CLUSTERING BY ORTHOGONAL NON-NEGATIVE MATRIX FACTORIZATION: A SEQUENTIAL NON-CONVEX PENALTY APPROACH

Shuai Wang[†], *Tsung-Hui Chang*[†⋆], *Ying Cui*[‡] *and Jong-Shi Pang*[‡]

[†]School of Science & Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China
[⋆]Shenzhen Research Institute of Big Data, Shenzhen 518172, China
[‡]Dept. of Industrial and Systems Engineering, University of Southern California, CA 90089, U.S.A.

ABSTRACT

The non-negative matrix factorization (NMF) model with an additional orthogonality constraint on one of the factor matrices, called the orthogonal NMF (ONMF), has been found to provide improved clustering performance over the K-means. The ONMF model is a challenging optimization problem due to the orthogonality constraint, and most of the existing methods directly deal with the constraint in its original form via various optimization techniques. In this paper, we propose an equivalent problem reformulation that transforms the orthogonality constraint into a set of norm-based non-convex equality constraints. We then apply a penalty approach to handle these non-convex constraints. The penalized formulation is smooth and has convex constraints, which is amenable to efficient computation. We analytically show that the penalized formulation will provide a feasible stationary point of the reformulated ONMF problem when the penalty is large. Numerical results show that the proposed method greatly outperforms the existing methods.

Index Terms— Clustering, orthogonal NMF, penalty method.

1. INTRODUCTION

Clustering is one of the most fundamental data mining tasks and has an enormous number of applications [1]. Among the existing clustering methods, the K-means [2] is the most widely used one, thanks to its simplicity [3]. However, the K-means may not always yield satisfactory clustering results. On one hand, from an optimization perspective, the iterative steps of finding the cluster centroids and cluster assignment in K-means are equivalent to solving a binary constrained matrix factorization problem [4, 5] by alternating optimization. Due to the non-convex matrix factorization model and binary constraint, the iterates of K-means are likely to be stuck in an unsatisfactory local point and are sensitive to the choice of initial points [6]. On the other hand, the K-means overlooks the inherent low-rank structure and prior information which are usually owned by high-dimensional real data. Therefore, various dimension-reduction techniques such as principal component analysis (PCA), spectral clustering [7], non-negative matrix factorization (NMF) [8, 9] and deep neural networks [10, 11] are proposed for improving data clustering performance. However, these methods are merely used as a preprocessing stage to find a low-dimensional data representation, and the K-means is still used for clustering the dimension-reduced data. Thus, the intrinsic drawback of the K-means caused by the non-convex and discrete nature of data clustering is not addressed.

Recently, as an variant of NMF, the orthogonal NMF (ONMF) model has been considered for data clustering [12–18]. The ONMF model imposes an additional orthogonality constraint on one of the factor matrices in NMF. It turns out that, like the K-means, the orthogonally constrained factor matrix functions the same as an indicator matrix that shows how the data samples are assigned to different clusters [12, 16]. Therefore, the ONMF model can be regarded as a continuous formulation (which has no discrete constraint but is still non-convex) of K-means. Studies on various data mining tasks have found that the ONMF model can outperform the K-means and NMF based clustering methods [13, 15–20].

The ONMF model is challenging to solve. Many of the existing ONMF algorithms extend upon the classical multiplicative rule [8] to accommodate the additional orthogonality constraint. For example, reference [12] used a penalty approach for fulfilling the orthogonality constraint followed by applying the multiplicative rule. The authors of [14] derived the multiplicative rule directly using the gradient vector in Stiefel manifold. Besides, the work [16] employed the augmented Lagrangian method to deal with the non-negative constraint and applied the gradient projection method for an orthogonally constrained subproblem. Reference [15] proposed a hierarchical alternating least squares method that updates the row vectors of the factor matrix one by one subject to orthogonality constraints with the other rows. Reference [17] solves a sequence of non-negative PCA problems for finding good candidates for the ONMF model. The procedure however is computationally inefficient.

In this paper, we propose a new optimization framework, which we call the sequential non-convex penalty (SNCP) method, for handling the ONMF problem. The first ingredient of the SNCP method is to equivalently reformulate the matrix orthogonality constraint as a set of squared ℓ_1-norm-minus-ℓ_2-norm equality constraints. The second ingredient is to apply the penalty method [21] which adds these non-convex smooth constraints as penalty terms in the objective function, leaving only simple convex constraints. Then, the proximal alternating linearized minimization (PALM) method in [22] is employed to compute a stationary point of the penalized non-convex optimization problem. In addition, we study analytical conditions for which the proposed SNCP method can yield a feasible stationary point to the reformulated ONMF problem. Numerical results based on synthetic and real datasets show that the proposed SNCP method outperforms existing ONMF methods.

