Graph and Sparse-Based Robust Nonnegative Block Value Decomposition for Clustering

Yaser Esmaeili Salehani, *Member, IEEE*, Ehsan Arabnejad, and Mohamed Cheriet, *Senior Member, IEEE*

***Abstract*—In this paper, we first investigate the nonnegative block value decomposition (NBVD) approach through graph-based representation for clustering called G-NBVD. Then, we propose our three-step graph and sparse-based robust NBVD (GSR-NBVD) via robust NBVD (R-NBVD) framework. The robustness to outliers is obtained by converting the Frobenius norm of error function to the $\ell_{2,1}$-norm for NBVD structure that compensates the effect of samples that are not conforming to NBVD. To exploit the connection between the learning matrix and its corresponding coefficients through sparse representation, we enforce the sparse constraints on the middle matrix in the R-NBVD framework called SR-NBVD. To enhance the geometrical information from data space to the new space, we add a term to our objective minimization function through a regularized graph representation compact form called GSR-NBVD. Then, we prove the convergence of our proposed methods and show a visualization of the effectiveness of G-NBVD and GSR-NBVD step-by-step. Finally, we evaluate our proposed clustering methods over different kinds of data sets. The experimental results confirm that our methods outperforms several state-of-the-art methods through different metrics.**

***Index Terms*—Clustering, graph regularizer, sparse, robustness, $\ell_{2,1}$-norm, nonnegative matrix factorization (NMF), nonnegative block value decomposition (NBVD).**

I. Introduction

CLUSTERING is one of the most powerful technique in data mining for grouping set of objects into different groups through some similarity measures. Clustering methods can be divided into two groups: *hard* and *soft*. While in a hard clustering method (e.g. K-means) a sample is associated to only one cluster, soft clustering methods such as Nonnegative Matrix Factorization (NMF) [1]–[3] associate a sample to more than one cluster with values that show the degree of association to those clusters. On the other hand, there are four common categories for subspace clustering which is an extension of traditional clustering (e.g., principal component analysis (PCA) [4]) that seeks to find clusters in different subspaces within a dataset [5], [6]. They include algebraic algorithms such as matrix factorization-based methods and generalized PCA [7], [8], iterative methods (e.g., K-planes [9] and K-means projective clustering [10]), statistical methods [11], [12] and spectral-based framework such as spectral curvature clustering (SCC) [13], sparse subspace clustering (SSC) [14], [15] and low-rank representation (LRR) based clustering method [16], [17].

NMF framework is initially proposed to obtain a part-based representation of data with non-negativity constraints [2]. These constraints have some physical interpretations (e.g. presence and absence of object) [2], [18]. To solve an NMF-based problem, multiplicative updating (MU) rules have been proposed using Euclidean (Frobenius-norm) and/or Kullback Leibler (KL) divergence [2], [3]. Also, variants of NMF are proposed for various applications in image processing [19]–[21], biological data and document analysis [22]–[27]. These improvements are made by enforcing extra constraints to the objective function such as sparseness [22], [23], [28], smoothness [20], [27], robustness [29]–[31] and Graph regularizations [32], [33]. Indeed, sparse-based constraints is used vastly as an intrinsic property of given data in order to reduce the search space [15], [23], [28]. Graph regularization constraint is used to model the data sample space as a submanifold [33]. Robust NMF (RNMF) is proposed by adding additional term under sparse constraint in [34] and later, by changing the Frobenius norm of errors to $\ell_{2,1}$-norm in [29] in order to mitigate the effect of outliers.

Although the clustering methods such as K-means and 2-factor NMF-based clustering methods [2], [18], [32], [33] exploit the relation between samples or variables separately, co-clustering or bi-clustering methods [35]–[38] utilize the inter relation between samples and variables (features) to group data of rows (data instance) and columns (feature) simultaneously and to find which group of columns maximally corresponds to which group of rows [36], [37]. Hence, we may consider co-clustering structure as a kind of subspace clustering approach which localizes the search in a low-dimension subspace (i.e., different subspace rather than the original data space) in order to find clusters in multiple, possibly overlapping subspaces.

In recent years, different co-clustering methods have been proposed. A co-clustering algorithm through the information theory is proposed in [36]. Then, a method of co-clustering via non-negativity constraints called Block Value Decomposition (NBVD) is proposed in [35]. Later, a non-negative matrix tri-factorization (NMTF) as a 3-factor NMF is proposed in [39] and [40] which considers the orthogonality constraints on both bases and coefficients matrices. The main difference between these two methods is that the updating rules in the latter are

Manuscript received April 21, 2018; revised September 6, 2018 and October 12, 2018; accepted October 14, 2018. Date of publication October 19, 2018; date of current version December 17, 2018. This work was supported by the Natural Sciences and Engineering Research Council of Canada. The guest editor coordinating the review of this paper and approving it for publication was Prof. Thierry Bouwmans. *(Corresponding author: Yaser Esmaeili Salehani.)*

The authors are with the Synchromedia Laboratory, École de technologie supérieure, Montreal, QC H3C 1K3, Canada (e-mail: yaser.esmaeili@gmail.com; ehsan.arabnejad@gmail.com; mohamed.cheriet@etsmtl.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2018.2877041

obtained thorough stiefel manifold. A robust tri-factorization using sparse constraints is proposed in [41] for the cancer genomics application. The robustness to outliers is obtained by adding a term with sparsity constraints in objective function that compensate the effect of samples that are not conforming to NMTF.

Although tri-factorization approach [39], [40] seems to be more appropriate for clustering, because of using orthogonality constraints over bases (first) and coefficient (third) matrices simultaneously, the results in [40] show that NBVD structure performs close or better than tri-factorization methods. Furthermore, we observed that (from extensive simulation results) the NBVD structure is more suitable than NMTF method for our proposed framework since the former has less constrains over the matrices.

Motivated by the above discussions and observations, we first elaborate NBVD framework through the geometrical information of the data space via the nearest neighbor graph structure, called G-NBVD. Then, we introduce a new NBVD objective function along with the graph-based compact matrix and solve its corresponding minimization problem through multiplicative updating rules [3]. Indeed, the graph representation could expose the covered semantics through the intrinsic geometric structure of data similar to GNMF algorithm [33] but with more degree of freedom for three-factorized matrices in our proposed graph-based NBVD framework. In fact, the GNMF method clusters the data through the similarities along the features but the proposed G-NBVD clusters the data by considering the relations between the data samples and features.

Beyond the above described co-clustering approaches, a radically different way to investigate a robust structure to deal with the outliers has not been studied deeply to the best of our knowledge. This direction has been investigated in some details for NMF-based clustering method, e.g. [29], [34]. Thus, to have a further improvement for clustering and to mitigate the outliers, we propose a robust NBVD framework, called R-NBVD. To achieve the robustness, we employ the $\ell_{2,1}$-norm for the penalty function where the large errors (due to outliers) do not dominate the objective function.

To solve its corresponding minimization problem for R-NBVD, we use multiplicative updating rules to calculate the three involved matrices. Then, we prove the convergence of our proposed R-NBVD method. In addition to robustness, the sparse representation via the connector matrix (the middle matrix) so-called block value matrix [35] for a tri-factorization or NBVD framework has not been considered yet. Indeed, we impose the $\ell_{\frac{1}{2}}$-norm constraint [19], [21] on the connection (middle) matrix and we propose the sparse-based robust NBVD called SR-NBVD. The convergence of our proposed SR-NBVD method is also proved. Motivated by these two main achievements (robustness and sparseness), we will consider aforementioned graph-based representation matrix which boost the clustering results. Hence, in the final stage, we add the graph regularization constraints to SR-NBVD model in order to investigate the impact of using the geometric structure of data and propose a joint graph and SR-NBVD framework called GSR-NBVD. Then, we determine the updating rules for matrices and show the proof of convergence for GSR-NBVD. Furthermore, we visualize the improvements of our proposed methods for clustering using a simple data set in each stage.

The main contributions of the current work are summarized as follows:

- We present a novel graph regularization-based NBVD framework (GNBVD) which considers the geometric structure information contained in both data points and features simultaneously as a co-clustering approach.
- We also propose a robust graph-based NBVD by converting the Frobenius norm of errors to the $\ell_{2,1}$-norm under the sparse constraint of the middle matrix of three-factorized matrices (GSR-NBVD) in which improves the clustering results.
- We develop multiplicative updating rules to solve the corresponding optimization schemes of proposed G-NBVD and GSR-NBVD methods, and provide the convergence proofs of two minimization problems.

The remainder of the paper is organized as follows. In Section II, NMF and its variants are reviewed. The NBVD approach is briefly introduced in Section II. Then, we present our clustering methods in Section III including G-NBVD and GSR-NBD through three main steps: robust NBVD, sparse R-NBVD and graph SR-NBVD. We evaluate our proposed methods over different types of real-world data sets and compare their results with several state-of-the-art methods in Section IV. Section V concludes the paper and suggests paths for future plan/research.

II. NNMFs: NonNegative Matrix Factorizations

In this section, we first review the NMF based family methods and their solutions through the multiplicative updating rules approach. Then, the NBVD method is reviewed.

A. NMF and Its Variants

We start by reviewing the nonnegative matrix factorization (NMF) [2]. The goal is to decompose a given data matrix $\mathbf{X} \in \mathbb{R}^{L\times P}$ into two nonnegative matrices $\mathbf{U} \in \mathbb{R}_{+}^{L\times K}$ and $\mathbf{V} \in \mathbb{R}_{+}^{P\times K}$ such that $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$. The NMF problem is defined by

$$\underline{P_{NMF}} : \min_{\mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 \tag{1}$$

where $||.||_F$ denotes the Frobenius norm. The multiplicative updating rules has been proposed to solve (1) as follows [3]

$$\mathbf{U} \leftarrow \mathbf{U}. * \mathbf{X}\mathbf{V}./\mathbf{U}\mathbf{V}^T\mathbf{V} \tag{2}$$

$$\mathbf{V} \leftarrow \mathbf{V}. * \mathbf{X}^T\mathbf{U}./\mathbf{V}\mathbf{U}^T\mathbf{U} \tag{3}$$

where $.*$ and $./$ are the element-wise matrix multiplication and division, respectively. It can be shown that the objective function of $\underline{P_{NMF}}$ in (1) is nonincreasing under the update rules in (2) and (3) [3].

To consider the sparse property of the coefficient matrix $\mathbf{V}$, the ℓ_0-norm regularizer term might be added to the objective

function in (1) as follows

$$\underline{P_{SNMF}}: \min_{\mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X}-\mathbf{U}\mathbf{V}^T||_F^2+\lambda_s||\mathbf{V}||_0 \quad (4)$$

where $\lambda_s>0$ is the Lagrangian (regularizer) parameter and $||.||_0$ is the ℓ_0-norm function defined by finding the number of non-zero components of its given vector. Since the ℓ_0-norm problem is NP-hard because of the combinatorial exhaustive search of its solution, various approximations such as ℓ_p-norm $(0<p\leq 1)$ variants (e.g., ℓ_1-norm [28] and $\ell_{1/2}$-norm [19]) and $\arctan$ function [20] were proposed to tackle this problem. For instance, several sparsity constrained NMF methods have been proposed for the application of hyperspectral unmixing [19], [21], [32] and use the following $\ell_{1/2}$-norm term in their objective functions

$$\min_{\mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} \frac{1}{2}||\mathbf{X}-\mathbf{U}\mathbf{V}^T||_F^2+\lambda_s||\mathbf{V}||_{\frac{1}{2}}, \quad (5)$$

where the $\ell_{1/2}$-norm is defined as

$$||\mathbf{Z}||_{\frac{1}{2}}=\sum_i\sum_j |z_{ij}|^{\frac{1}{2}}, \quad (6)$$

with the appropriate dimensions of matrix $\mathbf{Z}$. Using the multiplicative updating rules, the values of $\mathbf{U}$ is updated similar to (2) and the value of $\mathbf{V}$ is updated as follows

$$\mathbf{V}\leftarrow \mathbf{V}.*\mathbf{X}^T\mathbf{U}./\left(\mathbf{V}\mathbf{U}^T\mathbf{U}+\frac{\lambda_s}{2}\mathbf{V}^{-\frac{1}{2}}\right) \quad (7)$$

where $[\cdot]^{-\frac{1}{2}}$ is the power of $-\frac{1}{2}$ element-wisely. In [19], it has been shown that the objective function in (5) is nonincreasing under the update rules in (2) and (7).

