Enhancing Clustering by Exploiting Complementary Data Modalities in the Medical Domain

Martin Schultz, Michael Krauthammer, Samah Jamal Fodeh, Ali Haddad, Cynthia Brandt

Enhancing Clustering by Exploiting Complementary Data Modalities in the Medical Domain

Samah Jamal Fodeh[1], Ali Haddad[2], Cynthia Brandt[3], Martin Schultz[4], Michael Krauthammer[5]

[(1,3)] Yale University School of Medicine, [2] Department of Mathematics, [4] Department of Computer Science, [5] Department of Pathology, (Yale University, New Haven, CT, USA) {samah.fodeh, ali.haddad, cynthia.brandt, Schultz-martin, michael.krauthammer} @yale.edu

Abstract. Data Clustering has been an active area of research in many different application areas, with existing clustering algorithms mostly focusing on partitioning one modality or representation of the data. In this study, we delineate and demonstrate a new, enhanced data clustering approach whose innovation is its exploitation of multiple data modalities. We propose BI-NMF, a bi-modal clustering approach based on Non Negative Matrix Factorization (NMF) that clusters two differing data modalities simultaneously. The strength of our approach is its combining of multiple aspects of the data when forming the final clusters. To assess the utility of our approach, we performed several experiments on two distinct biomedical datasets with two modalities each. Comparing the clusters of BI-NMF with NMF clusters of single data modality, we observed consistent performance enhancement across both datasets. Our experimental results suggest that BI-NMF is advantageous for boosting data clustering.

Keywords: BI-NMF, clustering, biomedical, images, non negative matrix factorization.

1 Introduction

Clustering has been an active area of research in data mining and machine learning due to the rapidly growing data in different domains such as biology and clinical medicine. In biology, for instance, there is an avalanche of data from novel high throughput and imaging technologies. When applied to cancer images, clustering has been effective in identifying malignant and normal breast images [1]. Biomedical publications often present the results of biological experiments in figures and graphs that feature detailed, explanatory footnotes and captions. This annotation comprises a simple, textual representation of the images. In the clinical literature, a new semantic representation has evolved as a result of mapping the words in physicians' clinical notes to the corresponding semantic descriptors in the Unified Medical Language System (UMLS). Each representation of the data e.g. images, captions and semantic descriptors, is a unique data modality generated by a particular process wherein the objects have different features, structure and dimensionality. Differential encoding of the features of each modality causes variability in the obtained partitions when clustering around the individual data modality. In this discussion we explore alternative methods of building clusters around the complementary data modalities of a particular dataset to obtain more cohesive clusters. Unlike current algorithms which cluster on a single data modality, our proposed approach creates clusters by extracting information from completely different domains of information that describe the same data.

There have been recent efforts to perform multi-modal clustering. For example, Chen, Wang and Dong [10] proposed a co-clustering method using textual data that employs non negative matrix factorization (NMF) that draws from two data modalities: textual documents and their corresponding categories. Their method, however, is semi-supervised and requires user input to allow the algorithm to "learn" the distance metric. Comar, Tan and Jain [5] proposed the joint clustering of multiple social networks to identify cohesive communities characterized by reduced levels of noise. In this paper, we propose BI-NMF that combines information from two complementary data modalities to enhance clustering. NMF is a matrix factorization approach that has been shown to be effective for improving data clustering [6] as it produces meaningful clusters due to the non-negative nature of the solution. Specifically, NMF aims to factorize a data matrix into two non-negative matrices which are more compact (with lower dimensionality) and their product approximates the original matrix. One hopes that the new representation uncovers the hidden clusters in a given dataset. In this study, we cluster by drawing information from two different data matrices pertaining to complementary data modalities, thereby allowing us to exploit different aspects of the data while simultaneously reducing the distortion associated with clustering on a single modality. BI-NMF can be useful for any data described with multiple sources of information, i.e., modalities. We demonstrate our algorithm on two clinical datasets that each has information from two modalities. The first dataset contains images and their corresponding text captions and the second features textual notes reported by a clinical radiologist and their complementary semantic descriptors. The major contribution in this paper is the demonstration of a new method that simultaneously clusters two data modalities by jointly factorizing their corresponding matrices. The chief advantage of our method is enhanced clustering via the exploitation of information from complementary data modalities

The remainder of this paper is organized as follows. Section 2 presents the related work on clustering using NMF. Section 3 derives the proposed method along with the formal proofs. Section 4 presents the experimental results, followed by Section 5 featuring some concluding remarks.

