Spherical Principal Component Analysis[*]

Kai Liu[†‡] Qiuwei Li[†§] Hua Wang[‡¶] Gongguo Tang[§]

Abstract

Principal Component Analysis (PCA) is one of the most broadly used methods to analyze high-dimensional data. However, most existing studies on PCA aim to minimize the reconstruction error measured by the Euclidean distance, although in some fields, such as text analysis in information retrieval, analysis using the angle distance is known to be more effective. In this paper, we propose a novel PCA formulation by adding a constraint on the factors to unify the Euclidean distance and the angle distance. Because the objective and constraints are nonconvex, the optimization problem is difficult to solve in general. To tackle the optimization problem, we propose an alternating linearized minimization method with guaranteed convergence and provable convergence rate. Experiments on synthetic data and real-world data sets have validated the effectiveness of our new method and demonstrated its advantages over state-of-art competing methods.

1 Introduction

In many real-world applications such as text categorization and face recognition, the dimensionality of the input data is usually very high. Because dealing with high-dimensional data is computationally expensive and noise or outliers in the data can increase dramatically as the dimension increases, dimension reduction plays a critical role when one analyzes high-dimensional data [4,17,19]. Among many dimension reduction methods, Principal Component Analysis (PCA) is one of the most broadly used one due to its simplicity and effectiveness.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of correlated variables into a set of linearly uncorrelated principal directions. Usually the number of principal directions is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal direction has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding direction has the highest variance under the constraint that it is orthogonal to the preceding directions. The resulting vectors are an uncorrelated orthogonal basis set.

When data points lie in a low-dimensional manifold and the manifold is linear or nearly-linear, the low-dimensional structure of data can be effectively captured by a linear subspace spanned by the principal PCA directions. More specifically, let $X = (x_1, \ldots, x_n)$ be n data points in m-dimensional space while $U = (u_1, \ldots, u_r)$ contains the principal directions and $V = (v_1, \ldots, v_k)$ contains the principal components (projected data along the principal directions). There exist two broadly used formulations for PCA:

- *Covariance-based approach* computes the covariance matrix $C = \sum_i (x_i - \bar{x})(x_i - \bar{x}) = XX^T$. By assuming that the data are centered, *i.e.*, $\bar{x} = 0$, we can drop the factor $\frac{1}{n-1}$ that does not affect U. The principal directions are obtained by:

$$\max_{U^T U = I} \operatorname{trace}(U^T X X^T U), \tag{1.1}$$

- *Low-rank approximation-based approach* solves the following optimization problem:

$$\min_{U^T U = I} J = \|X - UV\|_F^2 = \sum_{i,j} [X_{ij} - (UV)_{ij}]^2, \tag{1.2}$$

  where we approximate X by UV.

Taking the derivative w.r.t.V and setting it to zero, we have $V = X^T U$, by which Eq. (1.2) reduces to Eq. (1.1). Therefore, the solutions to these two approaches are identical. In our paper, we will focus on the second formulation.

2 Motivation

In Eq. (1.2), the objective function measures the difference between the original data X and its approximation of UV in the projected space, which is measured by the squared Euclidean distances and uses each feature with equal importance. However, in the real-world applications, there exist data sets which are preprocessed to be normalized and different features may have varied significances. Thus distance-based measurement method may

[*] This work was partially supported by NSF-IIS 1652943.

[†] Equal contribution.

[‡] Department of Computer Science, Colorado School of Mines, Golden, CO 80401, U.S.A.

[§] Department of Electrical Engineering, Colorado School of Mines, Golden, CO 80401, U.S.A.

[¶] Corresponding author: huawangcs@gmail.com.

yield poor results. On the other side, similarity-based measurement methods, such as the cosine similarity derived from the angle distance, have been proved to be more effective in some applications, such as information retrieval [18], signal processing [8], metric learning [15], *etc.*. Thus, deriving new methods that can directly measure the angle distance from PCA is of vital importance to solve real-world problems. However, this problem has not been well studied in literature yet.