2. CLUSTERING AND ORTHOGONAL NMF

Let $\mathbf{X} \in \mathbb{R}_+^{M \times N}$ be a non-negative data matrix that contains N data samples, and each of the samples has M features. The task of data clustering is to assign the N data samples into a predefined number of K clusters in the sense that the samples belonging to one cluster are close to each other based on certain distance metric. The most popular setting is to consider the Euclidean distance as the distance metric and the use of the K-means for data clustering.

The work is supported by the NSFC, China, under Grant 61571385 and 61731018, and by the Shenzhen Fundamental Research Fund under Grant ZDSYS201707251409055 and KQTD2015033114415450.

It is known [4, 5] that, from an optimization point of view, the K-means can be interpreted as an alternating optimization algorithm applied to the following matrix factorization problem

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

$$\text{s.t. } \|\mathbf{h}_j\|_1 = 1,\ [\mathbf{H}]_{ij}\in\{0,1\},\ \forall i\in\mathcal{K}, j\in\mathcal{N}, \tag{1b}$$

$$\mathbf{W}\geq 0,\ \mathbf{H}\geq 0, \tag{1c}$$

where $\mathcal{K}\triangleq\{1,\ldots,K\}$, $\mathcal{N}\triangleq\{1,\ldots,N\}$, $\|\cdot\|_F$ and $\|\cdot\|_1$ are the matrix Frobenius norm and vector 1-norm, respectively, $[\mathbf{H}]_{ij}$ is the (i,j)th entry of $\mathbf{H}$, and $\mathbf{W}\geq 0$ (resp. $\mathbf{H}\geq 0$) stands for that all elements of $\mathbf{W}$ (resp. $\mathbf{H}$) are non-negative. Here, columns of $\mathbf{W}\in\mathbb{R}^{M\times K}$ represent centroids of the K clusters. The matrix

$$\mathbf{H} = \big[\mathbf{h}_1,\ldots,\mathbf{h}_N\big] = \Big[\widetilde{\mathbf{h}}_1,\ldots,\widetilde{\mathbf{h}}_K\Big]^T \in\mathbb{R}^{K\times N}, \tag{2}$$

indicates the cluster assignment of samples. Specifically, under the clustering constraint (1b), each column of $\mathbf{H}$ has only one non-zero element; in particular, $[\mathbf{H}]_{ij}=1$ if the jth sample is uniquely assigned to cluster i, and $[\mathbf{H}]_{ij}=0$ otherwise. One can see from (1) that, when $\mathbf{W}$ is given, the optimal $\mathbf{H}$ is obtained by assigning each sample to the cluster that has the nearest centroid, and when $\mathbf{H}$ is given, the optimal $\mathbf{W}$ is given by the centroids of K clusters. The two steps are exactly the well-known K-means algorithm. However, due to the non-convex binary constraint (1b), the K-means is sensitive to the initial conditions and may not always yield satisfactory clustering performance. Therefore, there have been efforts to finding a better initial point for the K-means; see, e.g., the K-means++ [6].

The orthogonal NMF (ONMF) model proposed in [12] can provide better clustering performance over the K-means. In particular, the ONMF problem is given by

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_{\mathrm{F}}^2 \tag{3a}$$

$$\text{s.t. } \mathbf{W}\geq 0, \mathbf{H}\geq 0, \tag{3b}$$

$$\mathbf{H}\mathbf{H}^T = \mathbf{I}_K, \tag{3c}$$

where $\mathbf{I}_K$ is the K by K identity matrix. It has been shown [12, 16] that the ONMF model (3) is closely related to the K-means problem (1). Specifically, the orthogonality constraint $\mathbf{H}\mathbf{H}^T=\mathbf{I}_K$, together with the non-negativity constraint $\mathbf{H}\geq 0$, enforce each column of $\mathbf{H}$ to have at most one non-zero entry. Thus, matrix $\mathbf{H}$ in (3) functions similarly as that in (1) and indicates the cluster assignment of data samples. Nevertheless, different from (1), the non-zero entries of $\mathbf{H}$ in (3) are not limited to be one but can be scaled. Owing to the two facts, the ONMF model may outperform the K-means and other clustering methods that rely on the vanilla NMF model [5, 9, 16].

However, due to the orthogonality constraint, the ONMF problem (3) is challenging to solve. Unlike the existing methods [12, 14–16] which directly deal with the orthogonality constraint (3c), we present a novel problem reformulation of (3) and propose an SNCP method that is not only amenable to efficient computation but also able to provide promising clustering performance.

