regularization on $\mathbf{H}$, respectively, and neither of them take the group structure into account. Algorithms for solving (5.17) and (5.18) are described in [17]. For the group sparsity regularization method, we considered (3.2) with $q = \infty$ and $q = 2$. For all cases, parameters α and β need to be provided as input, and we determined them by cross validation: We iterated all possible combinations of $\alpha, \beta \in \left[1, 10^{-1}, \cdots, 10^{-7}\right]$ and chose a pair for which the reconstruction error is the minimum for another data matrix constructed in the same way. For each case of α and β pair, ten random initializations are tried, and the best is chosen.

In Figure 2-(a), it can be seen that the recovered images from the four different regularization methods are visually similar to each other. However, in the coefficient matrices shown in Figure 2-(b), the drawback of conventional regularization methods stands out. In the coefficient matrices recovered by the Frobenius norm or the l_1-norm regularization, the group structure was lost because nonzero (dark) elements appeared in the rows of zero (bright) values that present in the original matrix. In contrast, in the coefficient matrices recovered by the $l_{1,\infty}$-norm or the $l_{1,2}$-norm regularization, the group structure was preserved because the zero (bright) rows remained the same as the original matrix.

The failure to recover the group structure leads to a misinterpretation about the role of latent factors. In original matrices, the first group is constructed only with latent components $\{2, 3, 4, 5\}$, and the second group is constructed with only latent components

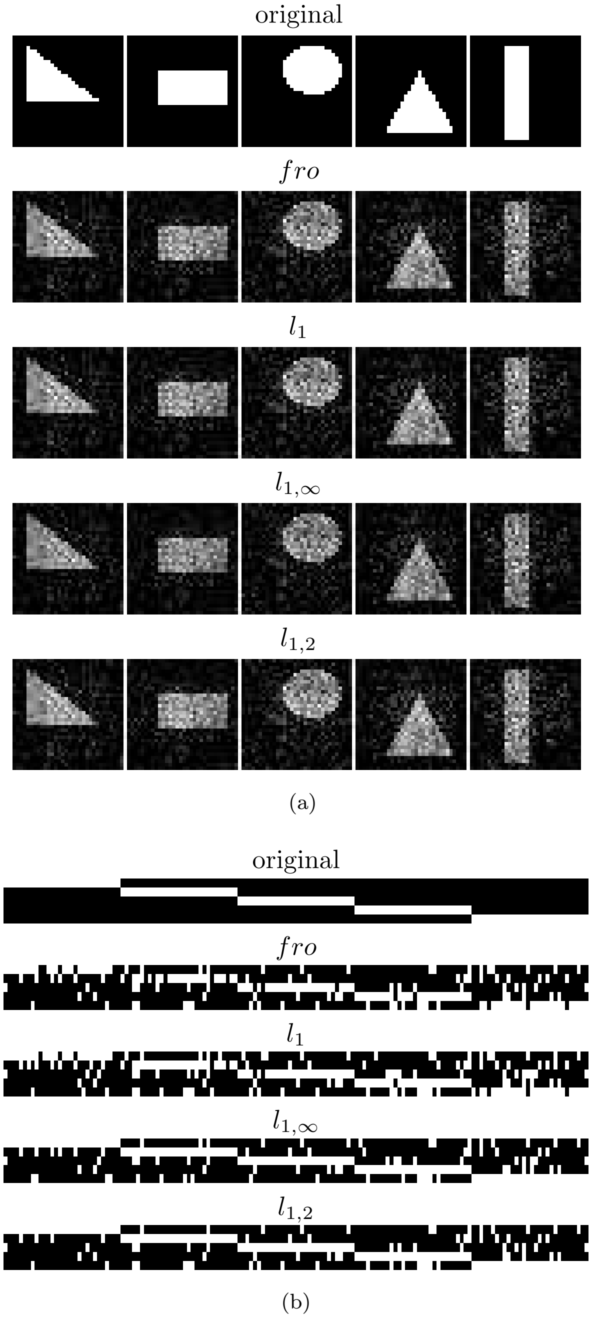


Figure 2: (a) Original latent factor images and recovered factor images with various regularization methods (b) Original coefficient matrix and recovered coefficient matrices with various regularization methods. In each of (a) and (b), first row: original factors, second row: recovered by (5.17), third row: recovered by (5.18), fourth row: recovered by (3.2) with $q = \infty$, fifth row; recovered by (3.2) with $q = 2$. See text for more details.