In addition to the sparse properties of coefficients matrix for the clustering purpose, the intrinsic geometric structure of data is also helpful for data clustering. Hence, the Graph Regularized Non-negative Matrix Factorization (GNMF) [33] is proposed as the following problem

$$\underline{P_{GNMF}}: \min_{\mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X}-\mathbf{U}\mathbf{V}^T||_F^2+\lambda_g \mathrm{Tr}(\mathbf{V}^T\mathbf{B}\mathbf{V}) \quad (8)$$

where $\lambda_g>0$ is the graph regularizer parameter, Tr(.) denotes the trace of a matrix, $\mathbf{B}=\mathbf{A}-\mathbf{W}$ is the graph Laplacian matrix with the weighting matrix $\mathbf{W}$ (i.e. binary, heat kernel, or dot-product, see [33] for more details) and the diagonal matrix $\mathbf{A}$ whose entries are column (or row) sums of $\mathbf{W}$, $\mathbf{A}_{ij}=\sum_l \mathbf{W}_{jl}$. Again, using the multiplicative updating rules, the values of $\mathbf{U}$ is updated similar to (2) and $\mathbf{V}$ is updated as follows

$$\mathbf{V}\leftarrow \quad \mathbf{V}.*(\mathbf{X}^T\mathbf{U}+\lambda_g\mathbf{W}\mathbf{V})./(\mathbf{V}\mathbf{U}^T\mathbf{U}+\lambda_g\mathbf{A}\mathbf{V}) \quad (9)$$

One can prove that the objective function of $\underline{P_{GNMF}}$ in (8) is nonincreasing under the update rules in (2) and (9) [33].

To mitigate outliers and noises in an appropriate model, the following robust $\ell_{2,1}$-norm approach is proposed in [29]

$$\underline{P_{RNMF}}: \min_{\mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X}-\mathbf{U}\mathbf{V}^T||_{2,1} \quad (10)$$

where $\ell_{2,1}$-norm of a matrix $\mathbf{Z}$ is defined as

$$||\mathbf{Z}||_{2,1}=\sum_{i=1}^{P}||z_i||=\sum_{i=1}^{P}\sqrt{\sum_{j=1}^{L}\mathbf{Z}_{ji}^2} \quad (11)$$

To solve the robust NMF problem in (10), the following multiplicative updating rules is proposed in [29]

$$\mathbf{U}\leftarrow \mathbf{U}.*\mathbf{X}\hat{\mathbf{D}}\mathbf{V}./\mathbf{U}\mathbf{V}^T\hat{\mathbf{D}}\mathbf{V} \quad (12)$$

$$\mathbf{V}\leftarrow \mathbf{V}.*\hat{\mathbf{D}}\mathbf{X}^T\mathbf{U}./\hat{\mathbf{D}}\mathbf{V}\mathbf{U}^T\mathbf{U} \quad (13)$$

where the diagonal matrix $\hat{\mathbf{D}}$ is defined with the following diagonal elements for $i\in\{1,\ldots,P\}$:

$$\hat{\mathbf{D}}_{ii}=\frac{1}{||x_i-\mathbf{U}v_i^T||}=\frac{1}{\sqrt{\sum_{j=1}^{L}(\mathbf{X}-\mathbf{U}\mathbf{V}^T)_{ji}^2}} \quad (14)$$

It can be shown that the objective function of $\underline{P_{RNMF}}$ in (10) is nonincreasing under the update rules in (12) and (13) [29]. The convergence theorems as well as their supporting lemmas are given in [29] for more details.

B. NBVD: Non-Negative Block Value Decomposition

Let $\mathbf{X}\in\mathbb{R}^{L\times P}$, $\mathbf{U}\in\mathbb{R}_+^{L\times K}$, $\mathbf{S}\in\mathbb{R}_+^{K\times N}$ and $\mathbf{V}\in\mathbb{R}_+^{P\times N}$. Then, the NBVD problem [35] is defined as follows

$$\underline{P_{NBVD}}: \min_{\mathbf{U}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T||_F^2 \quad (15)$$

The following updating rules are introduced in [35] to solve $\underline{P_{NBVD}}$ in (15)

$$\mathbf{U}\leftarrow \mathbf{U}.*\mathbf{X}\mathbf{V}\mathbf{S}^T./\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T \quad (16)$$

$$\mathbf{S}\leftarrow \mathbf{S}.*\mathbf{U}^T\mathbf{X}\mathbf{V}./\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V} \quad (17)$$

$$\mathbf{V}\leftarrow \mathbf{V}.*\mathbf{X}^T\mathbf{U}\mathbf{S}./\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S} \quad (18)$$

It has been proved that the objective function of $\underline{P_{NBVD}}$ in (15) is nonincreasing under the update rules in (16), (17) and (18) [35] where the steps of theorem for finding local minimizer of (15) is also given in [35].

III. G-NBVD AND GSR-NBVD: OUR PROPOSED METHODS FOR CLUSTERING

Motivated by NBVD framework (in spite of tri-factorization [39], [40] that the orthogonality constraints are imposed for the learning and coefficient matrices), we first propose a graph based NBVD method in Section III-A which improves significantly the clustering results of NBVD [35].

To further improve the clustering results for G-NBVD, we propose our regularized graph SR-NBVD through three main steps: i) robust NBVD through the $\ell_{2,1}$-norm model in Section III-B1 ii) sparse R-NBVD which adds the sparseness of the middle matrix in Section III-B2 iii) regularized graph SR-NBVD in Section III-B3.

Algorithm 1: Pseudocode of the Graph Based Nonnegative Block Value Decomposition (G-NBVD) Algorithm.

Input: data matrix $\mathbf{X}$ and parameters λ_g, ϵ, $I_{\max}$.
- Initialize matrices $\mathbf{U}, \mathbf{S}$ and $\mathbf{V}$.

Repeat:
- Compute $C_{\text{old}} = C(\mathbf{X}, \mathbf{U}, \mathbf{S}, \mathbf{V})$ using the objective function in (19).
- Update $\mathbf{U}$ using (16).
- Update $\mathbf{S}$ using (17).
- Update $\mathbf{V}$ using (20).
- Compute $C_{\text{new}} = C(\mathbf{X}, \mathbf{U}, \mathbf{S}, \mathbf{V})$ using the objective function in (19).
- Stop if either $|C_{\text{new}} - C_{\text{old}}|/C_{\text{old}} < \epsilon$ or the iteration number exceeds $I_{\max}$.

Output: The feature matrix $\mathbf{U}$, the coefficient matrix $\mathbf{V}$ and matrix $\mathbf{S}$.

A. G-NBVD: Graph Based NBVD Method

Inspired by GNMF clustering approach [33] which amplifies the clustering results via the intrinsic geometric properties of data in comparison with other NMF-based family clustering methods, we first investigate this property over NBVD framework. In a data representation through the graph-based matrix factorization, we can model the local similarity of data points using a nearest neighbor graph over data samples. It is based on the fact that if two data points are neighbors in original space, this relation should also be preserved by NMF in the projected space. We can use the nearest neighbors for each data sample (which we can consider each sample as a vertex of a graph) in order to construct the weight matrix. Indeed, we take advantage of geometrical structure of data space by finding a part-based representation space in which improves significantly the clustering results as mentioned later. Now, we introduce the following minimization problem for G-NBVD

$$\underline{P_{GNBVD}} : \min_{\mathbf{U}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X} - \mathbf{USV}^T||_F + \lambda_g \text{Tr}(\mathbf{V}^T\mathbf{BV}) \tag{19}$$

where $\lambda_g > 0$ is the graph regularizer parameter and $\mathbf{B}$ is the graph Laplacian matrix defined earlier in the problem of $\underline{P_{GNMF}}$ in (8). Our proposed G-NBVD algorithm constructs a nearest neighbor graph to make the manifold structure in which the weight matrix $\mathbf{B}$ is highly sparse. Hence, we may use the multiplicative updating rules to solve $\underline{P_{GNBVD}}$ which is very efficient. The values of $\mathbf{U}$ and $\mathbf{S}$ are updated as in (16) and (17), respectively. The values of $\mathbf{V}$ is updated as follows

$$\mathbf{V} \leftarrow \mathbf{V}. * \left(\mathbf{X}^T\mathbf{US} + \lambda_g\mathbf{WV}\right)./\left(\mathbf{VS}^T\mathbf{U}^T\mathbf{US} + \lambda_g\mathbf{AV}\right) \tag{20}$$

Algorithm 1 summarizes our proposed G-NBVD method.

B. GSR-NBVD: Graph Regularizer for Sparse-Based Robust NBVD

In this section, we propose our GSR-NBVD method through the following three steps.

1) R-NBVD: Robust NBVD Method: To mitigate the impacts of outliers and noises, we propose the following robust NBVD problem through $\ell_{2,1}$-norm approach

$$\underline{P_{RNBVD}} : \min_{\mathbf{U}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X} - \mathbf{USV}^T||_{2,1} \tag{21}$$

with the same dimensions of matrices described in $\underline{P_{NBVD}}$.

To solve the above problem (21), we use the multiplicative updating rules [3] and obtain the following updates rules

$$\mathbf{U} \leftarrow \mathbf{U}. * \mathbf{XDVS}^T./\mathbf{USV}^T\mathbf{DVS}^T \tag{22}$$

$$\mathbf{S} \leftarrow \mathbf{S}. * \mathbf{U}^T\mathbf{XDV}./\mathbf{U}^T\mathbf{USV}^T\mathbf{DV} \tag{23}$$

$$\mathbf{V} \leftarrow \mathbf{V}. * \mathbf{DX}^T\mathbf{US}./\mathbf{DVS}^T\mathbf{U}^T\mathbf{US} \tag{24}$$

where the diagonal matrix $\mathbf{D}$ is defined by

$$\mathbf{D}_{ii} = \frac{1}{\sqrt{\sum_{j=1}^{L}(\mathbf{X} - \mathbf{USV}^T)_{ji}^2}}, i = 1, \ldots, P \tag{25}$$

or equivalently

$$\mathbf{D}_{ii} = \frac{1}{||\mathbf{X}_{:,i} - \mathbf{USV}_{:,i}^T||}, i = 1, \ldots, P \tag{26}$$

where $\mathbf{Z}_{:,i}$ denotes the i-th column of matrix $\mathbf{Z}$.

2) SR-NBVD: Sparse-Based Robust NBVD: To address the sparse property of data, we enforce the sparse constraint on the middle matrix $\mathbf{S}$ through the robust NBVD approach in order to connect the feature matrix $\mathbf{U}$ to the coefficient matrix $\mathbf{V}$ efficiently and propose the following sparse-based Robust NBVD, called SR-NBVD problem,

$$\underline{P_{SRNBVD}} : \min_{\mathbf{U}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} ||\mathbf{X} - \mathbf{USV}^T||_{2,1} + \lambda_s||\mathbf{S}||_{\frac{1}{2}} \tag{27}$$

where $\lambda_s > 0$ denotes the Lagrangian parameter and $||\mathbf{S}||_{\frac{1}{2}}$ is defined as in (6).