3. PROPOSED METHOD

3.1. Problem Reformulation

As mentioned, the orthogonality constraint (3c) and the non-negative constraint (3b) imply that each column of $\mathbf{H}$ has at most one non-zero element. Since any vector $\boldsymbol{x}\in\mathbb{R}^n$ has at most one non-zero entry if and only if $\|\boldsymbol{x}\|_1=\|\boldsymbol{x}\|_2$, the constraint set (3c) and (3b) for $\mathbf{H}$ is equivalent to the following set

$$\left\{\mathbf{H}\in\mathbb{R}^{K\times N}\ \middle|\ \begin{array}{l}\mathbf{H}\geq 0,\\ \|\widetilde{\mathbf{h}}_i\|_2 = 1,\ i\in\mathcal{K},\\ \|\mathbf{h}_j\|_1 = \|\mathbf{h}_j\|_2,\ j\in\mathcal{N}\end{array}\right\}. \tag{4}$$

Firstly, under $\mathbf{H}\geq 0$, $\|\mathbf{h}_j\|_1=\|\mathbf{h}_j\|_2, j\in\mathcal{N}$ are the same as

$$(\mathbf{1}^T\mathbf{h}_j)^2 = \|\mathbf{h}_j\|_2^2,\ \forall j\in\mathcal{N}, \tag{5}$$

where $\mathbf{1}$ is the all-one vector. Secondly, note that the condition $\|\widetilde{\mathbf{h}}_i\|_2=1,\ i\in\mathcal{K}$, is not intrinsic to the data clustering task. In essence, both $\mathbf{H}$ and $\mathbf{Q}\mathbf{H}$, where $\mathbf{Q}\geq 0$ is a diagonal matrix, indicate the same cluster assignment, and both $(\mathbf{W},\mathbf{H})$ and $(\mathbf{W}\mathbf{Q}^{-1},\mathbf{Q}\mathbf{H})$ have the same objective values in (1a). Therefore, without loss of the clustering performance, we replace $\|\widetilde{\mathbf{h}}_i\|_2=1,\ i\in\mathcal{K}$, by

$$\|\widetilde{\mathbf{h}}_i\|_2\leq 1,\ \forall i\in\mathcal{K}. \tag{6}$$

By combing (4), (5) and (6), we have the following problem formulation for data clustering:

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2 \tag{7a}$$

$$\text{s.t. } (\mathbf{1}^T\mathbf{h}_j)^2 = \|\mathbf{h}_j\|_2^2,\ j\in\mathcal{N}. \tag{7b}$$

$$\|\widetilde{\mathbf{h}}_i\|_2\leq 1,\ i\in\mathcal{K}, \tag{7c}$$

$$\mathbf{W}\geq 0, \mathbf{H}\geq 0. \tag{7d}$$

Problem (7) is still difficult to solve. In particular, the constraints in (7b) are non-convex. Moreover, constraints in (7b) and in (7c) couple elements of $\mathbf{H}$ across rows and across columns, respectively. This makes it difficult to apply some decomposition methods for dealing with large-scale problems.

3.2. Proposed SNCP Method

To overcome the aforementioned issues, we propose the following penalized formulation

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2 + \frac{\rho}{2}\sum_{j=1}^N\Big((\mathbf{1}^T\mathbf{h}_j)^2 - \|\mathbf{h}_j\|_2^2\Big) \tag{8a}$$

$$\text{s.t. } \|\widetilde{\mathbf{h}}_i\|_2\leq 1,\ i\in\mathcal{K}, \tag{8b}$$

$$\mathbf{W}\geq 0, \mathbf{H}\geq 0, \tag{8c}$$

where $\rho>0$ is a penalty parameter. Notice that in (8), the non-convex constraints in (7b) are penalized in the objective function, which leads to a simple convex constraint set for problem (8). As will be introduced shortly, by applying the PALM method [22], one is able to solve (8) efficiently.

Here, let us analyze the conditions for which the penalized formulation (8) can yield a feasible solution to problem (7). We first assume that one is able to achieve a local minimum solution of the penalized problem (8).

Proposition 1 *For any $\rho>0$, if $(\mathbf{W}^\star,\mathbf{H}^\star)$ is a local minimizer of problem* (8)*, then $\mathbf{H}^\star$ satisfies* (7b)*.*

We omit the proof here. The idea of the proof is based on the observation that for any $\mathbf{H}$ that is not feasible to (7b), the objective value in (8a) at $(\frac{1}{\alpha}\mathbf{W}^\star,\alpha\mathbf{H}^\star)$ is strictly increasing in $\alpha>0$. While Proposition 1 shows that any local minimum solution of (8) is feasible to (7), problem (8) is non-convex and therefore a local minimum solution cannot be computed in general. Instead, most of the non-convex optimization algorithms can only yield a stationary point [23, (1.3.3)] under proper conditions [24].