Motivated by the above observations, in this paper we propose a spherical PCA model that can unify the Euclidean distance and the angle distance. By noticing that larger angle in the sphere in Fig. 1 also has larger Euclidean distance, we can add the normalization constraint to the component matrix, where the norm of each column in V is restricted to be 1 to guarantee the spherical distribution of components in the projected space:

(2.3)
$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{r\times n}} J = \|X - UV\|_F^2 = \sum_{i,j}[X_{ij} - (UV)_{ij}]^2, \quad s.t. \quad U \in \mathbb{U}, V \in \mathbb{V}.$$

where we define:

(2.4)
$$\mathbb{U} := \{U : U^TU = I\}, \mathbb{V} := \{V : \|V(:,j)\| = 1\, \forall j \in [0,n]\}.$$

Suppose the component is spherically distributed, then the Euclidean distance between v_i and v_j is:

(2.5)
$$||v_i - v_j||_2^2 = ||v_i||^2 + ||v_j||^2 - 2\langle v_i, v_j\rangle$$
$$= ||v_i||^2 + ||v_j||^2 - 2\frac{\langle v_i, v_j\rangle}{||v_i||||v_j||} = 2 - 2\ cos(\theta), \theta \in [0,\pi].$$

which is equivalent to angle distance such that a bigger angle θ will result in a larger Euclidean distance, and vice versa.

Remark 1. *In traditional PCA, without the normalization constraint on each column of V, the optimized solution to Eq. (1.2) can barely satisfy the spherical distribution. Since r is usually less than m, PCA will lose some component more or less, thus $x_i \neq Uv_i$ and usually $\|x_i\| \neq \|Uv_i\|$ (they may be equal, but it barely happens) . We have $||X_i||^2 = 1$ for normalized data and if $||v_i||^2 = 1$ then $||Uv_i||^2 = tr(v_i^TU^TUv_i) = tr(v_i^Tv_i) = ||v_i||^2 = 1$, which leads a contradiction. Thus we have to additionally enforce the constraint on V to guarantee our motivation.*

3 Formulation and Algorithm

3.1 Objective Function with Proximal Term

We first denote:

(3.6)
$$h(U,V) = \|X - UV\|_F^2, \quad s.t. \quad U \in \mathbb{U}, V \in \mathbb{V}.$$

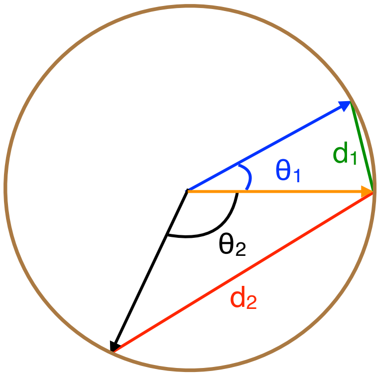

Figure 1: Larger angles ($\theta_2 > \theta_1$) in the sphere will have larger Euclidean distance, and vice versa, which unifies the cosine similarity and Euclidean distance simultaneously.

By noting the nonconvexity of Eq. (2.3), where no closed solution exists, we propose an alternating minimization method to get the optimized solution as:

(3.7)
$$U_{k+1} = \|X - UV^k\|_F^2, \quad U \in \mathbb{U},$$
$$v_{k+1} = \|x - U_{k+1}v\|_2^F, \quad V \in \mathbb{V}.$$

where v denotes the column of V and V in Eq. (2.3) can be decoupled into column-wise for optimization.

Note that because of the constraints $U \in \mathbb{U}, V \in \mathbb{V}$, the problem (3.6) is known as the nonconvex matrix factorization problems, which was previously studied in [12,24]. In this work we will focus on develop an efficient algorithm with provable convergence to solve (3.6). Note that the proximal algorithm recently has been successfully applied to a wide variety of situations: convex optimization, nonmonotone operators [6,10] with various applications to nonconvex programming. It was first introduced by Rockafellar [16] as an approximation regularization method in convex optimization and in the study of variational inequalities associated to maximal monotone operators.