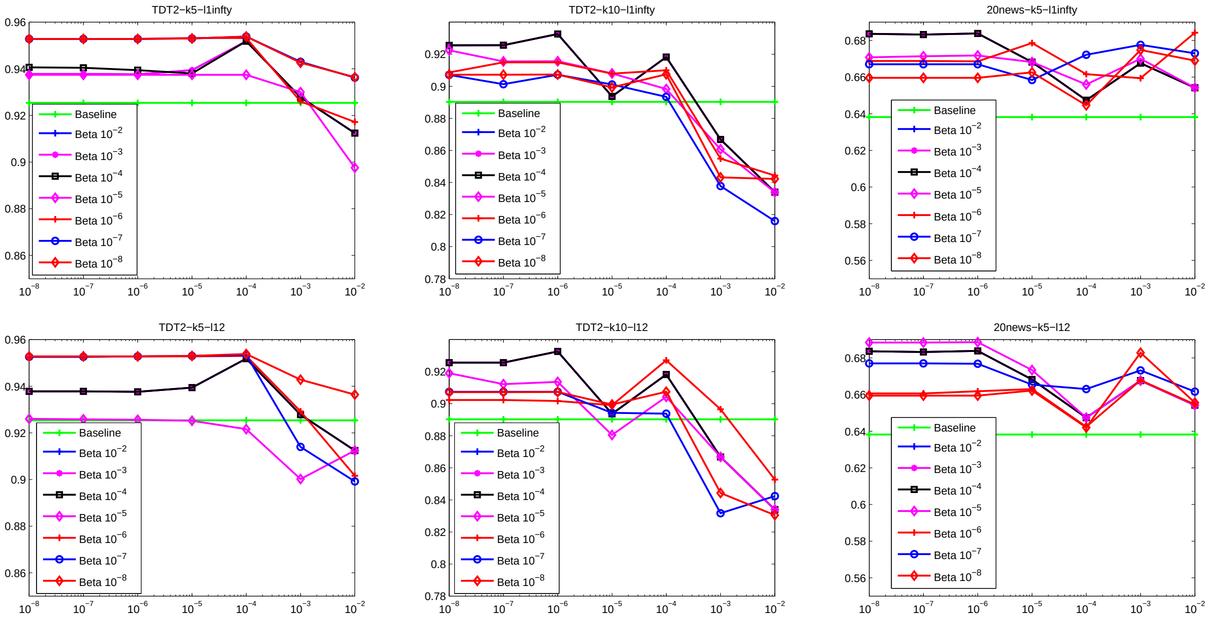


Figure 3: Accuracy of semi-supervised clustering with group sparsity regularization. The x-axis shows the values of α, and the y-axis shows clustering accuracy. Baseline represents the result of no regularization ($\alpha = \beta = 0$). Top: $q = \infty$, bottom: $q = 2$, left: TDT2 data set with $k = 5$, center: TDT2 data set with $k = 10$, right: 20 newsgroups data set with $k = 5$.

$\{1, 3, 4, 5\}$, and so on. However, the coefficient matrices recovered by the Frobenius norm or l_1-norm regularization suggest that all the five factors participate in all the groups, which is an incorrect understanding.

5.2 Semi-supervised clustering Our next demonstration is an application example of the group sparsity regularization with the column groups as shown in Figure 1-(a). One of successful applications of NMF is document clustering, and here we show that the group sparsity regularization can be used to incorporate side-information in clustering.

When NMF is used for clustering (see [34, 15]), after normalizing the columns of $\mathbf{W}$ and rescaling the rows of $\mathbf{H}$ correspondingly, the maximum element from each column of $\mathbf{H}$ is chosen to determine clustering assignments. That is, for a group of documents belonging to the same cluster, their representations in matrix $\mathbf{H}$ are similar to each other in a sense that the positions of elements having the maximum value in each column are the same. In particular, if a group of columns in $\mathbf{H}$ share the same sparsity pattern, it is likely that their clustering assignments are the same. Motivated by this observation, we propose to impose group sparsity regularization for the documents that are supervised to be in the same cluster (i.e., *'must-link'* constraints). In this way, the documents will be promoted to have the same clustering assignments, and latent factor matrix $\mathbf{W}$ will be accordingly adjusted. As a result, the accuracy of clustering assignments for the unsupervised part can be improved.