Let us denote $(\mathbf{W}^{\rho}, \mathbf{H}^{\rho})$ as a stationary point of (8), and assume $(\mathbf{W}^{\rho}, \mathbf{H}^{\rho}) \rightarrow (\mathbf{W}^{\infty}, \mathbf{H}^{\infty})$ when $\rho \rightarrow \infty$. By assuming that the Mangasarian-Fromovitz constraint qualification (MFCQ) [23, Sec. 3.2] holds for (7) at $(\mathbf{W}^{\infty}, \mathbf{H}^{\infty})$, we have the following theorem.

Theorem 1 *Assume that $(\mathbf{W}^{\rho}, \mathbf{H}^{\rho})$ is bounded and it has a limit point $(\mathbf{W}^{\infty}, \mathbf{H}^{\infty})$ when $\rho \rightarrow \infty$. Then, (a) $(\mathbf{W}^{\infty}, \mathbf{H}^{\infty})$ satisfies* (7b)*; (b) if the MFCQ holds for* (7) *at $(\mathbf{W}^{\infty}, \mathbf{H}^{\infty})$, then $(\mathbf{W}^{\infty}, \mathbf{H}^{\infty})$ is a stationary point of problem* (7)*.*

The proof is omitted here and will be reported in future publications. Theorem 1 implies that, by gradually increasing the penalty parameter ρ, any stationary point of the penalized problem (8) will eventually be feasible to (7) and it will also be a stationary point of (7) if the MFCQ holds. We should emphasize that Theorem 1 is different from the known results of the penalty methods [21, 24–27] where they all assume that either a local minimum or a global minimum of the penalized problem can be obtained.

Motivated by Theorem 1, we propose the SNCP method in Algorithm 1 for solving problem (7). In particular, we solve a sequence of the penalized problem (8) to a stationary point, with increased value of the penalty parameter ρ. To speed up the convergence, the obtained stationary point is used as the initial point for solving (8) in the next iteration.

Algorithm 1 Proposed SNCP method for solving (7).

1: **Set** $r = 0$, and given a parameter $\gamma > 1$, and initial values of $\rho > 0$ and $(\mathbf{W}^{(0)}, \mathbf{H}^{(0)})$.
2: **repeat**
3: Obtain a stationary point $(\mathbf{W}^{(r+1)}, \mathbf{H}^{(r+1)})$ of problem (8), e.g., by the PALM algorithm (Algorithm 2) with $(\mathbf{W}^{(r)}, \mathbf{H}^{(r)})$ being the initial point.
4: Update $\rho = \gamma\rho$.
5: Set $r = r + 1$.
6: **until** a predefined stopping criteria is satisfied.

3.3. Obtaining a Stationary Point of (8) by PALM

We employ the PALM method [22] to obtain a stationary point of problem (8). When applied to (8), the PALM involves performing two gradient projection steps for $\mathbf{H}$ and $\mathbf{W}$, respectively. Let us denote $F_{\rho}(\mathbf{W}, \mathbf{H}) \triangleq \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \frac{\rho}{2}\sum_{j=1}^{N}\left((\mathbf{1}^T\mathbf{h}_j)^2 - \|\mathbf{h}_j\|_2^2\right)$. At each iteration k of the PALM method, the gradient projection step for variable $\mathbf{H}$ is given by

$$\mathbf{H}^{k+1} = \arg\min_{\mathbf{H}} \|\mathbf{H} - \mathbf{B}^k\|_F^2 \quad \text{s.t. } \mathbf{H} \geq 0, \|\widetilde{\mathbf{h}}_i\|_2 \leq 1, \ i \in \mathcal{K}, \tag{11}$$

where $\mathbf{B}^k \triangleq \mathbf{H}^k - \frac{1}{t^k}\nabla_{\mathbf{H}} F_{\rho}(\mathbf{W}^k, \mathbf{H}^k)$ and $t^k > 0$ is a step size. Note that both the objective function and the constraint set of (11) are separable with respect to the rows of $\mathbf{H}$. Thus, update of $\mathbf{H}$ can be decomposed as K subproblems as in (9) (see Algorithm 2), where $(\widetilde{\mathbf{b}}_i^k)^T$ is the ith row of $\mathbf{B}^k$. The following proposition shows that (9) admits simple solutions:

Proposition 2 *Consider the following problem*

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{x} - \mathbf{b}\|_2^2 \tag{12a}$$

$$\text{s.t. } \boldsymbol{x} \geq 0, \|\boldsymbol{x}\|_2 \leq 1. \tag{12b}$$

where $\mathbf{b} = [b_1, \ldots, b_n]^T \in \mathbb{R}^n$ is given. Denote $\boldsymbol{x}^{\star} = [x_1^{\star}, \ldots, x_n^{\star}]^T$ as an optimal solution to (12).