Considering the fact that the objective function in Eq. (2.3) is nonconvex w.r.t.U and V, and the constraint on U and V are also nonconvex, we add the proximal term, which leads to the following alternating linearized minimization solutions:

(3.8)
$$U_{k+1} = \arg\min_{U^TU=I}\langle U - U_k, \nabla h(U_k)\rangle + \frac{\mu}{2}||U - U_k||_F^2,$$
$$v_{k+1} = \arg\min_{\|v\|=1}\langle v - v_k, \nabla h(v_k)\rangle + \frac{\lambda}{2}||v - v_k||^2.$$

Remark 2. *We add the proximal term to ensure that the updated solution will not go apart too far away* *from the previous step to avoid drastic changes. One can see that when the proximal term regularization parameters μ, λ are sufficiently large, they will dominate the objective function. Moreover, we can exploit the linearized minimization that minimizes the objective with Taylor expansion by making use of first order (linear) information.*

3.2 Proposed Algorithm Given the alternating minimization objective in Eq. (3.8), now we turn to derive the detailed updating algorithm with closed-form solutions in every step.

First, we derive the solution for U. Before the derivation, we prove the following useful lemma that is similar to [22, Theorem 1] and [21, Theorem 1]:

LEMMA 1. $\max_{X^TX=I} tr(X^TB)$ *is given by* $X = UV^T$, *where* $[U, \Sigma, V] = svd(B)$.

Proof. On one hand, we have:

$$\text{trace}(X^TB) = \text{trace}(X^TU\Sigma V^T) = \text{trace}(P\Sigma). \tag{3.9}$$

where $P = V^TX^TU$ is an orthonormal matrix since $PP^T = V^TX^TUU^TXV = I$. Thus every element including the diagonal of P is less than or equal to 1. Thus we have:

$$\text{trace}(P\Sigma) \leq \text{trace}(\Sigma). \tag{3.10}$$

On the other hand, when $X = UV^T$, we have $\text{trace}(X^TB) = \text{trace}(VU^TU\Sigma V^T) = \Sigma$. Thus $X = UV^T$ is the optimal solution that maximizes the objective. □

Accordingly, we have:

$$\begin{aligned} U_{k+1} &= \arg\min_{U^TU=I} \langle U - U_k, \nabla h(U_k)\rangle + \frac{\mu}{2}||U - U_k||_F^2 \\ &= \arg\max_{U^TU=I} \text{trace}(U^TM) = YZ^T. \end{aligned} \tag{3.11}$$

where $M = 2(X - U_kV_k)V_k^T + \mu U_k$ and Y, Z is obtained from $[Y, \Sigma, Z] = svd(M)$.

Second, we optimize v_{k+1} given that $U_{k+1}^TU_{k+1} = I$:

$$\begin{aligned} v_{k+1} &= \arg\min_{\|v\|=1} \langle v - v_k, \nabla h(v_k)\rangle + \frac{\lambda}{2}||v - v_k||_F^2 \\ &= \arg\max_{\|v\|=1} \langle v, q\rangle \\ &= \frac{q}{\|q\|_2}, \end{aligned} \tag{3.12}$$

where $q = 2U_{k+1}^Tx + (\lambda - 2)v_k$.

Algorithm 1 Alternating Linearized Minimization for Problem Eq. (3.6)

```
Input: data $X \in \mathbb{R}^{m\times n}$, rank of factors $r$, regularization parameters $\lambda, \mu$, number of iterations $K$
Initialization: $U_0 \in \mathbb{R}^{m\times r}, V_0 \in \mathbb{R}^{r\times n}$
while $k \leq K$ do
    optimize $U_{k+1}$ as Eq. (3.11)
    optimize each $v_{k+1}$ as Eq. (3.12)
end while
Output: $U_K$ and $V_K$
```

4 Convergence Analysis

In the following case, we let $\mathbb{U}$ and $\mathbb{V}$ be as defined in Eq. (2.4), and show the convergence of our proposed algorithm in the last section.

To begin with, we first show that $h(U, V)$ has Lipschitz continuous gradient at $U \in \mathbb{U}, V \in \mathbb{V}$, which will be very useful for the following convergence analysis.

PROPOSITION 1. *$h(U, V)$ has Lipschitz continuous gradient at $U \in \mathbb{U}, V \in \mathbb{V}$, where $\mathbb{U}$ and $\mathbb{V}$ are defined in Eq. (2.4). That is, there exists a constant L_c such that*

$$\|\nabla h(U, V) - \nabla h(U', V')\|_F \leq L_c\|(U, V) - (U', V')\|_F \tag{4.13}$$

for all $U, U' \in \mathbb{U}$ and $V, V' \in \mathbb{V}$. Here $L_c > 0$ is referred to as the Lipschitz constant.