2 Related Work

NMF has gained considerable attention recently in many domains such as pattern recognition and machine learning. Paatero and Tapper [6] proposed to use NMF algorithm to identify certain parts of objects like human faces. In a similar fashion, Xu, Liu and Gong used NMF to find clusters of documents [9]. They considered each dimension in the NMF space as one cluster and mapped a document *d* to the column *cluster* that has the maximum entry with *d*. As NMF performs learning in the Euclidean space, it fails to consider the intrinsic geometrical structure as suggested in [2], hence the authors extended NMF by imposing a new constraint that captures the geometrical representation of the data. Unlike previous methods which apply NMF to only one data modality, our proposed method aims to learn from two different modalities simultaneously. Reference [5] proposed to jointly cluster multiple networks using tri non-negative matrix factorization. Their updating rules, however, are different from ours since they minimized the KL-divergence metric in the cost function. In a similar context, the authors in [10] proposed a co-clustering method based on NMF that combines two modalities of the data. Their approach requires the user to provide input to learn a distance metric.

3 Methodology

BI-NMF is our proposed method for extracting information from two data modalities as a means of enhanced clustering. As our method is based on NMF, we describe NMF first and then discuss BI-NMF.

3.1 Non-negative matrix Factorization NMF

NMF [3] is a matrix factorization algorithm that deals with non-negative data matrices. Given a data matrix $X = [x_1, x_2, ...x_n] \in R^{(pxn)}$. NMF produces two non-negative matrices $\boldsymbol{U} \in R^{(pxk)}$ and $\boldsymbol{V} \in R^{(nxk)}$ as a result of minimizing the following objective function:

$$O = \|X - UV^T\|_F^2 \quad (1)$$

where $\|.\|_F$ denotes the matrix Frobenius norm. Lee and Seung [3] proposed an iterative approach using multiplicative rules to solve for *U* and *V*.

$$u_{ij}^{t+1} = u_{ij}^{t} \frac{(XV)_{ij}}{(UV^TU)_{ij}} \quad (2)$$

$$v_{ij}^{t+1} = v_{ij}^{t} \frac{(X^TU)_{ij}}{(VU^TU)_{ij}} \quad (3)$$

Each column in the original matrix *X* is a linear combination of the columns of *U* weighted by the components of the corresponding column in *V*. Therefore *U* can be regarded as containing a basis that is optimized for the linear approximation of the data in *X* [3]. It is proven by Lee and Seung that the objective function *O* in (1) is nonincreasing under the update rules (2) and (3).

3.2 BI-NMF

Our algorithm extends NMF using two modalities of the data. We argue that each modality covers certain aspects of the data, therefore utilizing two modalities maximizes the gained benefit and potentially improves the clusters. The two data modalities are represented by the matrices *A* and *B*. Let $A \in R^{(mxn)}$, $B \in R^{(pxn)}$, $U_1 \in R^{(mxk)}$, $U_2 \in R^{(pxk)}$, $V \in R^{(nxk)}$, we seek to approximate the new compact representation of the data by simultaneously factorizing *A* and *B*. BI-NMF minimizes the following objective function:

$$J = \|A - U_1V^T\|^2 + \|B - U_2V^T\|^2 \quad (4)$$

where *V* is anticipated to capture the agreement between *A* and *B* about the clusters. The objective function above can be rewritten as follows:

$$\begin{aligned} J &= (A - U_1V^T)(A - U_1V^T)^T + (B - U_2V^T) \\ &\quad (B - U_2V^T)^T \\ &= tr(AA^T) - 2tr(AVU_1^T) + tr(U_1V^TVU_1^T) + \\ &\quad tr(BB^T) - 2tr(BVU_2^T) + tr(U_2V^TVU_2^T) \end{aligned} \quad (5)$$

in the second step we used the matrix property $tr(XY)=tr(YX)$ and $tr(X)=tr(X^T)$. The objective function *J* needs to be solved under the constraints $u_1(i,j)>0$, $u_2(i,j)>0$ and $v(i,j)>0$. This is a typical constrained optimization problem that can be solved using