Algorithm 2 PALM for solving (8).

1: **Set** $k = 0, \mathbf{W}^0 = \mathbf{W}^{(r)}, \ \mathbf{H}^0 = \mathbf{H}^{(r)}$.
2: **repeat**
3: Update $\widetilde{\mathbf{h}}_i, i \in \mathcal{K}$, and $\mathbf{W}$ by

$$\widetilde{\mathbf{h}}_i^{k+1} = \arg\min_{\widetilde{\mathbf{h}}_i \geq 0, \ \|\widetilde{\mathbf{h}}_i\|_2 \leq 1} \|\widetilde{\mathbf{h}}_i - \widetilde{\mathbf{b}}_i^k\|_2^2 \tag{9}$$

$$\mathbf{W}^{k+1} = \max\left\{\mathbf{W}^k - \frac{1}{c^k}\nabla_{\mathbf{W}} F_{\rho}(\mathbf{W}^k, \mathbf{H}^{k+1}), 0\right\}, \tag{10}$$

4: Set $k = k + 1$,
5: **until** a predefined stopping criteria is satisfied.

(a) If $\mathbf{b} \leq 0$, then $\boldsymbol{x}^{\star} = \mathbf{0}$, and if $\mathbf{b} > 0$, then $\boldsymbol{x}^{\star} = \frac{\mathbf{b}}{\max\{\|\mathbf{b}\|_2, 1\}}$.

(b) If $\mathbf{b} \nleq 0$ and $\mathbf{b} \ngtr 0$, partition $\mathbf{b}$ as $\mathbf{b} = [\mathbf{b}_+^T, \mathbf{b}_-^T]^T$ without loss of generality, where $\mathbf{b}_+ > 0$ and $\mathbf{b}_- \leq 0$. Then, $\boldsymbol{x}_-^{\star} = \mathbf{0}$ and $\boldsymbol{x}_+^{\star} = \frac{\mathbf{b}_+}{\max\{\|\mathbf{b}_+\|_2, 1\}}$.

The proof is easy and is not presented. Analogously, as shown in (10), the update of $\mathbf{W}$ is also simple, where $c^k > 0$ is a step size. Therefore, the PALM method in Algorithm 2 is efficiently implementable. The convergence conditions of the PALM algorithm (Algorithm 2) has been analyzed in [22]. According to [22, Theorem 1], given proper values of c^k and t^k [22, (3.3), (3.4)] and given the fact that $F_{\rho}(\mathbf{W}, \mathbf{H})$ is a coercive function, the PALM algorithm can yield a bounded stationary point of problem (8) as $k \rightarrow \infty$.

4. NUMERICAL RESULTS AND DISCUSSIONS

We examine the clustering performance of the proposed SNCP method against seven existing clustering methods, namely, K-means (KM), K-means++ [6], NMF followed by K-means (NMF+KM) [5], DTPP [12], ONP-MF [16], ONMF-S [14] and HALS [15]. The purity [28], adjusted Rand index (ARI) [29] and clustering accuracy (ACC) [30] are adopted for performance evaluation.

4.1. Performance with Synthetic Data

The synthetic data is based on the linear model $\mathbf{X} = \mathbf{W}\mathbf{H} + \mathbf{E}$ where $\mathbf{E} \in \mathbb{R}^{M \times N}$ denotes the measurement noise. The signal to noise ratio (SNR) is defined as $10\log_{10}(\|\mathbf{W}\mathbf{H}\|_F^2 / \|\mathbf{E}\|_F^2)$. We follow the same procedure as in [5] to generate $\mathbf{W}, \mathbf{E}$ and the cluster assignment matrix $\mathbf{H}$, with $M = 2000$, $N = 1000$ and $K = 10$ (10 clusters). The number of data samples in the 10 clusters are $117, 62, 36, 124, 15, 24, 119, 43, 122$ and 338, respectively. Like [5], 5% of the data samples are replaced by the same number of randomly generated outliers.