We solve the $\underline{P_{SRNBVD}}$ problem in (27) using the multiplicative updating rules where the values of $\mathbf{U}$ and $\mathbf{V}$ are updated as in (22) and (24) respectively, and $\mathbf{S}$ is updated as follows

$$\mathbf{S} \leftarrow \mathbf{S}. * \mathbf{U}^T\mathbf{XDV}./\left(\mathbf{U}^T\mathbf{USV}^T\mathbf{DV} + \frac{\lambda_s}{2}\mathbf{S}^{-\frac{1}{2}}\right) \tag{28}$$

Indeed, adding the sparse constraint on the middle matrix of the proposed model showed the better clustering performances (i.e., in the most cases) during the extensive simulations over different data sets. We left theoretical discussion of how we may connect the sparseness of the middle matrix with the clustering performance as well as the structure of the possible outliers for our future work.

3) GSR-NBVD: A Graph Regularizer for Sparse-Based Robust NBVD: We enhance our proposed sparse-based robust NBVD method (SR-NBVD) mentioned in Section III-B2 by preserving the graph structure of data which explained earlier in Section III-A. To model our jointly Graph and Sparse-based Robust NBVD, called GSR-NBVD, we propose the following minimization problem

$$\underline{P_{GSRNBVD}} : \min_{\mathbf{U}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}} C(\mathbf{X}, \mathbf{U}, \mathbf{S}, \mathbf{V}) \tag{29}$$

Algorithm 2: Pseudocode of the Joint Graph and Sparse-based Robust Nonnegative Block Value Decomposition Method (GSR-NBVD).

Input: data matrix $\mathbf{X}$ and parameters λ_s, λ_g, ϵ, $I_{\max}$.
- Initialize $\mathbf{U}, \mathbf{S}$ and $\mathbf{V}$.
Repeat:
- Compute $C_{\text{old}} = C(\mathbf{X}, \mathbf{U}, \mathbf{S}, \mathbf{V})$ using (30).
- Update $\mathbf{D}$ using (25).
- Update $\mathbf{U}$ using (22).
- Update $\mathbf{S}$ using (28).
- Update $\mathbf{V}$ using (31).
- Compute $C_{\text{new}} = C(\mathbf{X}, \mathbf{U}, \mathbf{S}, \mathbf{V})$ using (30).
- Stop if either $|C_{\text{new}} - C_{\text{old}}| / C_{\text{old}} < \epsilon$ or the iteration number exceeds $I_{\max}$.
Output: The feature matrix $\mathbf{U}$, the coefficient matrix $\mathbf{V}$ and matrix $\mathbf{S}$.

where its cost function is defined as follows

$$C(\mathbf{X}, \mathbf{U}, \mathbf{S}, \mathbf{V}) = ||\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T||_{2,1} + \lambda_s ||\mathbf{S}||_{\frac{1}{2}} + \lambda_g \text{Tr}(\mathbf{V}^T \mathbf{B} \mathbf{V}) \quad (30)$$

where $\lambda_s > 0$ and $\lambda_g > 0$ are the constant parameters for the sparse and graph regularization terms respectively and $\mathbf{B}$ is the graph Laplacian matrix defined earlier in the problem of $\underline{P_{GNMF}}$ in (8).

The proposed problem is a general structure that can leverage the power of sparse-based robust NBVD (SR-NBVD) and graph Laplacian regularization which leads to more degree of freedom compared to the GNMF algorithm [33] and considers the relations between the data points and features. Our proposed GSR-NBVD can have more discriminating power than the NBVD by preserving the graph structure and robustness property through the $\ell_{2,1}$-norm of modeling error.

To solve $\underline{P_{GSRNBVD}}$, we use the multiplicative updating rules in which the values of $\mathbf{U}$ and $\mathbf{S}$ are updated as in (22) and (28), respectively and the values of $\mathbf{V}$ is updated as follows

$$\mathbf{V} \leftarrow \mathbf{V}. * \big(\mathbf{D}\mathbf{X}^T \mathbf{U}\mathbf{S} + \lambda_g \mathbf{W}\mathbf{V}\big)./\big(\mathbf{D}\mathbf{V}\mathbf{S}^T \mathbf{U}^T \mathbf{U}\mathbf{S} + \lambda_g \mathbf{A}\mathbf{V}\big) \quad (31)$$

Our proposed GSR-NBVD method for data clustering is summarized in Algorithm 2.

C. Convergence

In the first part, we show the convergence of our proposed G-NBVD method that used the updating rules in (16), (17) and (20) as follows.

Theorem 1: (i) The objective function of $\underline{P_{GNBVD}}$ in (19) using updating $\mathbf{U}$ in (16) monotonically decreases while fixing $\mathbf{S}$ and $\mathbf{V}$. (ii) Updating $\mathbf{S}$ using the rule of (17) while fixing $\mathbf{U}$ and $\mathbf{V}$, the objective function of (19) monotonically decreases. (iii) Updating $\mathbf{V}$ using the rule of (20) while fixing $\mathbf{U}$ and $\mathbf{S}$, the objective function of (19) monotonically decreases.

Proof: See Appendix A for the proof. ■

Then, the following theorem shows the convergence of R-NBVD based on updating rules in (22), (23) and (24).

Theorem 2: (i) The objective function of $\underline{P_{RNBVD}}$ in (21) using updating $\mathbf{U}$ in (22) monotonically decreases while fixing $\mathbf{S}$ and $\mathbf{V}$. (ii) Updating $\mathbf{S}$ using the rule of (23) while fixing $\mathbf{U}$ and $\mathbf{V}$, the objective function of (21) monotonically decreases. (iii) Updating $\mathbf{V}$ using the rule of (24) while fixing $\mathbf{U}$ and $\mathbf{S}$, the objective function of (21) monotonically decreases.

Proof: See Appendix B for the proof. ■

Afterwards, we show the convergence proposed SR-NBVD method thorough the updating rules in (22), (28) and (24).

Theorem 3: (i) The objective function in (27) using updating $\mathbf{U}$ in (22) monotonically decreases while fixing $\mathbf{S}$ and $\mathbf{V}$. (ii) The objective function in (27) monotonically decreases using updating $\mathbf{S}$ in (28) while fixing $\mathbf{U}$ and $\mathbf{V}$. (iii) Updating $\mathbf{V}$ using the rule of (24) while fixing $\mathbf{U}$ and $\mathbf{S}$, the objective function of (27) monotonically decreases.

Proof: See Appendix C for the proof. ■

Finally, the following theorem shows the convergence of our proposed GSR-NBVD method based on updating rules in (22), (28) and (31).

Theorem 4: (i) The objective function of $\underline{P_{GSRNBVD}}$ in (29) using updating $\mathbf{U}$ in (22) monotonically decreases while fixing $\mathbf{S}$ and $\mathbf{V}$. (ii) Updating $\mathbf{S}$ using the rule of (28) while fixing $\mathbf{U}$ and $\mathbf{V}$, the objective function of (29) monotonically decreases. (iii) Updating $\mathbf{V}$ using the rule of (31) while fixing $\mathbf{U}$ and $\mathbf{S}$, the objective function of (29) monotonically decreases.

Proof: See Appendix D for the proof. ■

We must note that we replace zero values of matrices $\mathbf{U}, \mathbf{S}$ and $\mathbf{V}$ with a small positive value during the iterations to avoid the zero-lock problem similar to one used in [42]. Also we must mention that the convergence proofs show that the involved constraints are satisfied and the (existing) optimal solution have the sparseness and robustness properties through the graph structure and $\ell_{2,1}$-norm terms. We left the direct proof of robustness property terms against outliers as our future work.

D. Complexity

The computational cost is an important issue for the clustering, specifically for the large size of data samples with high dimensions. In this subsection, we count the arithmetic operations per each iteration for the updating rules of NBVD and our proposed methods of G-NBVD and GSR-NBVD. These operations include three major floating-point-operations (FLOPs): addition, multiplication, division and square-root and its inverse (i.e. for calculating $(.)^{\frac{1}{2}}$ and $(.)^{-\frac{1}{2}}$).

We must note that the most of block value decomposition updates of the proposed framework are identical or similar to the corresponding NBVD model [35] and GNMF structure [33] except computing matrix $\mathbf{D}$. However, it can be computed efficiently since it is diagonal and pretty sparse in the implementation. Therefore, the overall complexity follows the basic NBVD model in [35].

Table I compares the computational costs of our two main proposed methods with NMF and GNMF in terms of FLOPs and the big O notation.

TABLE I
COMPUTATIONAL OPERATION COUNTS FOR EACH ITERATION, IN TERMS OF A FLOATING-POINT-OPERATION (FLOP) AND SQUARE-ROOT AND ITS INVERSE, IN NMF, GNMF, NBVD AND OUR PROPOSED METHODS

	FLOP (+)	FLOP (×)	FLOP (÷)	SQR/SQR^{-1}	Overall
NMF [3]	$2PLK + 2(L+P)K^2$	$2PLK + 2(L+P)K^2 + (L+P)K$	$(L+P)K$	-	$O(PLK)$
GNMF [33]	$2PLK + 2(L+P)K^2 + P(q+3)K$	$2PLK + 2(L+P)K^2 + (L+P)K + P(q+1)K$	$(L+P)K$	-	$O(PLK)$
NBVD [35]	$PL(2K+3N) + 5LNK + 2PN(K+N) + L(K^2+N^2) + KN(K+N) + KN$	$PL(2K+3N) + LN(5K+N) + 2PN(K+N) + LK^2 + KN(K+N+1) + LK$	$PN + LK + KN$	-	$O(PLK)$
G-NBVD	$PL(2K+3N) + 5LNK + PN(2K+2N+q+2) + L(K^2+N^2) + KN(K+N) + KN$	$PL(2K+3N) + LN(5K+N) + PN(2K+2N+1) + LK^2 + KN(K+N+1) + LK + N$	$PN + LK + KN$	-	$O(PLK)$
GSR-NBVD	$2PL(K+2N+1) + LN(5K+N) + PN(2K+2N+q+2) + LK^2 + KN(2K+N+1)$	$2PL(K+2N+1) + LN(6K+N) + PN(2K+2N+q+5) + N^2K + K(L+N)$	$PN + LK + KN$	$P + KN$	$O(PLK)$

• The value of q comes from the Graph-based structure implementation that used a q-nearest neighbor graph to construct the sparse matrix $\mathbf{W}$, see [33] for more details.
• The first two rows of table is recomputed which is compatible exactly with the computations in [33].
• The overall complexity of the last three methods is the order of LPK or LPN where the value of K and N is the same or a factor (by 2 or 0.5) of each other through our extensive simulations for setting the dimension of $\mathbf{S}$.

IV. EXPERIMENTS

A. Parameter Selection and Initialization

Our proposed GSR-NBVD framework has two types of parameters: the dimensions values K and N, the regularization parameters for sparsity (λ_s) and graph regularizer (λ_g). There is not a unique prescription to select those parameters though some methods have been studied for choosing regularizer parameters [43]–[45] and dimensions [33], [35], [46].

In this paper, we propose to choose K and N as the factors of clusters' number (by assuming that it is known/given) as follows

$$K = \alpha_K \times \gamma \tag{32}$$

$$N = \alpha_N \times \gamma \tag{33}$$

where α_K and α_N are empirically obtained (depending the types of data sets) and γ is the number of clusters. We selected $\alpha_K = \alpha_N = 1$ for biological data set where this number is the same as used in [35], $\alpha_K = \alpha_N = 2$ for text data sets and $\alpha_N = 2\alpha_K = 3$ for image data set throughout our experiments. We choose the values of λ_s and λ_g by cross-validation in which the majority of four metrics are maximized. Those values are in the range of $(0, 10]$ and $(0, 10^5]$ for λ_s and λ_g, respectively, for all types of data sets.