Lagrange multiplier method. Let $\alpha = [\alpha_{ij}]_{mxk}$, $\beta = [\beta_{ij}]_{pxk}$ and $\varphi = [\varphi_{ij}]_{nxk}$ be the Lagrange multipliers for the constraints $u_1(i,j)>0$, $u_2(i,j)> 0$ and $v(i,j)>0$, respectively. For notational convenience, we are using the same indices i and j even though the dimensions of U_1, U_2 and V are not necessarily the same. The Lagrange L is:

$$\begin{aligned} L \;=\; & tr(AA^T) - 2tr(AVU_1^T) + tr(U_1V^TVU_1^T) \\ & + tr(BB^T) - 2tr(BVU_2^T) + tr(U_2V^TVU_2^T) \\ & + tr(\alpha U_1^T) \;+\; tr(\beta U_2^T) \;+\; tr(\varphi V^T) \end{aligned} \tag{6}$$

the partial derivatives of the Lagrange function L with respect to U_1, U_2 and V are:

$$\frac{\partial L}{\partial U_1} \;=\; -2AV + 2U_1V^TV + \;\alpha \tag{7}$$

$$\frac{\partial L}{\partial U_2} \;=\; -2BV + 2U_2V^TV + \;\beta \tag{8}$$

$$\begin{aligned} \frac{\partial L}{\partial V} \;=\; & -2A^TU_1 + 2VU_1^TU_1 - \\ & 2B^TU_2 + 2VU_2^TU_2 + \varphi \end{aligned} \tag{9}$$

Solving with respect to α, β, φ and utilizing the Kuhn-Tucker conditions $\alpha_{ij}u_{1(i,j)} = 0$, $\beta_{ij}u_{2(i,j)} = 0$, and $\varphi_{ij}v_{(i,j)} = 0$, we get the following equations:

$$-(AV)_{(i,j)}u_{1(i,j)} + (U_1V^TV)_{(i,j)}u_{1(i,j)} = 0 \tag{10}$$

$$-(BV)_{(i,j)}u_{2(i,j)} + (U_2V^TV)_{(i,j)}u_{2(i,j)} = 0 \tag{11}$$

$$-(A^TU_1)v_{(i,j)} - (B^TU_2)v_{(i,j)} + (VU_1^TU_1)v_{(i,j)} + (VU_2^TU_2)v_{(i,j)} = 0 \tag{12}$$

after rearranging the last 3 equations we obtain the following update rules:

$$u_{1(i,j)} = u_{1(i,j)}\,\frac{(AV)_{(i,j)}}{(U_1V^TV)_{(i,j)}} \tag{13}$$

$$u_{2(i,j)} = u_{2(i,j)}\,\frac{(BV)_{(i,j)}}{(U_2V^TV)_{(i,j)}} \tag{14}$$

$$v_{(i,j)} = v_{(i,j)}\,\frac{(A^TU_1 + B^TU_2)_{(i,j)}}{(VU_1^TU_1 \;+\; VU_2^TU_2\,)_{(i,j)}} \tag{15}$$

The objective function J in (4) is nonincreasing under the update rules in (13) (14) (15) (see appendix). The update rules of U_1, U_2 and V converge and the final solution is a local minimum. Lee and Seung [3] used an auxiliary function to prove the convergence of (1); which essentially minimizes a distance function. Equation (4), however, is the summation of two distance functions. Following the steps in [5] show that minimizing the auxiliary function of the summation is sufficient to decrease the objective function of the sum of distances. The matrix V computed in (15) is used to define the clusters as proposed by [9]. Each column $V_{.j}$ corresponds to one cluster and each row $V_{i.}$ to a point. A point is assigned to the cluster associated to the maximum value in its corresponding row. Formally, assign x_i to cluster c if $c = arg\max_j V_{ij}$. Note that the clusters in V are computed by joining

information from two data modalities represented by the matrices *A* and *B*. It is important to mention that we normalized the matrices *A* and *B* using *TFIDF*. Further, we rescaled both matrices using the following formula:

$$X = X * [\text{diag}(X^T X e)]^{-1/2} \quad (16)$$

where *X* is a data matrix and *e* is a unit vector. Transforming the matrices using (16) before applying BI-NMF was proposed in [9]. We noticed that this transformation helped improve the clustering results. The pseudo code of our algorithm is summarized below.