Algorithm convergence: Let us first examine the converge behavior of the proposed SNCP method. We define the following two terms for accessing the convergence and satisfaction of the orthogonality constraint

$$\text{Normalized Residual } = \frac{\|\mathbf{W}^{(r)} - \mathbf{W}^{(r-1)}\|_F}{\|\mathbf{W}^{(r-1)}\|_F} + \frac{\|\mathbf{H}^{(r)} - \mathbf{H}^{(r-1)}\|_F}{\|\mathbf{H}^{(r-1)}\|_F},$$

$$\text{Normalized Orthogonality } = \frac{\|\mathbf{Q}^{(r)}\mathbf{H}^{(r)}(\mathbf{Q}^{(r)}\mathbf{H}^{(r)})^T - \mathbf{I}_K\|_F}{K^2},$$

where $\mathbf{Q}^{(r)}$ is a diagonal matrix such that rows of $\mathbf{Q}^{(r)}\mathbf{H}^{(r)}$ have unit 2-norm. The stopping condition of Algorithm 2 is the normalized residual of $(\mathbf{W}^k, \mathbf{H}^k)$ less than ϵ, where $\epsilon \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. The initial penalty parameter ρ is set to 10^{-8} and the parameter γ for increasing ρ is set to 1.1. One can see from Fig. 1 that the proposed SNCP method indeed converges and

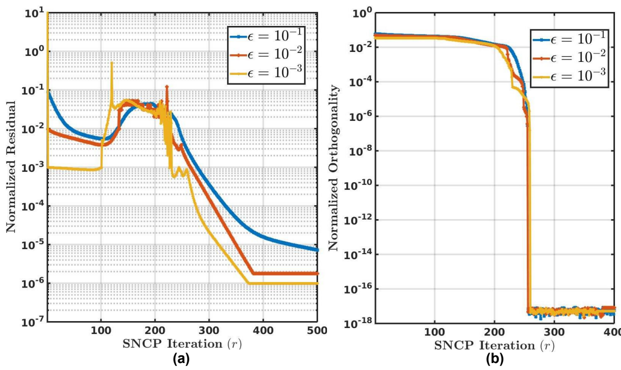

Fig. 1: Convergence curves of normalized residual and orthogonality achieved by SNCP on the synthetic data with SNR = -1 dB.

Table 1: Clustering performance (%) on the synthetic data for different values of SNR.

SNR (dB)		-5	-3	-1	1	3	5
Purity	KM	68.7	75.9	79.2	78.1	77.7	76.1
	KM++	75.9	79.7	79.7	76.9	78.5	76.7
	NMF+KM	91.1	90.2	91.2	92.5	93.3	93.3
	DTPP	90.3	91.7	92.0	92.6	92.5	93.3
	ONP-MF	72.0	75.2	81.6	81.1	81.8	83.7
	ONMF-S	86.9	87.9	81.6	91.2	91.9	92.0
	HALS	90.8	**92.5**	92.5	93.0	93.0	93.4
	SNCP	**92.0**	92.4	**92.6**	**93.0**	**93.7**	**93.8**
ARI	KM	57.0	65.8	73.8	73.6	73.4	68.7
	KM++	57.7	60.0	61.8	64.8	68.2	69.8
	NMF+KM	90.6	90.4	90.9	91.5	91.9	92.0
	DTPP	78.8	84.3	84.5	87.1	85.0	89.5
	ONP-MF	42.6	56.9	65.1	65.8	68.2	70.7
	ONMF-S	71.2	73.4	73.6	77.6	78.4	78.6
	HALS	67.9	80.3	85.2	87.2	87.0	86.2
	SNCP	**91.1**	**91.3**	**91.5**	**91.6**	**91.9**	**92.0**
ACC	KM	63.4	69.8	74.5	74.3	75.6	75.9
	KM++	64.8	68.1	68.7	69.8	71.7	73.8
	NMF+KM	89.2	88.5	89.4	90.8	91.4	91.4
	DTPP	82.6	86.9	87.1	89.4	88.1	91.6
	ONP-MF	57.4	64.5	75.4	75.0	75.1	78.4
	ONMF-S	77.8	79.1	79.9	82.5	83.3	83.7
	HALS	76.2	84.9	88.5	89.6	89.4	89.4
	SNCP	**91.5**	**91.7**	**92.1**	**92.7**	**93.3**	**93.6**

satisfies well the orthogonality constraint. Moreover, if the stopping criterion for the inner PALM is more stringent, then the SNCP takes less iterations to converge.