We tested this task with two text data sets as follows. The Topic Detection and Tracking corpus 2[1] (TDT2) is a collection of English news articles from various sources such as NYT, CNN, and VOA in 1998. The 20 Newsgroups data set[2] is a collection of newsgroup documents in 20 different topics. From term-document matrices constructed from these data sets[3] [6], we randomly selected $k = 5$ (and $k = 10$) topics that contain at least 60 documents each and extracted random subsamples of 60 documents from each topic. Then, 10 documents from each topic were used as a supervised set, and the rest 50 were used an unsupervised (i.e., test) set. That is, we constructed a matrix $\mathbf{X} = \left(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(k)}, \mathbf{X}^{(k+1)}\right)$ where $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(k)}$ represent the supervised parts from each topic and $\mathbf{X}^{(k+1)}$ represents the unsupervised part from all the topics. For the first k groups each having 10 super-

[1] `http://projects.ldc.upenn.edu/TDT2/`

[2] `http://people.csail.mit.edu/jrennie/20Newsgroups/`

[3] `http://www.zjucadcg.cn/dengcai/Data/TextData.html`

vised documents, group sparsity regularization is applied, whereas the last group having total $50 \times k$ unsupervised documents was given no regularization. As a result, we used the following formulation
(5.19)

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} f(\mathbf{W},\mathbf{H}) + \alpha \left\|\mathbf{W}\right\|_F^2 + \beta \sum_{b=1}^{k} \left\|\mathbf{H}^{(b)}\right\|_{1,q},$$

where $\mathbf{H} = \left(\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(k)}, \mathbf{H}^{(k+1)}\right)$. Observe that no regularization is imposed for the last group $\mathbf{H}^{(k+1)}$. The goal is to solve (5.19) and accurately assign clustering labels to the unsupervised part from the final solution of $\mathbf{H}^{(k+1)}$. We selected the most frequent 10,000 terms to reduce the matrix size. We repeated with 10 different random subsamples and evaluated average clustering accuracy with the Hugarian method[4].

The execution results are shown in Figure 3. In semi-supervised clustering, choosing a good parameter setting is difficult because a standard method such as cross validation is not straightforward to apply. Therefore, instead of showing results for specific choices of α and β, we present how the performance of the suggested approach depends on α and β. The results shown in the figures demonstrate a reasonable trend. The group sparsity regularization does boost the clustering performance, but too strong regularization such as $\alpha \geq 10^{-3}$ is often harmful. It can be seen that a wide selection of the parameter values, $\alpha \in [10^{-8}, 10^{-5}]$ and $\beta \in \left[10^{-8}, 10^{-2}\right]$, can be used to improve the clustering accuracy.

Note that the goal of our demonstration is not to argue that the group sparsity regularization is the best semi-supervised clustering approach. Such an investigation requires in-depth consideration on other semi-supervised clustering methods, and it is beyond the scope of this paper. In fact, the group sparsity regularization can be potentially combined with other matrix factorization-based semi-supervised clustering methods [33], and the combination would be an interesting future work. In addition, the group sparsity regularization takes into account only '*must-link*' constraints, and combining with another approach for handling '*cannot-link*' constraints would also be a promising avenue for further study.

5.3 Multilingual text analysis Now, we turn to an application of the group sparsity regularization with the row groups (as apposed to the column groups in the previous subsection). We consider the task of analyzing multilingual text corpus, which is becoming important under the trend of rapidly increasing amount of web text information. Demand for a multilingual text analysis system is particularly high in a nation or a community, such as EU, where multiple official languages are used. An effective approach in multilingual modeling is to make use of parallel (i.e., translated) corpus to discover aligned latent topics. Aligned latent topics can then be used for topic visualization, cross-lingual retrieval, or classification. In this subsection, we show how group-sparsity regularization can be used to improve the interpretation of aligned latent topics.