Algorithm 1 BI-NMF

Input: data modality A, data modality B, maximum number of iterations I_{max}. Clusters C.

1. **Initialize U_1^t, U_2^t, V^t , normalize A, B** using (16)
2. **for** $t = 1$ **to** I_{max} **do**
   - compute u_1^{t+1} using (13) , u_2^{t+1} using (14), v^{t+1} using (15)
   - set $u_1^t = u_1^{t+1}$, $u_2^t = u_2^{t+1}$, $v^t = v^{t+1}$
3. **end**
4. **C = AssignClusters(V)**

---

4 Experimental evaluation

We evaluated the proposed algorithm on two biomedical datasets. We demonstrate the effectiveness of BI-NMF by comparing its output clusters with the two NMF clustering solutions of each individual data modality, and with the NMF clusters of the two modalities merged. In the latter method, classic NMF [3] is applied to the merged matrices *A* and *B* after normalizing using *TFIDF*. We also compare BI-NMF with the two ensemble clusters computed for each individual data modality and with the combined ensemble clustering proposed in [8]. Combined ensemble clustering is fundamentally based on combining two data modalities using ensemble clustering. In this method, the co-association matrices are generated for each individual data modality and subsequently combined into one co-association matrix whereupon k-means is applied to obtain the consensus clustering. We also report the clusters of each data modality based on k-means.

4.1 Datasets

Pubmed Images Dataset. It consists of 3000 images extracted from articles of PubMed Central. Images with no captions were dropped and 2607 were retained. The images in the dataset were classified into 5 different categories by domain expert annotators. Discrepancies among the annotators were resolved by assigning the image to the category with the majority of votes. The list of annotations is: 564 images were assigned to the experimental category, 1131 images to the graph category, 645 images to the diagrams category, 86 images to the clinical category, and 181 images were assigned to the others category. We generated two modalities for the images. In one modality the images were represented using the pictorial and textural features computed using the Haralick method [7]. The other modality is a Bag of Words *BOW* representation generated using captions.

Radiology Reports Dataset. It consists of radiology reports collected from clinical records of patients for research purposes. The radiology reports were annotated by domain experts and classified into four categories. The categories and the counts of their content reports are: 35 abdominal MRI reports, 486 abdominal CT reports, 248 abdominal ultrasound reports and 500 non-abdominal radiology reports. For simplicity, we will call these MRI, CT, Ultrasound, and non-abdominal, respectively. The reports are represented using two data modalities: Textual features BOW and Bag of Concepts (*BOC*). In the BOW modality, the reports are represented using the original words that appear in the clinical narratives and weighted using their *TFIDF* score. In the *BOC* modality, the vectors are indexed by semantic concepts derived from cTAKES [4], a natural language processing tool that maps text to concepts from the *UMLS* ontology.

4.2 Evaluation metrics

The clustering results are evaluated by comparing to gold standard annotations of images and radiology reports. We use three measures to evaluate the quality of the clusters: micro-averaged precision, purity and Normalized Mutual Information (*NMI)*. Micro-averaged precision is an average over data points, which by default gives higher weight to those classes with many data points. *NMI* measures the amount of information by which our knowledge about the classes increases upon definition of the clusters.

$$micro\ averaged\ precision = \frac{\sum_{i=1}^{k} TP_i}{\sum_{j=1}^{k} TP_j + FP_j}$$
$$purity = \sum_{i} \frac{|C_i|}{n} max(precision(C_i, L_j)) \quad (17)$$
$$NMI = \frac{I(X;Y)}{log\ k + log\ c}$$

where *TP* is true positive, *FP* is false positive, *n* is the number of data points, *k* is the number of clusters, *c* is the number of classes, C_i is the i^{th} cluster, L_j is the j^{th} class, *I(X;Y)* is the mutual information between two random variables *X* (the cluster) and *Y* (the class).