To initialize our methods, we may use random initialization with standard normal distribution where maximized by 0.01 or the similar approach introduced in [29]. In our experiments, we use the similar approach proposed for robust $\ell_{2,1}$-norm NMF method in [29] for the initializations of matrices $\mathbf{U}, \mathbf{S}$ and $\mathbf{V}$. First, a projection of original data set is produced by PCA method with the dimension of $2 \times K$. Then, we employ K-means clustering method on the projected data to achieve the clustering results, say $\tilde{\mathbf{V}}$, and initialize $\mathbf{V}$ with $\tilde{\mathbf{V}} + c_1$ where c_1 is the first constant value. Afterwards, we obtain $\mathbf{US}$ matrix by computing the cluster center for each category. Again, we apply K-means clustering method on the obtained $\mathbf{US}$ matrix, and initialize $\mathbf{U}$ with $\tilde{\mathbf{U}} + c_2$ where c_2 is the second constant value. Finally, $\mathbf{S}$ is initialized by computing the clustering centroid for each category of resulted $\mathbf{US}$ matrix. Empirically, we run K-means 20 times during initializations.

We implement all of these methods over MATLAB platform, on an Intel Core i7-4790 (at 3.6 GHz) and 16 GB of RAM.

B. Clustering Visualization

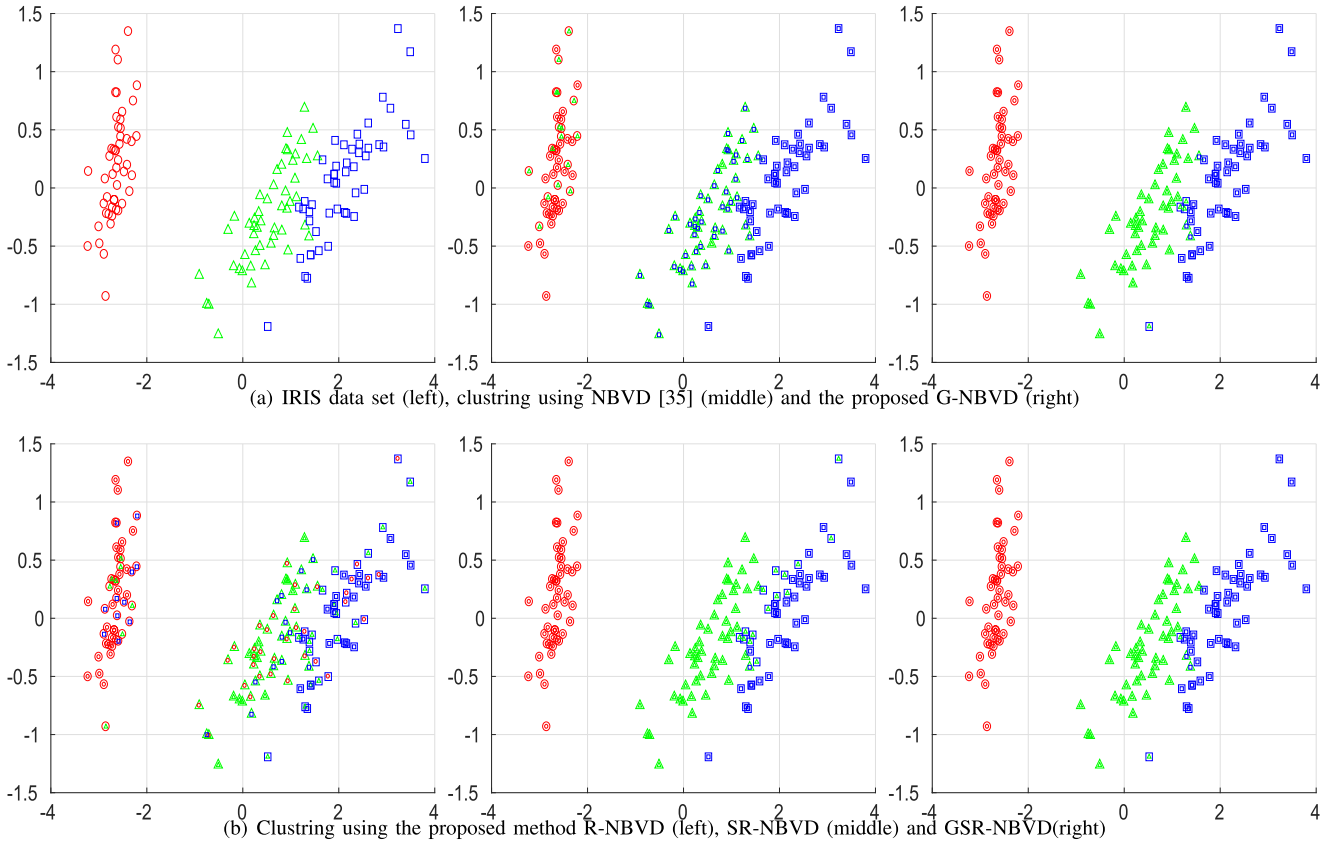
To visualize the effectiveness of our proposed G-NBVD and three-step GSR-NBVD methods, we employ the IRIS dataset [47], [48]. It contains 4 measures of 3 classes and 50 samples per class. We use PCA [49] to project data to 2-D space as shown in Figure 1(a)-left. The distribution of three classes are displayed via different colors and shapes of markers. We apply NBVD [35], the proposed G-NBVD and GSR-NBVD, step-by-step, to the original IRIS data set and use K-means to obtain the cluster index for each sample. The matched grouped classes are shown with the same color and shapes (i.e. with larger markers).

We must note that for the clustering purpose using NBVD and NMF families throughout the paper, we apply K-means clustering to the representation of data obtained by those methods.

As shown in Figure 1, we observe that G-NBVD and GSR-NBVD methods group three different classes with the highest accuracy (Figure 1(a)-right and (b)-right) compared with NBVD shown in Figure 1(a)-middle. Furthermore, we observe that the improvements of clustering during three steps of GSR-NBVD approach as depicted in Figure 1(b), from left to right.

C. Robustness to Outliers

To verify the effectiveness of our proposed G-NBVD and GSR-NBVD methods in the presence of outliers and corruptions, we generate a dataset by combining the Columbia Object Image Library (COIL-20) [50] which contains 20 objects of 32×32 gray scale images and a number of face images selected randomly from the Extended Yale Database [51]. We add different number of face images (i.e., outliers) to COIL-20 in the range of 5% to 40% of the number of samples in the original data set (i.e., 1440 samples) by step of 1%. We call the added images as outliers and consider them as outlier category. Thus, the corrupted data set has the number of (1 + (5% to 40%)) × 1440 samples with 21 clusters in total. In other words, we produce 36 noisy data sets with outliers where each contains a specified



(a) IRIS data set (left), clustring using NBVD [35] (middle) and the proposed G-NBVD (right)

(b) Clustring using the proposed method R-NBVD (left), SR-NBVD (middle) and GSR-NBVD(right)

Fig. 1. Illustration of clustering over IRIS data set through NBVD, proposed G-NBVD and three-steps clustering method: R-NBVD, SR-NBVD and GSR-NBVD, the similar shapes and colors confirm the correct clustering data points.

number of outliers. Then, we measure the accuracy of different clustering methods.

First, we show the outliers detection accuracy of our proposed G-NBVD and GSR-NBVD methods and several state-of-the-art clustering methods as shown in Figure 2(a)-left. We observe that G-NBVD and GSR-NBVD and GNMF clustering methods outperforms the other clustering methods for the outliers between 5% to 20% with the accuracy range of 90% to 76%. Then, the accuracy of proposed G-NBVD method decreases smoothly in the range of 20% to 40% of outliers with the higher accuracy results compared with the other methods. Second, we show the accuracy results of two selected classes of COIL-20 in the presence of outliers in Figures 2(a)-middle and Figures 2(a)-right. We observe that both proposed methods outperform the other methods. In particular, GSR-NBVD and G-NBCD keep the clustering accuracy of the class-3 of COIL-20 with around 58% from 5% to 18% and 55% from 5% to 27% outliers, respectively.

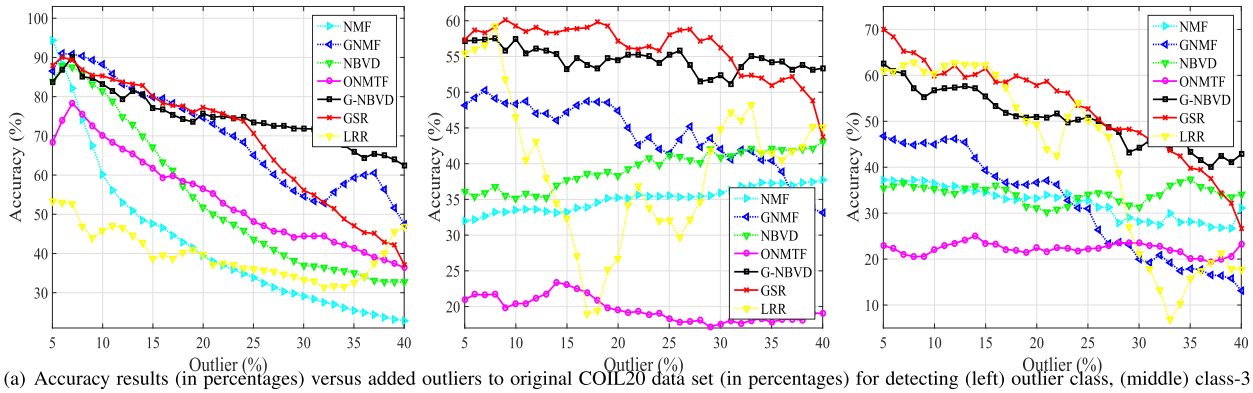
Moreover, we show the performance of clustering methods over corrupted data sets mentioned above through different clustering metrics including the clustering's accuracy, purity and ARI as given in equations (34), (38) and (39), respectively. The main observations are as follows:

- The proposed GSR-NBVD has the highest average clustering accuracy form 5% to 27% of outliers compared with the other methods and the performance is falling down sharply when outliers are more than 27% as shown in Figure 2(b)-left.
- The proposed G-NBVD has the second highest accuracy from 5% to 27% of outliers and has the first rank of average accuracy after that as shown in Figure 2(b)-left.
- Both GSR-NBVD and G-NBVD have the highest purity results with the first and the second rank, respectively as shown in Figure 2(b)-middle. Moreover, G-NMF meets the purity performance of G-NBVD from 20% of outliers.
- The proposed GSR-NBVD has the highest ARI with around 73% form 5% to 26% of outliers and this metric is falling down sharply when outliers are increased from 26% as shown in Figure 2(b)-right.
- The proposed G-NBVD has the best ARI from 26% to 40% of outliers compared with the other methods and has the second rank from 5% to 26% of outliers as shown in Figure 2(b)-right.

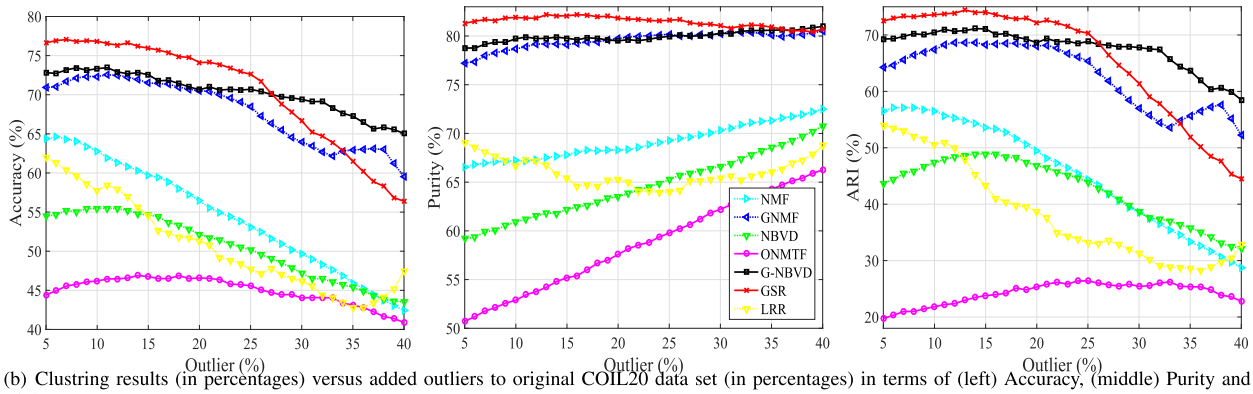
D. Results on Real Data Sets

1) Data Sets Description: We select eight different real-world data sets from document, image and biological data sets as follows

- **CSTR** consists of the abstracts of technical reports in 4 research topics published in the Department of Computer Science at Rochester University between 1991 and 2002



(a) Accuracy results (in percentages) versus added outliers to original COIL20 data set (in percentages) for detecting (left) outlier class, (middle) class-3 and (right) class-5



(b) Clustring results (in percentages) versus added outliers to original COIL20 data set (in percentages) in terms of (left) Accuracy, (middle) Purity and (right) ARI

Fig. 2. The clustering results as a function of added outliers (chosen from YALE data set) to original data set (COIL20) in percentages through different clustering methods NMF, GNMF, NBVD, ONMTF, LRR and our proposed methods of G-NBVD and GSR-NBVD.

and used for text categorization and clustering binary data in [52] and [53], respectively.