We have used the DGT Multilingual Translation Memory (DGT-TM)[5] in our analysis. This corpus contains the body of EU law, which is partially translated into 22 languages. We used documents in English, French, German, and Dutch, which will be denoted by EN, FR, DE, and NL, respectively. Applying stop-words and stemmer for each language, we selected the most frequent 10,000 terms in each language to construct term-document matrices. Matrix factorization problem was set up as in Figure 1-(b). As we deal with four languages, the source matrix $\mathbf{X}$ consists of four row blocks: $\mathbf{X}^T = \left(\left(\mathbf{X}^{(1)}\right)^T, \cdots, \left(\mathbf{X}^{(4)}\right)^T\right)$. Columns of these matrices contain the term-document representation of the same document in four different languages. Not all documents are translated into all languages, so the source matrix $\mathbf{X}$ is not fully observed in this case. Missing parts were ignored by treating them with zero weights. Once a low-rank factorization is obtained, the columns of $\mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(4)}$ with the same column index are interpreted as aligned latent topics that convey the same meaning but in different natural languages.

The expected benefit of group sparsity regularization is removing noisy alignments of the latent factors. That is, if a certain topic component appear in documents only in a subset of languages, we would like to detect a zero column in the latent factor for the language where the topic is missing. To test this task, we used a partial corpora from DGT-TM as follows. We collected pairwise translation corpora for EN-FR, EN-DE, and EN-NL (of sizes 1,273, 1,295, and 632, respectively), and appended single language documents in EN, FR, DE, and NL (of sizes 1,300, 930, 610, and 699, respectively). Using $q = \infty$, $k = 500$, $\alpha = 10^{-3}$, and $\beta = 5 \times 10^{-3}$, the algorithm described in Section 4 was applied to $\mathbf{X}^T$. The

[4] http://en.wikipedia.org/wiki/Hungarian_method

[5] http://langtech.jrc.it/DGT-TM.html

Table 1: Summary of topics analyzed by group sparsity regularized NMF.

Id		Keywords
2	EN	member,state,institut,benefit,person,legisl,resid,employ,regul,compet,insur,pension
	FR	procédur,march,de,passat,membr,adjud,recour,d'un,consider,une,aux,concili
	DE	akt,gemeinschaft,rechtsakt,bestimm,europa,leitlini,organ,abfass,dies,sollt,erklar,artikel
	NL	regel,bevoegd,artikel,grondgebied,stat,organ,lid-stat,tijdvak,wettelijk,uitker,werknemer,krachten
14	EN	test,substanc,de,use,en,toxic,prepar,soil,concentr,effect,may,method
	DE	artikel,nr,verordn,flach,eg,absatz,mitgliedstaat,flachenzahl,gemass,erzeug,anhang,wirtschaftsjahr
	NL	word,and,effect,stoff,test,preparat,teststof,stof,la,per,om,kunn
231	EN	brake,shall,vehicl,test,system,point,trailer,control,line,annex,requir,type
	NL	de,moet,voertuig,punt,bijlag,aanhangwag,op,remm,wordt,niet,dor,mag
302	EN	statist,will,develop,polici,european,communiti,programm,inform,data,need,work,requir
	FR	statist,européen,une,polit,programm,un,développ,don,aux,communautair,l'union,mis
	DE	statist,europa,dat,entwickl,programm,information,erford,bereich,neu,dies,gemeinschaft,arbeit
	NL	vor,statistisch,statistiek,europes,word,ontwikkel,over,zull,om,gebied,communautair,programma
392	EN	shall,requir,provid,class,system,space,door,deck,fire,bulkhead,ship,regul
	DE	so,schiff,raum,muss,klass,tur,absatz,vorhand,deck,stell,maschinenraum,regel
452	EN	must,machineri,design,use,oper,safeti,manufactur,risk,requir,construct,direct,person
	NL	moet,machin,zijn,de,dor,om,fabrikant,niet,lidstat,overeenstemm,eis,elk
488	EN	clinic,case,detect,antibodi,isol,compat,diseas,demonstr,specimen,fever,pictur,specif
	FR	détect,cliniqu,une,cas,mis,évident,malad,isol,part,échantillon
	DE	nachweis,klinisch,prob,isolier,bild,vereinbar,fall,spezif,fieb,krankheit,akut,ohn
	NL	klinisch,geval,dor,ziekt,aanton,isolatie,beeld,detectie,monster,bevestigd,niet,teg
494	EN	european,council,schengen,union,treati,visa,decis,articl,provis,nation,protocol,common
	FR	européen,conseil,l'union,vis,décis,présent,trait,schengen,commun,état,communaut,protocol
	DE	europa,rat,union,beschluss,vertrag,gemeinsam,ubereinkomm,artikel,dies,schengen-besitzstand
	NL	de,europes,rad,besluit,overeenkomst,verdrag,protocol,bepal,betreff,lidstat,unie,gemeenschap