4.3 Single Modality Clustering: BI-NMF vs NMF, k-means and Ensemble Clustering

We compare the clustering solutions produced by BI-NMF which draws information from different data modalities with the output clusters obtained using single data modality in order to demonstrate the benefit of leveraging multiple representations of the data. We show the performance of regular NMF on single modalities, along with comparable approaches such as k-means and ensemble clustering [8]. In ensemble clustering, a number of clustering solutions are aggregated in a co-association matrix that measures the number of times each pair of data points are placed into the same cluster. K-means is applied to the co-association matrix to get the final clusters. Table 1 shows a comparison between the performances of several clustering methods on single data modalities: K-means clusters of each data modality, the cluster ensembles of each data

modality and NMF applied to each individual data modality. For the sake of clarity, the method descriptor has two parts: the applied method and the data modality used. For radiology reports, we observed that the ensemble clustering method applied to one data modality performed poorly when compared to NMF of single data modality, while outperforming single-modality k-means. With the exceptions discussed below, BI-NMF clusters were significantly better than single modality NMF, single modality ensemble clusters and k-means clusters as shown in Table 1 vs Table 2. It is important to mention that for the Pubmed images data, the clusters of k-means for the captions modality yield comparative clusters to BI-NMF based on purity as shown in Table 1. Nevertheless, *NMI* and micro averaged precision measures suggest that BI-NMF clusters are better than k-means clusters. To further assure this result, we computed the average of 100 BI-NMF runs and got consistent results. This result strongly emphasizes the benefit of our method that draws information from two data modalities.

Table 1: One data modality: Performance of different clustering methods of each data modality

Data	Method Descriptor	Micro Avg Precision	Purity	NMI
Radiology Reports	k-means_words	0.506	0.639	0.240
	Ensemble_words	0.506	0.640	0.238
	NMF_words	**0.676**	**0.791**	**0.599**
	k-means_concepts	0.555	0.758	0.490
	Ensemble_concepts	0.581	0.764	0.503
	NMF_concepts	**0.665**	**0.884**	**0.787**
Pubmed Images	k-means_Haralick	0.318	0.505	0.141
	Ensemble_Haralick	0.306	0.513	**0.150**
	NMF_Haralick	**0.331**	**0.516**	0.145
	k-means_captions	0.456	**0.558**	**0.180**
	Ensemble_captions	**0.479**	0.519	0.153
	NMF_captions	0.445	0.518	0.134

4.4 Two modality clusterting: BI-NMF vs NMF_merged and Combined Ensemble Clustering

To assess the effectiveness of BI-NMF, we compared its performance against another bi-modality clustering approach called combined ensemble clustering. In combined ensemble clustering, two co-association matrices are generated from two data modalities then linearly combined into one co-association matrix upon which k-means is applied to obtain the final clusters. We also compare the output clusters of our method with the clusters obtained when applying NMF to the merged data modalities. We implemented the combined ensemble clustering algorithm in [8] and applied it to our biomedical datasets. Table 2 shows a comparison in performance between NMF_merged, combined ensemble clustering and BI-NMF for radiology reports data and PubMed images data. Recall that in the NMF_merged method the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ pertaining to both data modalities are first combined and *NMF* is subsequently applied to the combined matrix after normalization. The performance of the two methods depends on their respective emphases on forming the BI-NMF clusters from various modalities versus combining different features of the data modalities prior to the formation of clusters.

Table 2: Two data modalities: Performance of different clustering methods for both modalities

Data	Method Descriptor	Micro Avg Precision	Purity	NMI
Radiology Reports	NMF_merged	0.584	0.793	0.599
	Combined Ensemble Clustering	0.582	0.761	0.513
	BI-NMF	**0.777**	**0.903**	**0.825**
Pubmed Images	NMF_merged	0.367	0.461	0.119
	Combined Ensemble Clustering	0.483	0.542	0.190
	BI-NMF	**0.551**	**0.558**	**0.200**