- **k1a** contains web pages in 20 subject directories of Yahoo! and it is built for the WebACE project [54], included in the CLUTO clustering toolkit [55].
- **k1b** is similar to k1a but more general directory hierarchy with 6 document categories.
- **re0** is the subset of standard Reuters-215785 dataset [56] which consists of news articles on the Reuters newswire in 1987 and contains 13 topics and used for co-clustering in [40].
- **re1** is another subset of Reuters-215785 dataset with 25 subjects and different size of instances and dimension [40].
- **COIL20** contains 32×32 gray scale images of 20 objects viewed from varying angles.
- **Ecoli** is a multiclass classification dataset belongs to UCI machine learning repository [57] including 8 attributes.
- **movements** contains 15 classes (out of 24 instances each) and each class corresponds to a typical hand movement [58].

Table II summarizes the characteristics of these data sets.

2) Measurements: We use four different metrics to measure the quality of clustering including clustering accuracy (CA), normalized mutual information (NMI), purity (PU), and adjusted rand index (ARI). They are defined as follows.

CA is a measure to calculate how accurately samples are grouped together. This measure needs clustering alignment and

TABLE II
DATASET INFORMATION

dataset	type	No. samples	dimension	No. clusters
CSTR	text	475	1000	4
k1a	text	2340	21839	20
k1b	text	2340	21839	6
re0	text	1504	2886	13
re1	text	1657	3758	25
COIL20	image	1440	1024	20
Ecoli	bio	336	7	8
movements	bio	360	90	15

we use Monkres algorithm [59] to align the clustering results and the ground truth to measure the clustering accuracy as

$$\text{CA} = \frac{\sum_{j=1}^{P} \delta(c_j, \tilde{c}_j)}{P} \tag{34}$$

where P is total number of samples, c_j is the ground truth cluster and $\tilde{c}_j$ is matched cluster obtained by clustering algorithm.

NMI is a normalized mutual information between two sets of clusters and does not require the alignment of the ground truth and clustering results. By considering two cluster sets of $\mathcal{C}$ and $\tilde{\mathcal{C}}$ which are the set of ground truth and the one obtained by clustering method respectively, and C_i and $\tilde{C}_j$ denote two sets of documents belong to ground truth cluster i-th and the obtained cluster of j-th, the mutual information (MI) is

TABLE III
NUMERICAL CLUSTERING RESULTS (MEAN VALUES) FOR K-MEANS, NMF, GNMF, NBVD, ONMTF [40], LRR [17], SSC [16] AND OUR PROPOSED METHODS OVER DIFFERENT DATA SETS IN TERMS OF FOUR METRICS: CLUSTERING ACCURACY, NORMALIZED MUTUAL INFORMATION, PURITY AND ADJUSTED RAND INDEX

dataset	mesure	K-means	NMF	GNMF	NBVD	ONMTF	LRR	SSC	G-NBVD	GSR-NBVD
CSTR	CA	80.98±7.56	72.17±0.83	86.02±7.16	86.05±8.58	79.48±9.27	87.79±0.001	85.68±0.001	91.69±0.76	**91.87±0.17**
	NMI	70.10±4.95	61.58±1.08	72.72±5.09	73.79±9.42	64.55±8.62	72.38±0.001	70.72±0.001	81.05±1.52	**81.42±0.46**
	PU	84.21±4.40	77.79±0.64	87.07±4.85	86.81±6.83	81.32±6.68	87.79±0.001	85.68±0.001	91.69±0.76	**91.87±0.17**
	ARI	68.98±9.09	54.93±1.56	76.31±7.97	73.27±15.43	60.43±14.30	75.34±0.001	76.69±0.001	84.48±2.04	**84.98±0.44**
k1a	CA	46.48±5.52	39.06±2.85	34.97±2.90	44.72±4.38	46.35±2.78	41.48±1.76	**48.44±1.80**	47.57±3.62	43.80±3.58
	NMI	56.14±1.89	51.69±2.13	44.83±1.24	52.92±2.26	47.03±3.52	53.28±0.49	51.81±0.47	**56.30±1.81**	54.23±1.77
	PU	**64.20±3.28**	58.12±1.87	51.20±0.88	60.71±2.20	54.12±2.96	60.19±1.17	56.88±0.47	62.50±2.40	60.92±2.45
	ARI	36.61±9.34	25.24±4.00	25.96±5.31	28.75±5.40	36.02±7.12	30.42±2.15	**45.78±4.64**	36.67±6.45	37.01±7.70
k1b	CA	78.99±8.51	72.25±0.72	88.91±3.85	88.92±3.23	73.62±11.63	88.50±0.001	74.97±0.001	89.45±3.01	**91.33±0.35**
	NMI	69.84±5.57	62.11±1.13	75.82±2.39	75.85±5.14	61.62±10.30	74.36±0.001	54.59±0.004	76.58±5.10	**80.11±0.56**
	PU	83.58±4.83	78.11±0.60	89.33±2.61	88.91±3.23	78.46±7.75	89.91±0.001	83.47±0.001	89.45±3.01	**91.33±0.35**
	ARI	66.83±10.64	55.52±1.34	80.12±3.36	77.68±7.64	54.84±16.73	77.72±0.001	58.47±0.23	79.00±6.67	**83.79±0.71**
re0	CA	36.15±3.32	38.05±2.12	21.02±0.58	38.91±6.42	35.68±4.71	29.78±0.65	30.09±1.13	37.81±3.92	**39.11±4.48**
	NMI	33.78±1.88	**33.99±1.61**	14.74±0.15	32.72±3.94	29.37±3.60	22.10±0.66	27.42±0.29	32.59±1.74	31.33±2.97
	PU	60.59±3.48	**63.68±1.50**	44.00±0.54	61.26±3.48	55.04±3.32	45.25±0.67	59.59±0.29	61.54±3.84	61.23±4.58
	ARI	13.86±2.65	14.64±2.11	0.01±0.001	14.78±6.88	10.36±5.07	6.58±0.53	12.77±0.59	15.53±4.11	**15.68±5.18**
re1	CA	34.51±1.63	33.19±1.66	23.02±1.10	33.12±3.55	30.07±1.74	37.80±2.54	31.64±0.64	**37.92±1.52**	37.49±2.04
	NMI	43.84±2.21	43.13±1.19	29.25±1.23	39.51±3.79	34.92±2.54	43.31±1.52	42.36±0.42	46.66±0.76	**46.74±0.61**
	PU	57.75±2.36	55.26±1.34	44.47±0.87	50.66±3.41	47.54±1.54	55.82±0.89	56.88±0.47	57.63±0.80	**57.94±0.78**
	ARI	16.84±2.64	12.38±1.67	9.54±0.80	9.35±2.68	5.71±1.75	17.23±1.74	17.25±0.46	22.62±1.17	**22.91±1.80**
COIL20	CA	63.73±2.30	63.57±3.00	68.41±2.85	54.22±3.02	45.45±4.10	63.73±3.39	72.22±0.001	74.60±6.58	**80.17±3.35**
	NMI	74.38±0.95	73.47±1.61	82.59±2.64	68.72±1.92	57.68±4.03	77.08±2.56	84.65±0.001	87.43±2.94	**89.53±1.35**
	PU	66.44±1.63	65.24±2.51	75.46±2.59	58.85±2.13	52.55±3.52	69.37±2.86	79.10±0.001	79.95±4.21	**82.97±2.65**
	ARI	55.94±2.04	55.89±3.31	63.20±5.98	44.71±3.84	22.01±5.66	53.21±5.62	56.97±0.001	70.97±7.61	**77.36±2.82**
Ecoli	CA	57.51±2.18	55.96±3.21	47.87±1.57	57.46±2.03	57.34±1.60	56.56±0.90	49.40±0.001	**62.78±0.89**	58.37±2.37
	NMI	54.14±0.78	49.09±2.35	46.38±1.79	52.69±0.47	52.94±0.71	49.04±0.44	48.03±0.001	**56.55±0.53**	54.52±2.11
	PU	82.69±0.94	78.91±2.16	75.94±1.57	82.20±0.80	81.91±0.88	81.64±0.20	80.65±0.001	**84.11±0.62**	82.94±1.04
	ARI	41.36±1.51	42.68±3.66	32.53±3.39	40.58±1.10	40.01±1.16	39.63±0.56	36.82±0.001	**44.87±1.21**	41.04±2.26
movements	CA	46.01±1.50	48.39±2.26	45.79±2.39	44.39±2.39	42.93±2.67	41.31±2.03	47.61±1.12	49.94±1.48	**50.36±1.63**
	NMI	58.42±1.38	58.01±2.14	59.93±1.67	56.89±1.77	55.12±1.98	50.77±1.90	59.68±0.69	63.79±0.84	**63.82±0.73**
	PU	48.47±1.44	50.40±2.48	48.72±1.68	46.83±2.20	45.33±2.31	45.35±1.81	51.40±1.05	52.94±0.97	**53.10±1.08**
	ARI	31.26±1.49	31.17±2.53	33.53±2.38	29.41±2.37	28.22±2.31	24.16±2.27	34.32±0.95	38.70±1.08	**38.82±1.07**

defined by

$$\mathrm{MI}(\mathcal{C},\tilde{\mathcal{C}}) = \sum_i \sum_j p(C_i,\tilde{C}_j)\log_2 \frac{p(C_i,\tilde{C}_j)}{p(C_i)p(\tilde{C}_j)} \tag{35}$$

or equivalently

$$\mathrm{MI}(\mathcal{C},\tilde{\mathcal{C}}) = \sum_i \sum_j \frac{|C_i \cap \tilde{C}_j|}{P}\log_2 \frac{P|C_i \cap \tilde{C}_j|}{|C_i||\tilde{C}_i|} \tag{36}$$

Then, NMI is defined as

$$\mathrm{NMI}(\mathcal{C},\tilde{\mathcal{C}}) = \frac{\mathrm{MI}}{\max(E(\mathcal{C}),E(\tilde{\mathcal{C}}))} \tag{37}$$

where $E(.)$ denotes the entropy function.

PU is used to measure how samples belongs to only one cluster are grouped together and it is computed by

$$\mathrm{PU} = \frac{1}{P}\sum_{i=1}^{P} \max_j |c_i \cap \tilde{c}_j| \tag{38}$$

ARI is a modification of rand index (RI). RI measures the agreement between two sets of labels and ARI considers the agreement by chance as well, defined by

$$\mathrm{ARI} = \frac{(\text{index}) - (\text{expected index})}{\max((\text{index}) - (\text{expected index})} \tag{39}$$

3) Clustering Results: In this experiment, we compare the clustering results of NBVD-based frameworks (NBVD method [35], the proposed G-NBVD and GSR-NBVD methods) with the NMF-based methods (NMF [3] and GNMF [33]), K-means, LRR [17] and SSC [16] as the benchmarks. We measure the clustering results of all these methods through four mentioned metrics: accuracy, purity, NMI and ARI.