columns in **W** are sorted in decreasing amounts of explained variance, and keywords in each topic are listed in a decreasing order of the weights given to each term. The results are summarized in Table 1.

Out of $k = 500$ columns, six of them resulted empty, making the 494th topic the last one in Table 1. Two aspects of the results can be noted as a summary. First, the keywords in each language of the same topic appeared quite well-aligned in general. Second, zero columns indeed were detected in some of the discovered topics. For example, the 231th topic, which is regarding vehicles and trailers, appeared only in English and Dutch documents. Similarly, the 452th topic, which is regarding ships, appeared only in English and German documents. When we tried without group sparsity regularization, however, all the columns of **W** appeared as nonzero.

5.4 Timing comparison Our last experiments are comparisons of Algorithm 1 and Algorithm 3 in terms of computational efficiency. Using data sets from the three previous demonstrations, we executed the two methods and compared time-vs-objective value graphs. In NMF, it is typical to try several initializations, and the execution of one initial value appears as a piecewise-linear decreasing function. We averaged the functions from 10 initializations to generate the plots shown in Figure 4.

From the figure, it can be seen that the vector-block BCD method in Algorithm 3 converges to a minimum faster than the matrix-block BCD method in Algorithm 1. The trend is consistent in both dense (synthetic data set) and sparse (text data sets) matrices. In a non-convex optimization problem such as NMF, each execution may converge to a different local minimum, but the converged minima found by the two methods were in general close to each other.

6 Conclusions and Discussion

In this paper, we proposed mixed-norm regularization methods for promoting group sparsity in NMF. Regularization by $l_{1,q}$-norm successfully promotes that sparsity pattern is shared among data items or features within a group. Efficient convex optimiza-