The quality of the clusters obtained by BI-NMF was superior compared to that of combined ensemble clustering and NMF_merged for both datasets in terms of all reported measures. On radiology reports, compared to combined ensemble clustering, BI-NMF achieved a relative improvement of the order of 33%, 18% and 60% in terms of micro averaged precision, purity and NMI, respectively. Similarly, it outperformed NMF_merged and yield a better clustering solution with a difference of 32%, 13% and 38% in terms of micro averaged precision, purity and NMI, respectively. BI-NMF also outperformed combined ensemble clustering and NMF_merged for Pubmed images as shown in Table 2. The micro averaged precision reported for BI-NMF was .551 compared to .483 for combined ensemble clustering. Likewise, purity and NMI showed a relative improvement of 3% and 5%, respectively. Superior performance is also observed for the proposed method compared to NMF_merged, it yield a relative improvement of 50%, 21% and 68% in terms of micro averaged precision, purity and NMI, respectively.

5 Conclusion

In this paper, we demonstrate an enhanced data clustering approach whose innovation is its exploitation of multiple data modalities called BI-NMF. Our proposed method is a bi-modal clustering algorithm based on non negative matrix factorization. It utilizes two modalities of the data to improve clustering. We applied the method on two biomedical datasets and demonstrated enhanced performance relative to ensemble clustering and NMF based on single and merged data modalities, on three standard metrics. Given our results, we conclude that BI-NMF is advantageous for enhanced biomedical data clustering and potentially useful for data from other domains.

Acknowledgements. This study was funded by NIH/NLM 5R01LM009956 (MK), and a VA grant HIR 08-374/HSR&D: Consortium for Healthcare Informatics (CB,SF,MK).

References

1. B. Chandra, S. Nath, A Mlhothra, Classification and Clustering of Cancer Images. The 6th International Joint Conference on Neural Networks, 3843-3847 (2006)
2. D. Cai, X. He, X. Wang, H. Bao, and J. Han: Locality preserving non-negative matrix factorization, Proc. 27th annual inte'l ACM SIGIR, 96-103 (2004)

3. D. D. Lee, and H. S. Seung, Algorithms for non-negative matrix factorization, Advances in neural information processing systems, (13) (2001)
4. G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, Journal AMIA, (17):507 (2010)
5. P. Mandayam-Comar, P. N. Tan, A.K. Jain, Identifying Cohesive Subgroups and Their Correspondences in Multiple Related Networks, (1), 476-483, WI-IAT (2010)
6. P. Paatero and U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5(2):111-126 (1994)
7. R. M. Haralick, Statistical and structural approaches to texture, IEEE, 67:786-804 (1979)
8. S. J. Fodeh, W. F. Punch, and P. N. Tan, Combining statistics and semantics via ensemble model for document clustering, ACM symposium on Applied Computing, 1446-1450 (2009)
9. W. Xu, X. Liu, and Y. Gong, Document clustering based on non-negative matrix factorization, " Proc. 26th annual int'l ACM SIGIR, 267-273 (2003)
10. Y. Chen, L. Wang, and M. Dong, Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering, TKDE, 1459-1474 (2009)

Appendix

Theorem 1. The objective function J is non-increasing under the rules (13) (14) (15).
The proof follows the one given by Lee and Sung [3] since update rules for U_1 and U_2 do not change. For the update rule (15), we use the auxiliary function trick.

Definition 1. $G(v,v^t)$ is an auxiliary function for $F(v)$ if the following are satisfied:

$$G(v, v^t) \geq F(v), \qquad G(v, v) = F(v) \tag{18}$$

Lemma 1. If $G_1(v, v^t)$ and $G_2(v, v^t)$ are auxiliary functions for $F_1(v)$ and $F_2(v)$ respectively, then: (a) $G(v, v^t) = G_1(v, v^t) + G_2(v, v^t)$ is the auxiliary function for $F(v) = F_1(v) + F_2$, (b) $F(v)$ is non-increasing under the update:

$$v^{t+1} = arg \min_v G(v, v^t) \tag{19}$$

The proof of (a) is trivial, for (b) we have:

$$\begin{aligned} F(v^{t+1}) &= F_1(v^{t+1}) + F_2(v^{t+1}) \\ &\leq G_1(v^{t+1}, v^t) + G_2(v^{t+1}, v^t) \\ \leq G_1(v^t, v^t) + G_2(v^t, v^t) &= F_1(v^t) + F_2(v^t) = F(v^t) \end{aligned} \tag{20}$$