Table III shows these results and the best and the second best results are shown with **underline-bold** and underline-underline, respectively.

The main results are summarized as follows:

- The proposed GSR-NBVD has the best clustering results (all four mentioned metrics) over CSTR, k1b, COIL20 and movements data sets. Also, it outperforms the other methods over re1 through average NMI, purity and ARI and over re0 through average accuracy and ARI.
- The proposed G-NBVD has the second best clustering performances over CSTR, k1b (except purity), k1a (except NMI and ARI), re0 (except accuracy and NMI), re1 (except accuracy), COIL20 and movements.
- The proposed G-NBVD and GSR-NBVD have the best and the second best clustering results over Ecoli data set, respectively.
- In general, the proposed GSR-NBVD and G-NBVD clustering methods have the best clustering performances, in 21 and 6 cases, respectively, out of 32 possible cases over

all 8 data sets. Moreover, they have the second best clustering performances, in 4 and 22 cases, respectively, out of 32 possible cases.

In summary, we observe that our proposed algorithm outperforms substantially the other co-clustering and clustering algorithms in almost all cases, in term of four mentioned metric values.

V. Conclusion and Future Work

In this paper, we proposed two clustering methods based on 3-factorization matrices: a graph based NBVD and a robust NBVD approach through jointly graph and sparse regularizers for clustering. In our first proposed method, we used the geometric properties of data along with NBVD framework called G-NBVD. In the second proposed clustering method, we used the $\ell_{2,1}$-norm term to model the error that compensates the outliers effects. To improve our proposed R-NBVD, we investigated the sparseness property over robust NBVD framework and enforced the $\ell_{\frac{1}{2}}$-norm constrains on the connector matrix between the feature and coefficient matrix. To have further enhancement over SR-NBVD, we proposed to use graph representation of data matrix and introduced our final GSR-NBVD model. To solve our proposed minimization problems (i.e., both G-NBVD and GSR-NBVD frameworks) at each step, we employed multiplicative updating rules and proved the convergence through several theorems. The evaluations of our proposed methods over various types of data sets affirmed the effectiveness of our proposed co-clustering methods which outperformed several state-of-the-art methods.

In our future works, we will investigate how to determine the values of matrix dimensions where they have a linear relation with the number of clusters currently. Also, the regularizers' estimation is the other open problem. Furthermore, we are working to find a theoretical approach for the possible boundary conditions for the outliers that can be handled through our proposed algorithms as well as the possible assumptions on the nature of outliers such as sparseness property. Finally, the direct proof of robustness against outliers/noises through the $\ell_{2,1}$-norm is left for our future work.

Appendix A
Proof of Theorem 1

Proof: The proof of this theorem is similar to the proof of theorems mentioned later for the robust based NBVD approach by substituting $\mathbf{D} = 1$ in all parts. For the proof of part (i), we can follow Appendix B. For parts (ii) and (iii) , we use the similar approach in Appendix C and Appendix D, respectively and put $\lambda_s = 0$. ■

Appendix B
Proof of Theorem 2

Proof: (i) First, we show that the following cost function monotonically decreases.

$$C(\mathbf{U}) = \mathrm{Tr}\big((\mathbf{X} - \mathbf{USV}^T)\mathbf{D}(\mathbf{X} - \mathbf{USV}^T)^T\big) \tag{40}$$

Let $\mathbf{S}^r$ and $\mathbf{V}^r$ be the old values of $\mathbf{S}$ and $\mathbf{V}$ on the right-hand-side (RHS) of (23) and (24), respectively, $\mathbf{S}^{r+1}$ and $\mathbf{V}^{r+1}$ be the new values of $\mathbf{S}$ and $\mathbf{V}$ on the left-hand-side (LHS) of (23) and (24). So, we must show $C(\mathbf{U}^{r+1}) \leq C(\mathbf{U}^r)$. Using the axillary function approach in [3], $Z(\mathbf{U}, \hat{\mathbf{U}})$ is an auxiliary function of $C(\mathbf{U})$ if the following relations are satisfied

$$Z(\mathbf{U}, \hat{\mathbf{U}}) \geq C(\mathbf{U}), \quad \forall \hat{U}, \tag{41}$$

$$Z(\mathbf{U}, \mathbf{U}) = C(\mathbf{U}). \tag{42}$$

By defining

$$\mathbf{U}^{r+1} = \underset{\mathbf{U}}{\operatorname{argmin}}\, Z(\mathbf{U}, \mathbf{U}^r) \tag{43}$$

we have

$$C(\mathbf{U}^{r+1}) = Z(\mathbf{U}^{r+1}, \mathbf{U}^{r+1}) \leq Z(\mathbf{U}^{r+1}, \mathbf{U}^r) \leq C(\mathbf{U}^r) \tag{44}$$

So, this proves that $C(\mathbf{U}^r)$ monotonically decreases.

Now, we show that the following function is an auxiliary function of $C(\mathbf{U})$

$$Z(\mathbf{U}, \hat{\mathbf{U}}) = \mathrm{Tr}\big(\mathbf{XDX}^T - 2\mathbf{U}^T\mathbf{XDVS}^T\big) + \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{(\hat{\mathbf{U}}\mathbf{SV}^T\mathbf{DVS}^T)_{ij}\mathbf{U}_{ij}^2}{\hat{\mathbf{U}}_{ij}} \tag{45}$$

By rearranging the RHS of $C(\mathbf{U})$ in (40) and using the fact that the trace operator is invariant under cyclic permutations (i.e. $\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}) = \mathrm{Tr}(\mathbf{BCA})$), we have

$$C(\mathbf{U}) = \mathrm{Tr}\big(\mathbf{XDX}^T - 2\mathbf{U}^T\mathbf{XDVS}^T + \mathbf{U}^T\mathbf{USV}^T\mathbf{DVS}^T\big) \tag{46}$$

Then, by using the following matrix inequality [46]

$$\mathrm{Tr}(\mathbf{G}^T\mathbf{AGQ}) \leq \sum_{i,j}(\mathbf{A}\hat{\mathbf{G}}\mathbf{Q})_{ij}\frac{\mathbf{G}_{ij}^2}{\hat{\mathbf{G}}_{ij}} \tag{47}$$

where the matrices $\mathbf{A}, \mathbf{Q}, \mathbf{G}$ are nonnegative matrices with the suitable dimensions and $\mathbf{A} = \mathbf{A}^T$, $\mathbf{Q} = \mathbf{Q}^T$ and the equality holds if $\mathbf{G} = \hat{\mathbf{G}}$, and setting $\mathbf{A} = 1, \mathbf{Q} = \mathbf{SV}^T\mathbf{DVS}^T, \mathbf{G} = \mathbf{U}$ and $\hat{\mathbf{G}} = \hat{\mathbf{U}}$, then the third term of (46) is always smaller than or equal to the third term of (45). The equality holds if $\mathbf{U} = \hat{\mathbf{U}}$. Thus, $Z(\mathbf{U}, \hat{\mathbf{U}})$ of (45) is an auxiliary function of cost function $C(\mathbf{U})$ in (46).

To find the global minima of (45), let $f(\mathbf{U}) = Z(\mathbf{U}, \hat{\mathbf{U}})$, then

$$\frac{\partial f(\mathbf{U})}{\partial \mathbf{U}_{ij}} = -2\mathbf{XDVS}^T + 2\frac{(\hat{\mathbf{U}}\mathbf{SV}^T\mathbf{DVS}^T)_{ij}\mathbf{U}_{ij}}{\hat{\mathbf{U}}_{ij}} \tag{48}$$

and the Hessian matrix (the second order derivatives) is

$$\frac{\partial^2 f(\mathbf{U})}{\partial \mathbf{U}_{ij}\partial \mathbf{U}_{kl}} = \left(2\frac{(\hat{\mathbf{U}}\mathbf{SV}^T\mathbf{DVS}^T)_{ij}}{\hat{\mathbf{U}}_{ij}}\right)\delta_{jl}\delta_{ik} \tag{49}$$

Hence, the Hessian matrix is semi-positive which implies that the function $f(\mathbf{U})$ is a convex function and it has a unique global minima. It is obtained by setting the LHS of (48) to zero that

gives

$$\mathbf{U}_{ij} = \hat{\mathbf{U}}_{ij} \frac{(\mathbf{XDVS}^T)_{ij}}{(\hat{\mathbf{U}}\mathbf{SV}^T\mathbf{DVS}^T)_{ij}} \tag{50}$$

The above equation recovers the updating rule for $\mathbf{U}$ in (22) by replacing $\mathbf{U}^{r+1} \leftarrow \mathbf{U}$ and $\mathbf{U}^r \leftarrow \hat{\mathbf{U}}$. Thus, the objective function $C(\mathbf{U})$ in (40) is nonincreasing monotonically under the updating rules (50).

Second, we show that the following inequation holds under the update rule of (22).

$$||\mathbf{X} - \mathbf{U}^{r+1}\mathbf{SV}^T||_{2,1} - ||\mathbf{X} - \mathbf{U}^r\mathbf{SV}^T||_{2,1} \leq \frac{1}{2}\big(C(\mathbf{U}^{r+1}) - C(\mathbf{U}^r)\big) \tag{51}$$

where $C(\mathbf{U})$ is defined in (40).

We rewrite $C(\mathbf{U}^{r+1})$ and $C(\mathbf{U}^r)$ as follows, respectively:

$$\begin{aligned} C(\mathbf{U}^{r+1}) &= \mathrm{Tr}\big((\mathbf{X} - \mathbf{U}^{r+1}\mathbf{SV}^T)\mathbf{D}(\mathbf{X} - \mathbf{U}^{r+1}\mathbf{SV}^T)^T\big) \\ &= \sum_{i=1}^{P}\sum_{j=1}^{L}(\mathbf{X} - \mathbf{U}^{r+1}\mathbf{SV}^T)_{ji}^2\mathbf{D}_{ii} \\ &= \sum_{i=1}^{P}||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T||^2\mathbf{D}_{ii} \end{aligned} \tag{52}$$

$$\begin{aligned} C(\mathbf{U}^{r}) &= \mathrm{Tr}\big((\mathbf{X} - \mathbf{U}^{r}\mathbf{SV}^T)\mathbf{D}(\mathbf{X} - \mathbf{U}^{r}\mathbf{SV}^T)^T\big) \\ &= \sum_{i=1}^{P}\sum_{j=1}^{L}(\mathbf{X} - \mathbf{U}^{r}\mathbf{SV}^T)_{ji}^2\mathbf{D}_{ii} \\ &= \sum_{i=1}^{P}||\mathbf{X}_{:,i} - \mathbf{U}^{r}\mathbf{SV}_{:,i}^T||^2\mathbf{D}_{ii} \end{aligned} \tag{53}$$

Then, the RHS of (51) becomes

$$\begin{aligned} &\frac{1}{2}\big(C(\mathbf{U}^{r+1}) - C(\mathbf{U}^r)\big) = \\ &\frac{1}{2}\sum_{i=1}^{P}\Big(||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T||^2\mathbf{D}_{ii} - ||\mathbf{X}_{:,i} - \mathbf{U}^{r}\mathbf{SV}_{:,i}^T||^2\mathbf{D}_{ii}\Big) \\ &= \frac{1}{2}\sum_{i=1}^{P}\left(||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T||^2\mathbf{D}_{ii} - \frac{1}{\mathbf{D}_{ii}}\right) \end{aligned} \tag{54}$$

where we use the definition of diagonal matrix $\mathbf{D}$ in (26) by replacing $\mathbf{U}$ with $\mathbf{U}^r$ in the last relation of (54).