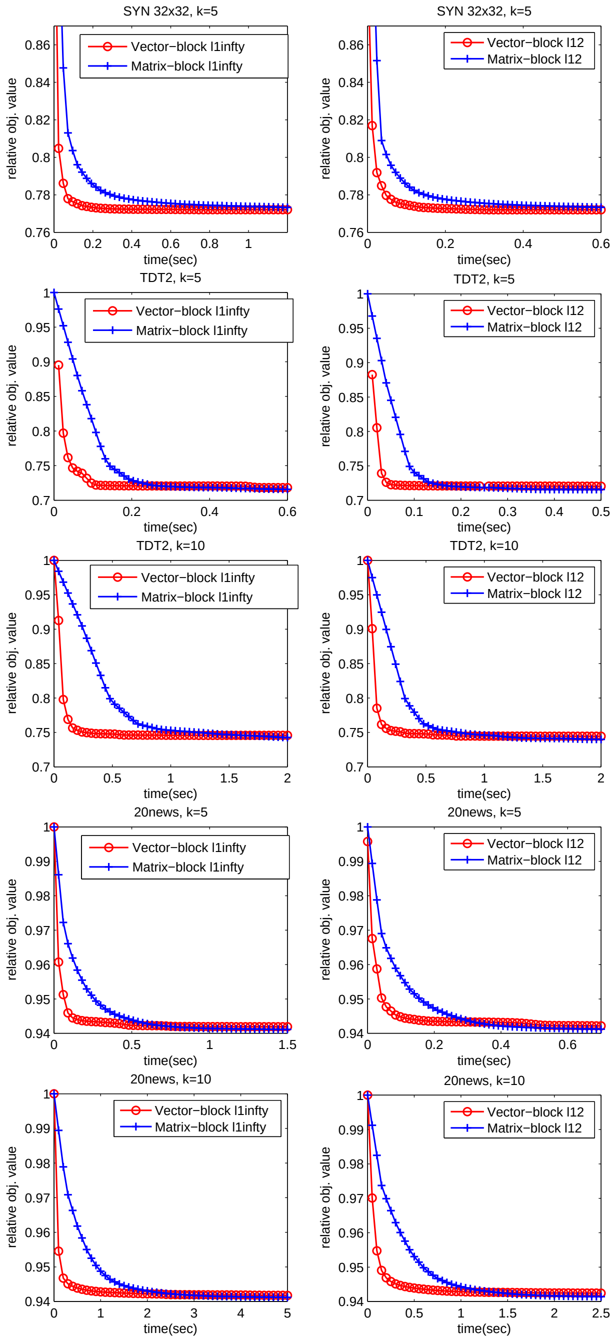


Figure 4: Computational comparisons of the matrix-block and the vector-block BCD methods. The x-axis shows execution time, and the y-axis shows the value of the objective function of (3.1) divided by its evaluation with initial random inputs. All graphs show average results from 10 random initializations. Left: $q = \infty$, right: $q = 2$, first row: synthetic data set used in Section 5.1, second row: TDT2 data set with $k = 5$, third row: TDT2 data set with $k = 10$, fourth row: 20 newsgroup data set with $k = 5$, fifth row: 20 newsgroup data set with $k = 10$.

tion methods based on the block coordinate descent (BCD) method are presented, and the comparisons of them are also provided. Effectiveness of group sparsity regularization is demonstrated with application examples for factor recovery, semi-supervised clustering, and multilingual analysis.

A few interesting directions of future investigation has been learned. First, our study addressed only non-overlapping group structure, and further extending our work to algorithms and applications with overlapping and hierarchical groups will be interesting. In addition, although $l_{1,q}$-norm regularization has been applied to many supervised and unsupervised learning tasks, how its effect depends on q remains to be studied further.

References

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In S. Sra, S. Nowozin, and S. J. Wright., editors, *Optimization for Machine Learning*, pages 19–54. MIT Press, 2011.

[2] L. Badea. Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In *Proceedings of the Pacific Symposium on Biocomputing 2008*, pages 267–278, 2008.

[3] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In *Advances in Neural Information Processing Systems 22*, pages 82–89. 2009.

[4] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[5] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: Theory and Examples.* Springer-Verlag, 2006.

[6] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, December 2005.

[7] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell. Accelerated gradient method for multi-task sparse learning problem. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, pages 746–751, 2009.

[8] A. Cichocki and A.-H. Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E92-A(3):708–721, 2009.

[9] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* Wiley, 2009.

[10] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-ball for

learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 2008.

[11] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning*, pages 487–494, 2010.

[12] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP*, volume 9, pages 366–373, 2010.

[13] Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. In *Advances in Neural Information Processing Systems 23*, pages 982–990. 2010.

[14] D. Kim, S. Sra, and I. S. Dhillon. Fast Newton-type methods for the least squares nonnegative matrix approximation problem. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 343–354, 2007.

[15] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. Technical report, Georgia Institute of Technology Technical Report GT-CSE-08-01, 2008.

[16] J. Kim and H. Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 353–362, 2008.

[17] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[18] S. Kim and E. P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine Learning*, pages 543–550, 2010.

[19] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375–390, 2006.

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

[21] T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Bridging domains with words: Opinion analysis with matrix tri-factorizations. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 293–302, 2010.

[22] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[23] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 649–656, 2009.

[24] J. Liu and J. Ye. Moreau-yosida regularization for grouped tree structure learning. In *Advances in Neural Information Processing Systems 23*, pages 1459–1467. 2010.

[25] M. Masaeli, Y. Yan, Y. Cui, G. Fung, and J. G. Dy. Convex principal feature selection. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 619–628, 2010.

[26] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[27] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.

[28] P. M. Pardalos and N. Kovoor. An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds. *Mathematical Programming*, 46(1-3):321–328, 1990.

[29] A. Quattoni, X. Carreras, M. Collins, and T. Darrell. An efficient projection for l1,infty regularization. In *Proceedings of the 26th International Conference on Machine Learning*, pages 857–864, 2009.

[30] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 650–658, 2008.

[31] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[32] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

[33] F. Wang, T. Li, and C. Zhang. Semi-supervised clustering via matrix factorization. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 1–12, 2008.

[34] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 267–273, 2003.

[35] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.