Note that the third line is a result of the fact that v^{t+1} minimizes the auxiliary function G, then $G(v^{t+1}, v^t) \leq G(v^t, v^t)$ and $F(v^{t+1}) \leq F(v^t)$ as shown in [5]. To conclude the proof of *Theorem 1*, we show that the update rule (15) is the update given by (19), i.e.

$$v_{(i,j)}^{t+1} = arg \min_v G(v, v_{(i,j)}^t) \tag{21}$$

for a suitable auxiliary function $G(v, v^t)$. The objective function of eq.(5) can be written:

$$J = \sum_{i,j} F_{i,j}(v_{i,j}) \tag{22}$$

where $F_{i,j}$ is a quadratic function that depends only on $v_{i,j}$, the generic term of the matrix V. We need to show that the function $F_{i,j}$ is non-increasing under the update rule (15), or equivalently find an auxiliary function for $F_{i,j}$ such that the update rule (15) corresponds to (21). We compute the first and second order derivatives of $F_{i,j}$. One can easily check that:

$$F'_{i,j} = 2(-A^T U_1 - B^T U_2 + V(U_1^T U_1 + U_2^T U_2))_{i,j} \tag{23}$$

$$F''_{i,j} = 2(U_1^T U_1 + U_2^T U_2)_{j,j} \tag{24}$$

Then we consider:

$$G(v, v_{i,j}^t) = F_{i,j}(v_{i,j}^t) + F'_{i,j}(v_{i,j}^t)(v - v_{i,j}^t) + \frac{(V(U_1^T U_1 + U_2^T U_2))_{i,j}}{v_{i,j}^t}(v - v_{i,j}^t)^2 \tag{25}$$

now we need to show that $G(v, v_{i,j}^t)$ corresponds to an auxiliary function for $F_{i,j}$:

It is obvious that $G(v_{i,j}^t, v_{i,j}^t) = F_{i,j}(v_{i,j}^t)$. We only need to show $G(v_{i,j}, v_{i,j}^t) \geq F_{i,j}(v_{i,j}^t)$. Since $F_{i,j}$ is a quadratic form, consider the following Taylor series for $F_{i,j}$:

$$F_{i,j}(v) = F_{i,j}(v_{i,j}^t) + F'_{i,j}(v_{i,j}^t)(v - v_{i,j}^t) + \frac{1}{2}F''_{i,j}(v_{i,j}^t)(v - v_{i,j}^t)^2 \tag{26}$$

we need to show that:

$$\frac{(V(U_1^T U_1 + U_2^T U_2))_{i,j}}{v_{i,j}^t} \geq (U_1^T U_1 + U_2^T U_2)_{j,j} \tag{27}$$

the inequality (27) is obvious since

$$\frac{(V(U_1^T U_1))_{i,j}}{v_{i,j}^t} = \sum_k \frac{v_{i,k}^t}{v_{i,j}^t}(U_1^T U_1)_{k,j} \geq (U_1^T U_1)_{j,j} \tag{28}$$

and the same inequality holds for U_2:

$$\frac{(V(U_2^T U_2))_{i,j}}{v_{i,j}^t} = \sum_k \frac{v_{i,k}^t}{v_{i,j}^t}(U_2^T U_2)_{k,j} \geq (U_2^T U_2)_{j,j} \tag{29}$$

Thus $G(v_{i,j}, v_{i,j}^t) \geq F_{i,j}(v_{i,j}^t)$. We conclude the proof of *Theorem 1* by checking that (21) corresponds to (15). Indeed, given (22) and (26), we can get by solving $G'(v_{i,j}, v_{i,j}^t) = 0$.

$$v_{i,j}^{t+1} = v_{i,j}^t\left(1 - \frac{F'_{i,j}(v_{i,j}^t)}{2(V(U_1^T U_1 + U_2^T U_2))_{i,j}}\right) \tag{30}$$

After arranging the equation, one can easily show that (30) is equivalent to (15).