Also, we rewrite the LHS of (51) as follows

$$\begin{aligned} &||\mathbf{X} - \mathbf{U}^{r+1}\mathbf{SV}^T||_{2,1} - ||\mathbf{X} - \mathbf{U}^{r}\mathbf{SV}^T||_{2,1} \\ &= \sum_{i=1}^{P}\Big(||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T|| - ||\mathbf{X}_{:,i} - \mathbf{U}^{r}\mathbf{SV}_{:,i}^T||\Big) \\ &= \sum_{i=1}^{P}\left(||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T|| - \frac{1}{\mathbf{D}_{ii}}\right) \end{aligned} \tag{55}$$

Thus, by subtracting (55) from (54)

$$\begin{aligned} &\sum_{i=1}^{P}\left(||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T|| \right. \\ &\quad \left. - \frac{1}{2}||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T||^2\mathbf{D}_{ii} - \frac{1}{2\mathbf{D}_{ii}}\right) \\ &= \sum_{i=1}^{P} -\frac{\mathbf{D}_{ii}}{2}\left(||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T||^2 \right. \\ &\quad \left. - 2||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T||\frac{1}{\mathbf{D}_{ii}} + \frac{1}{\mathbf{D}_{ii}^2}\right) \\ &= \sum_{i=1}^{P} -\frac{\mathbf{D}_{ii}}{2}\left(||\mathbf{X}_{:,i} - \mathbf{U}^{r+1}\mathbf{SV}_{:,i}^T|| - \frac{1}{\mathbf{D}_{ii}}\right)^2 \\ &\leq 0 \end{aligned} \tag{56}$$

and this ensures that (51) holds. Thus the value of RHS (51) must be negative or zero. Then

$$||\mathbf{X} - \mathbf{U}^{r+1}\mathbf{SV}^T||_{2,1} - ||\mathbf{X} - \mathbf{U}^{r}\mathbf{SV}^T||_{2,1} \leq 0 \tag{57}$$

Consequently, the objective function of (21) monotonically decreases using updating $\mathbf{U}$ in (22) while fixing $\mathbf{S}$ and $\mathbf{V}$.

We could follow the same procedure mentioned in the next Appendices to prove parts (ii) and (iii) of Theorem 2 by letting $\lambda_s = 0$ and $\lambda_g = 0$, respectively, in Appendix C and Appendix D. ■

APPENDIX C
PROOF OF THEOREM 3

Proof: The first part (i) is easily proved based on the the proof of updating rules for $\mathbf{U}$ in Appendix B. Then, we follow the same procedure mentioned earlier in Appendix B and using the new auxiliary function and some manipulations to prove part (ii) which explained as follows.

First, we rewrite the objective function of (27) as the following $\mathbf{S}$-dependent cost function

$$C(\mathbf{S}) = \mathrm{Tr}\big((\mathbf{X} - \mathbf{USV}^T)\mathbf{D}(\mathbf{X} - \mathbf{USV}^T)^T\big) + \lambda_s\sum_{i=1}^{K}\sum_{j=1}^{N}|s_{ij}|^{\frac{1}{2}} \tag{58}$$

Then, we show that the following function is an auxiliary function of $C(\mathbf{S})$

$$Z(\mathbf{S}, \hat{\mathbf{S}}) = \mathrm{Tr}\big(\mathbf{XDX}^T - 2\mathbf{XDVS}^T\mathbf{U}^T\big) + \lambda_s\sum_{i=1}^{K}\sum_{j=1}^{N}|s_{ij}|^{\frac{1}{2}} + \sum_{i=1}^{K}\sum_{j=1}^{N}\frac{(\mathbf{U}^T\mathbf{U}\hat{\mathbf{S}}\mathbf{V}^T\mathbf{DV})_{ij}\mathbf{S}_{ij}^2}{\hat{\mathbf{S}}_{ij}} \tag{59}$$

Now, we show that the required conditions of an auxiliary function must be satisfied as follows

$$Z(\mathbf{S},\hat{\mathbf{S}}) \geq C(\mathbf{S}), \quad \forall \hat{S}, \tag{60}$$

$$Z(\mathbf{S},\mathbf{S}) = C(\mathbf{S}). \tag{61}$$

We can rewrite the RHS of (58) as follows

$$C(\mathbf{S}) = \mathrm{Tr}\big(\mathbf{X}\mathbf{D}\mathbf{X}^T - 2\mathbf{X}\mathbf{D}\mathbf{V}\mathbf{S}^T\mathbf{U}^T\big) + \lambda_s \sum_{i=1}^{K}\sum_{j=1}^{N} |s_{ij}|^{\frac{1}{2}} + \mathrm{Tr}(\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{D}\mathbf{V}\mathbf{S}^T\mathbf{U}^T) \tag{62}$$

Using the matrix inequality of (47) and assigning $\mathbf{A} = \mathbf{U}^T\mathbf{U}$, $\mathbf{Q} = \mathbf{V}^T\mathbf{D}\mathbf{V}$, $\mathbf{G} = \mathbf{S}$ and $\hat{\mathbf{G}} = \hat{\mathbf{S}}$, we see that the forth term of (62) is always smaller than or equal to the forth term of (59) and the equality holds if $\mathbf{S} = \hat{\mathbf{S}}$ .Thus, $Z(\mathbf{S},\hat{\mathbf{S}})$ in (59) is an auxiliary function of $C(\mathbf{S})$ in (58).

Now, we construct the first and second order derivatives of (59) to find its global minimia. Let $f(\mathbf{S}) = Z(\mathbf{S},\hat{\mathbf{S}})$, so

$$\frac{\partial f(\mathbf{S})}{\partial \mathbf{S}_{ij}} = -2\mathbf{U}^T\mathbf{X}\mathbf{D}\mathbf{V} + \frac{\lambda_s}{2}\mathbf{S}_{ij}^{-\frac{1}{2}} + 2\frac{(\mathbf{U}^T\mathbf{U}\hat{\mathbf{S}}\mathbf{V}^T\mathbf{D}\mathbf{V})_{ij}\mathbf{S}_{ij}}{\hat{\mathbf{S}}_{ij}} \tag{63}$$

and the Hessian matrix is

$$\frac{\partial^2 f(\mathbf{S})}{\partial \mathbf{S}_{ij}\partial \mathbf{S}_{kl}} = \frac{1}{4}\bigg(-\lambda_s\mathbf{S}_{ij}^{-\frac{3}{2}} + 8\frac{(\mathbf{U}^T\mathbf{U}\hat{\mathbf{S}}\mathbf{V}^T\mathbf{D}\mathbf{V})_{ij}}{\hat{\mathbf{S}}_{ij}}\bigg)\delta_{jl}\delta_{ik} \tag{64}$$

By choosing the appropriate values of λ_s, (64) will be semi-positive, and hence, $Z(\mathbf{S},\hat{\mathbf{S}})$ in (59) has a unique global minima which obtained by

$$\mathbf{S}_{ij} = \hat{\mathbf{S}}_{ij}\frac{(\mathbf{U}^T\mathbf{X}\mathbf{D}\mathbf{V})_{ij}}{(\mathbf{U}^T\mathbf{U}\hat{\mathbf{S}}\mathbf{V}^T\mathbf{D}\mathbf{V} + \frac{\lambda_s}{2}\mathbf{S}^{\frac{1}{2}})_{ij}} \tag{65}$$

By replacing $\mathbf{S}^{r+1} \leftarrow \mathbf{S}$ and $\mathbf{S}^r \leftarrow \hat{\mathbf{S}}$ in (65), it is easy to show that $C(\mathbf{S})$ in (58) is monotonically nonincreasing under the update rule of (28).

In the next step, we need to show that

$$||\mathbf{X} - \mathbf{U}\mathbf{S}^{r+1}\mathbf{V}^T||_{2,1} - ||\mathbf{X} - \mathbf{U}\mathbf{S}^r\mathbf{V}^T||_{2,1} \leq \frac{1}{2}\big(C(\mathbf{S}^{r+1}) - C(\mathbf{S}^r)\big) \tag{66}$$

where $C(\mathbf{S})$ is defined in (58).

We can prove this by showing that the RHS of (66) is negative or zero which the steps of the proof are similar to the second part of Appendix B and we skip them. ■

Appendix D
Proof of Theorem 4

Proof: We follow the same procedure mentioned earlier in Appendix B and Appendix C to prove all parts of theorem. Since the second and the third terms of (30) depends on $\mathbf{S}$ and $\mathbf{V}$, we have exactly the same update rule as in R-NBVD and the proof is similar. For the second part of Theorem, we have the similar procedure as in Theorem 3 and it is skipped. We prove the third part of Theorem by defining the following $\mathbf{V}$-dependent cost function (by removing the second part of original objective function of GSR-RNBVD in (30)

$$C(\mathbf{V}) = \mathrm{Tr}\big((\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{D}(\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T)^T + \lambda_g\mathbf{V}^T\mathbf{B}\mathbf{V}\big) \tag{67}$$

Then, we show that the following function is an auxiliary function of $C(\mathbf{V})$

$$Z(\mathbf{V},\hat{\mathbf{V}}) = \mathrm{Tr}\big(\mathbf{X}\mathbf{D}\mathbf{X}^T - 2\mathbf{X}\mathbf{D}\mathbf{V}\mathbf{S}^T\mathbf{U}^T + \lambda_g\mathbf{V}^T\mathbf{B}\mathbf{V}\big) + \sum_{i=1}^{P}\sum_{j=1}^{N}\frac{(\mathbf{D}\hat{\mathbf{V}}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S})_{ij}\mathbf{V}_{ij}^2}{\hat{\mathbf{V}}_{ij}} \tag{68}$$

Again we have to show that the following conditions are satisfied:

$$Z(\mathbf{V},\hat{\mathbf{V}}) \geq C(\mathbf{V}), \quad \forall \hat{V}, \tag{69}$$

$$Z(\mathbf{V},\mathbf{V}) = C(\mathbf{V}). \tag{70}$$

The RHS of (67) can be written as

$$C(\mathbf{V}) = \mathrm{Tr}\big(\mathbf{X}\mathbf{D}\mathbf{X}^T - 2\mathbf{X}\mathbf{D}\mathbf{V}\mathbf{S}^T\mathbf{U}^T + \lambda_g\mathbf{V}^T\mathbf{B}\mathbf{V} + \mathbf{V}^T\mathbf{D}\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\big) \tag{71}$$

By using the matrix inequality mentioned in (47) and assigning $\mathbf{A} = \mathbf{D}$, $\mathbf{Q} = \mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}$, $\mathbf{G} = \mathbf{V}$ and $\hat{\mathbf{G}} = \hat{\mathbf{V}}$, we see that the forth term of (71) is always smaller than or equal to the forth term of (68). Therefore, $Z(\mathbf{V},\hat{\mathbf{V}})$ in (68) is an auxiliary function of $C(\mathbf{V})$ in (71).

Then, we construct the first and second order derivatives of (68) to find its global minimia. Let $f(\mathbf{V}) = Z(\mathbf{V},\hat{\mathbf{V}})$, so

$$\frac{\partial f(\mathbf{V})}{\partial \mathbf{V}_{ij}} = -2\mathbf{D}\mathbf{X}^T\mathbf{U}\mathbf{S} + 2\lambda_g\mathbf{B}\mathbf{V} + 2\frac{(\mathbf{D}\hat{\mathbf{V}}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S})_{ij}\mathbf{V}_{ij}}{\hat{\mathbf{V}}_{ij}} \tag{72}$$

and the Hessian matrix is

$$\frac{\partial^2 f(\mathbf{V})}{\partial \mathbf{V}_{ij}\partial \mathbf{V}_{kl}} = 2\bigg(\lambda_g\mathbf{B} + \frac{(\mathbf{D}\hat{\mathbf{V}}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S})_{ij}}{\hat{\mathbf{V}}_{ij}}\bigg)\delta_{jl}\delta_{ik} \tag{73}$$

Since (73) is semi-positive, $Z(\mathbf{V},\hat{\mathbf{V}})$ in (68) has a unique global minima which obtained by

$$\mathbf{V}_{ij} = \hat{\mathbf{V}}_{ij}\frac{(\mathbf{D}\mathbf{X}^T\mathbf{U}\mathbf{S} + \lambda_g\mathbf{W}\mathbf{V})_{ij}}{(\mathbf{D}\hat{\mathbf{V}}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S} + \lambda_g\mathbf{A}\mathbf{V}))_{ij}} \tag{74}$$

where $\mathbf{B}$ is already substituted by $\mathbf{A} - \mathbf{W}$. So, the updating rule $\mathbf{V}$ in (31) is recovered by (74) by exchanging $\mathbf{V}^{r+1} \leftarrow \mathbf{V}$ and $\mathbf{V}^r \leftarrow \hat{\mathbf{V}}$ where it confirms $C(\mathbf{V})$ in (67) is monotonically nonincreasing under the update rule (74).