Clustering quality: Table 1 lists the clustering performance of the methods under test on the synthetic data with different SNR values. All results are obtained by averaging over 20 simulation trials. In each trial, all methods use the same randomly generated initial point. For the proposed SNCP method, the stopping condition is when both the normalized residual and orthogonality are less than 2×10^{-6}, and $\epsilon = 3 \times 10^{-3}$ for the PALM. First of all, one can observe that K-means++ does not perform better than the K-means due to the presence of outliers. The NMF based methods (i.e., NMF+KM, DTPP, ONP-MF, ONMF-S, HALS, and proposed SNCP) significantly outperform the K-means and K-means++. Nevertheless, one can see from Table 1 that the proposed SNCP method consistently yield the best clustering performance, especially for the clustering accuracy. Among all the other methods, the NMF+KM performs most closely with the proposed SNCP, which echos the result that dimension reduction can greatly improve the clustering performance.

Clustering stability: We also evaluate the stability of the clustering methods against different initial points. In particular, we adopt

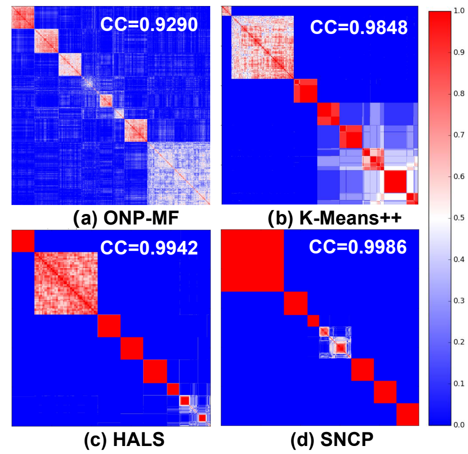

Fig. 2: The consensus map of clustering results for SNR = -3. Here CC stands for the cophenetic correlation coefficient.

Table 2: Clustering performance (%) on 6 real datasets.

Data		1	2	3	4	5	6
Purity	DTPP	90.2	96.2	97.3	96.7	92.5	91.4
	ONMF-S	93.1	79.7	96.8	97.9	91.4	90.2
	ONP-MF	98.1	**97.2**	91.9	98.5	92.3	89.1
	HALS	97.2	94.2	97.1	98.2	**92.8**	91.9
	SNCP	**98.3**	96.3	**97.5**	**98.8**	92.4	**94.3**
ARI	DTPP	88.0	66.1	92.1	90.7	31.2	41.2
	ONMF-S	87.5	60.0	91.1	95.9	32.5	41.5
	ONP-MF	84.2	**72.1**	91.0	96.8	29.9	**57.1**
	HALS	93.2	63.0	91.7	96.1	32.5	42.7
	SNCP	**93.6**	69.9	**93.3**	**97.4**	**32.6**	48.7
ACC	DTPP	80.6	71.0	89.1	90.5	52.6	50.8
	ONMF-S	88.6	68.1	89.5	98.0	56.1	55.1
	ONP-MF	91.3	79.3	91.4	98.5	53.2	**65.8**
	HALS	90.7	66.9	89.2	98.2	53.1	54.5
	SNCP	**94.1**	**80.9**	**93.1**	**98.8**	**58.0**	58.8

the consensus map and cophenetic correlation (CC) coefficient [31] to measure the stability. Roughly speaking, in the consensus map, the (i, j)th entry will be close to 1 if sample i and sample j are consistently assigned to the same cluster even under different initial conditions. The CC coefficient (between 0 and 1) is used to quantize the overall quality of the consensus map. Fig. 2 shows the results for SNR = -3 dB, and one can see that the proposed SNCP method can give consistent clustering results except for small-sized clusters.

4.2. Performance with Real Dataset

We consider the real dataset TDT2 corpus [30] which consists of 10212 on-topic documents in total with 56 semantic categories. We extract 6 subsets each of which contains 10 randomly picked categories ($K = 10$). Table 2 displays the best clustering result of each method obtained from 10 different initial points. As seen, in most of the cases, the SNCP method provides improved clustering performance over the existing ONMF based clustering methods. Interestingly, it is observed that the ONP-MF in [16] performs best for the 6th dataset in terms of ARI and ACC.

In summary, in this paper we have proposed the SNCP method for data clustering. The numerical results presented above have shown that the proposed method mostly outperforms the existing seven methods under test for either synthetic data and the real dataset in [30]. Due to limited space, the current paper cannot present results about computational complexity and computation time. Experimental experience indicates that the proposed SNCP method is competitive and more efficient when compared to the other ONMF based methods.

5. REFERENCES

[1] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*, Chapman & Hall/CRC Press, Boca Raton, FI, USA, 2013.