Similar to the second part of Appendix B, we can easily show that the following inequation holds under the update rule of (31)

$$||\mathbf{X} - \mathbf{U}\mathbf{S}(\mathbf{V}^{r+1})^T||_{2,1} - ||\mathbf{X} - \mathbf{U}\mathbf{S}(\mathbf{V}^r)^T||_{2,1} \leq \frac{1}{2}\big(C(\mathbf{V}^{r+1}) - C(\mathbf{V}^r)\big) \tag{75}$$

where $C(\mathbf{V})$ is defined in (67). Indeed, we could show that the RHS of (75) is negative or zero that completes the proof. ■

REFERENCES

[1] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, Jun. 1994.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[4] I. Jolliffe, "Principal component analysis," in *International Encyclopedia of Statistical Science*. Berlin, Germany: Springer, 2011, pp. 1094–1096.

[5] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[6] M. Yin, S. Xie, Z. Wu, Y. Zhang, and J. Gao, "Subspace clustering via learning an adaptive low-rank graph," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3716–3728, Aug. 2018.

[7] T. E. Boult and L. G. Brown, "Factorization-based segmentation of motions," in *Proc. IEEE Workshop Vis. Motion*, 1991, pp. 179–186.

[8] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.

[9] P. S. Bradley and O. L. Mangasarian, "K-plane clustering," *J. Global Optim.*, vol. 16, no. 1, pp. 23–32, 2000.

[10] P. K. Agarwal and N. H. Mustafa, "K-means projective clustering," in *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, 2004, pp. 155–165.

[11] A. Y. Yang, S. R. Rao, and Y. Ma, "Robust statistical estimation and segmentation of multiple subspaces," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2006, pp. 99–106.

[12] H. Derksen, Y. Ma, W. Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy coding and compression," in *Proc. Vis. Commun. Image Process.*, 2007, vol. 6508, Art. no. 65080H.

[13] G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *Int. J. Comput. Vis.*, vol. 81, no. 3, pp. 317–330, 2009.

[14] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 2790–2797.

[15] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[16] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[18] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning spatially localized, parts-based representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. I-207–I-212.

[19] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly, "Hyperspectral unmixing via $l_{\frac{1}{2}}$ sparsity-constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4282–4297, Nov. 2011.

[20] Y. Esmaeili Salehani and S. Gazor, "Smooth and sparse regularization for NMF hyperspectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3677–3692, Aug. 2017.

[21] Y. Esmaeili Salehani and S. Gazor, "Collaborative unmixing hyperspectral imagery via nonnegative matrix factorization," in *Proc. Int. Conf. Image Signal Process.*, Jun. 2016, pp. 118–126.

[22] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[23] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," College Comput., Georgia Inst. Technol., Atlanta, GA, USA, Tech. Rep., 2008.

[24] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2013.

[25] 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*. Hoboken, NJ, USA: Wiley, 2009.

[26] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 267–273.

[27] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Inf. Process. Manage.*, vol. 42, no. 2, pp. 373–386, 2006.

[28] P. O. Hoyer, "Non-negative sparse coding," in *Proc. 12th IEEE Workshop Neural Netw. Signal Process.*, 2002, pp. 557–565.

[29] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using L21-norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 673–682.

[30] Z. Li, J. Tang, and X. He, "Robust structured nonnegative matrix factorization for image representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1947–1960, May 2018.

[31] Z. Lin, C. Xu, and H. Zha, "Robust matrix factorization by majorization minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 208–220, Jan. 2018.

[32] Y. Yuan, M. Fu, and X. Lu, "Substance dependence constrained sparse NMF for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 2975–2986, Jun. 2015.

[33] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[34] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," *Frontiers Elect. Electron. Eng. China*, vol. 6, no. 2, pp. 192–200, Jun. 2011. [Online]. Available: https://doi.org/10.1007/s11460-011-0128-0

[35] B. Long, Z. M. Zhang, and P. S. Yu, "Co-clustering by block value decomposition," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 635–640.

[36] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2003, pp. 89–98. [Online]. Available: http://doi.acm.org/10.1145/956750.956764

[37] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 269–274.

[38] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Co-clustering of biological networks and gene expression data," *Bioinformatics*, vol. 18, no. suppl_1, pp. S145–S154, 2002.

[39] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix T-factorizations for clustering," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2006, pp. 126–135.

[40] J. Yoo and S. Choi, "Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds," *Inf. Process. Manage.*, vol. 46, no. 5, pp. 559–570, 2010.

[41] S. J. Kim, T. Hwang, and G. B. Giannakis, "Sparse robust matrix tri-factorization with application to cancer genomics," in *Proc. 3rd Int. Workshop Cogn. Inf. Process.*, May 2012, pp. 1–6.

[42] A. Mirzal, "A convergent algorithm for orthogonal nonnegative matrix factorization," *J. Comput. Appl. Math.*, vol. 260, pp. 149–166, 2014.

[43] Y. Esmaeili Salehani, S. Gazor, I.-M. Kim, and S. Yousefi, "ℓ_0-norm sparse hyperspectral unmixing using arctan smoothing," *Remote Sens.*, vol. 8, no. 3, pp. 1–20, Feb. 2016.

[44] Y. Esmaeili Salehani and S. Gazor, "Sparse data reconstruction via adaptive ℓ_p-norm and multilayer NMF," in *Proc. IEEE 7th Annu. Inf. Technol., Electron. Mobile Commun. Conf.*, Oct. 2016, pp. 1–6.

[45] Y. Esmaeili Salehani and M. Cheriet, "Non-dictionary aided sparse unmixing of hyperspectral images via weighted nonnegative matrix factorization," in *Proc. Int. Conf. Image Anal. Recognit.*, Montreal, QC, Canada, 2017, pp. 596–604.

[46] C. H. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[47] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x

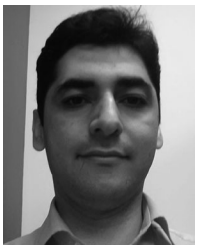
[48] E. Anderson, "The species problem in iris," *Ann. Missouri Botanical Garden*, vol. 23, pp. 457–509, 1936. [Online]. Available: https://www.biodiversitylibrary.org/part/4079

[49] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal Component Analysis*. Berlin, Germany: Springer, 1986, pp. 115–128.

[50] S. A. Nene, S. K. Nayar, and H. H. Murase, "Columbia object image library (coil-20)," Comput. Vis. Lab., Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-005-96, 1996.

[51] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[52] T. Li, S. Zhu, and M. Ogihara, "Efficient multi-way text categorization via generalized discriminant analysis," in *Proc. 12th Int. Conf. Inf. Knowl. Manage.*, 2003, pp. 317–324.
[53] T. Li, "A general model for clustering binary data," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 188–197.
[54] E.-H. Han *et al.*, "Webace: A web agent for document categorization and exploration," in *Proc. 2nd Int. Conf. Auton. Agents*, 1998, pp. 408–415.
[55] 2006. [Online]. Available: http://www.glaros.dtc.umn.edu/gkhome/cluto/cluto/overview
[56] 1990. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection
[57] P. Horton and K. Nakai, "A probabilistic classification system for predicting the cellular localization sites of proteins," in *Proc. 4th Int. Conf. Intell. Syst. Mol. Biol.*, 1996, pp. 109–115. [Online]. Available: http://dl.acm.org/citation.cfm?id=645631.662879
[58] D. B. Dias, R. C. B. Madeo, T. Rocha, H. H. Bíscaro, and S. M. Peres, "Hand movement recognition for brazilian sign language: A study using distance-based neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Piscataway, NJ, USA, 2009, pp. 2355–2362. [Online]. Available: http://dl.acm.org/citation.cfm?id=1704555.1704610
[59] W. Wessel, "Lovász, L.; Plummer, M. D., Matching Theory. Budapest, Akadémiai Kiadó 1986. XXXIII, 544 S., FT 680,—. ISBN 9630541688," *ZAMM—J. Appl. Math. Mech. / Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 68, no. 3, pp. 146–146, 1988. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/zamm.19880680310



Yaser Esmaeili Salehani received the B.Sc. degree with *summa cum laude* from the Electrical Engineering Department, Khaje-Nasir Toosi University of Technology, Tehran, Iran, and the M.Sc. degree from the Electrical Engineering Department, Sharif University of Technology, Tehran, Iran, both in the communications engineering in 2002 and 2004, respectively, the M.A.Sc. degree from Electrical and Computer Engineering Department, Concordia University, Montreal, QC, Canada, with the early graduate award in 2011, and the Ph.D. degree from Electrical and Computer Engineering Department, Queen's University, Kingston, ON, Canada, in 2016 with the IEEE Ph.D. Research Excellence Award. He is currently a Post-Doctoral Fellow with Synchromedia Laboratory, École de Technologie Supérieure, Montreal, QC, Canada. His current research interests include signal and image processing in general, machine learning techniques and sparse problems, big data analysis such as hyperspectral images and medical imaging, manuscript dating/ageing, and document analysis.

Ehsan Arabnejad received the B.Sc. degree in electrical engineering (electronic) in 2003, the M.Sc. degree in electrical engineering (telecommunication) in 2009, and the Ph.D. degree in computer science in 2018.

He is currently working with Synchromedia Laboratory for multimedia communication in telepresence, École de Technologie Supérieure (University of Quebec), Montreal, QC, Canada. His research interests include pattern recognition and document understanding.

Mohamed Cheriet (SM'95) received the M.Sc. and Ph.D. degrees in computer science from the University of Pierre et Marie Curie (Paris VI), Paris, France, in 1985 and 1988, respectively.

Since 1992, he has been a Professor with the Automation Engineering Department, École de Technologie Supérieure (ETS) (University of Quebec), Montreal, QC, Canada, and was a Full Professor in 1998. Since 1998 he has been the founder and Director of Synchromedia Laboratory, ETS, which targets multimedia communication in telepresence applications. He also co-founded the Laboratory for Imagery, Vision and Artificial Intelligence, ÉTS, and was the Director from 2000 to 2006. He is an expert in computational intelligence, pattern recognition, mathematical modeling for image processing, cognitive and machine learning approaches, and perception. His research has acquired extensive experience in cloud computing and network virtualization. In addition, he has authored or coauthored more than 350 technical papers in the field. He serves on the editorial boards of several renowned journals and international conferences. He co-authored a book entitled *Character Recognition Systems: A guide for Students and Practitioners* (Wiley, Spring 2007). He is a Fellow of the International Association for Pattern Recognition, a Fellow of the Canadian Academy of Engineering, the founder and former Chair of the IEEE Montreal Chapter of Computational Intelligent Systems, the Steering Committee Member of the IEEE Green ICT Initiative, and the Chair of the IEEE ICT Emissions Working Group. He was the recipient of the 2016 IEEE J. M. Ham Outstanding Engineering Educator Award and the 2012 Queen Elizabeth II Diamond Jubilee Medal.