[2] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Info. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[3] H. Ding, Y. Liu, L. Huang, and J. Li, "K-means clustering with distributed dimensions," in *Proc. ICML*, New York, USA, Jun. 19-24, 2016, pp. 1339–1348.

[4] C. Bauckhage, "K-means clustering is matrix factorization," *arXiv preprint arXiv:1512.07548*, 2015.

[5] B. Yang, X. Fu, and N. D. Sidiropoulos, "Learning from hidden traits: joint factor analysis and latent clustering," *IEEE Trans. Signal Process.*, vol. 65, pp. 256–269, Jan. 2017.

[6] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. SODA*, Philadelphia, PA, USA, Jan. 07-08, 2007, pp. 1027–1035.

[7] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, Denver, CO, USA, Dec. 2000, pp. 556–562.

[9] A. C. Turkmen, "A review of non-negative matrix factorization methods for clustering," *CoRR*, vol. abs/1507.03194, 2015.

[10] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. ICML*, New York, NY, USA, Jun. 19-24, 2016, pp. 478–487.

[11] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards K-means-friendly spaces: simultaneous deep learning and clustering," in *Proc. ICML*, Sydney, Australia, Aug. 06-11, 2017, pp. 3861–3870.

[12] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal non-negative matrix t-factorizations for clustering," in *Proc. ACM KDD*, Philadelphia, PA, USA, Aug. 20-23, 2006, pp. 20–23.

[13] J. Yoo and S. Choi, "Non-negative matrix factorization with orthogonality constraints," *J. Comp. Sci. Eng.*, vol. 4, no. 2, pp. 97–109, May 2010.

[14] S. Choi, "Algorithms for orthogonal non-negative matrix factorization," in *Proc. IEEE IJCNN*, Hong Kong, China, Jun. 01-08, 2008, pp. 1828–1832.

[15] K. Kimura, Y. Tanaka, and M. Kudo, "A fast hierarchical alternating least squares algorithm for orthogonal non-negative matrix factorization," in *Proc. ACML*, Nha Trang City, Vietnam, Nov. 26-28, 2015, pp. 129–141.

[16] F. Pompili, N. Gillis, P. A. Absil, and F. Glineur, "Two algorithms for orthogonal non-negative matrix factorization with application to clustering," *Neurocomputing*, vol. 141, no. 2, pp. 15–25, Oct. 2014.

[17] M. Asteris, D. Papailiopoulos, and A. G. Dimakis, "Orthogonal NMF through subspace exploration," in *Proc. NIPS*, Canada, Dec. 07-12, 2015, pp. 343–351.

[18] S. Wang, P. Wu, M. Zhou, T.-H. Chang, and S. Wu, "Cell subclass identification in single-cell RNA-sequencing data using orthogonal non-negative matrix factorization," in *Proc. IEEE ICASSP*, Calgary, Canada, Apr. 15-20, 2018, pp. 876–880.

[19] H. Mansour, S. Rane, P. T. Boufounos, and A. Vetro, "Video querying via compact descriptors of visually salient objects," in *Proc. IEEE ICIP*, Paris, France, Oct. 27-30, 2014, pp. 2789–2793.

[20] A. Mirzal, "Nonparametric orthogonal NMF and its application in cancer clustering," in *Proc. DaEng*, Kuala Lumpur, Malaysia, Dec. 15, 2013, pp. 177–184.

[21] S.-P. Han and O. L. Mangasarian, "Exact penalty functions in nonlinear programming," *Math Program.*, vol. 17, pp. 251–269, Dec. 1979.

[22] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for non-convex and non-smooth problems," *Math Program.*, vol. 146, no. 1, pp. 459–494, Aug. 2014.

[23] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems, Vol. I*, Springer-Verlag, New York Berlin Heidelberg, 2003.

[24] D. P. Bertsekas, *Nonlinear Programming, 2nd Edition*, Athena Scientific, Belmont, Massachusetts, 1999.

[25] G. D. Pillo and L. Grippo, "On the exactness of a class of non-differentiable penalty functions," *JOTA*, vol. 57, no. 3, pp. 399–410, Jun. 1988.

[26] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1997.

[27] N. Zhang, Y.-F. Liu, H. Farmanbar, T.-H. Chang, M. Hong, and Z.-Q. Luo, "Network slicing for service-oriented networks under resource constraints," *IEEE JSAC*, vol. 35, no. 11, pp. 2512–2521, Nov. 2017.

[28] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, UK, 2008.

[29] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, May 2001.

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

[31] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," in *Proc. Natl. Acad. Sci. USA*, Mar. 23, 2004, pp. 4164–4169.