Proof. Proof of Proposition 1 is equivalent to show $\|\nabla^2 h(U, V)\|_2 \leq L_c$ for all $U \in \mathbb{U}, V \in \mathbb{V}$. Standard computations give the Hessian quadrature form $[\nabla^2 h(U, V)](\Delta, \Delta)$ for any $\Delta = \begin{bmatrix}\Delta_U \\ \Delta_V^{\mathrm{T}}\end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ (where $\Delta_U \in \mathbb{R}^{m\times r}$ and $\Delta_V \in \mathbb{R}^{r\times n}$) as

$$\begin{aligned} &[\nabla^2 h(U, V)](\Delta, \Delta) \\ &= \|\Delta_U V + U\Delta_V\|_F^2 + 2\langle UV - X, \Delta_U\Delta_V\rangle, \end{aligned} \tag{4.14}$$

which gives:

$$\begin{aligned} \|\nabla^2 h(U, V)\|_2 &= \max_{\|\Delta\|_F=1} \left|[\nabla^2 h(U, V)](\Delta, \Delta)\right| \\ &\leq \max_{\|\Delta\|_F=1} \|\Delta_U V + U\Delta_V\|_F^2 + 2\left|\langle UV - X, \Delta_U\Delta_V\rangle\right| \\ &\leq 2(\|U\|_F^2 + \|V\|_F^2 + \|U\|_F\|V\|_F + \|X\|_F) := L_c, \end{aligned} \tag{4.15}$$

where the inequality follows from $|\langle A, B\rangle| \leq \|A\|_F\|B\|_F$ and $\|CD\|_F \leq \|C\|_F\|D\|_F$. Due to the constraints on U and V, we have $\|U\|_F^2 = tr(U^TU) = tr(I) = r, \|V\|_F^2 = \sum_{j=1}^n \|v\|^2 = n$. □

To analyse the convergence, we rewrite Eq. (3.6) as

$$\min_{U,V} f(U, V) = h(U, V) + \delta_{\mathbb{U}}(U) + \delta_{\mathbb{V}}(V), \tag{4.16}$$

where $\delta_{\mathbb{U}}(U) = \begin{cases} 0, & U \in \mathbb{U} \\ \infty, & U \notin \mathbb{U} \end{cases}$ is the indicator function of the set $\mathbb{U}$ and therefore nonsmooth, so is $\delta_{\mathbb{V}}(V)$.

The following result establishes that the subsequence convergence property of the proposed algorithm, *i.e.*, the sequence generated by Algorithm 1 is bounded and any of its limit point is a critical point of Eq. (4.16).

Theorem 1. (Subsequence convergence) *Let $\{W_k\}_{k\geq 0} = \{(U_k, V_k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1 with constant step size $\lambda, \mu > L_c$. Then the sequence $\{W_k\}_{k\geq 0}$ is bounded and obeys the following properties:*

(P1) sufficient decrease:

$$f(W_{k-1}) - f(W_k) \geq \frac{\min(\lambda,\mu) - L_c}{2}\|W_k - W_{k-1}\|_F^2, \tag{4.17}$$

which implies that

$$\lim_{k\to\infty} \|W^{k-1} - W^k\|_F = 0. \tag{4.18}$$

(P2) the sequence $\{f(W_k)\}_{k\geq 0}$ is convergent.

(P3) for any convergent subsequence $\{W_{k'}\}$, its limit point $W^\star$ is a critical point of f and

$$\lim_{k'\to\infty} f(W_{k'}) = \lim_{k\to\infty} f(W_k) = f(W^\star). \tag{4.19}$$

Before proving Theorem 1, we give out some necessary definition.

Definition 1. *[3] Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a proper and lower semi-continuous function, whose domain is defined as*

$$\operatorname{dom} f := \{u \in \mathbb{R}^n : f(u) < \infty\}.$$

The (Fréchet) subdifferential ∂f of f at u is defined by

$$\partial f(u) = \left\{ z : \liminf_{v\to u} \frac{f(v) - f(u) - \langle z, v-u\rangle}{\|u-v\|} \geq 0 \right\}$$

for any $u \in \operatorname{dom} h$ and $\partial f(u) = \emptyset$ if $u \notin \operatorname{dom} f$.

We say u is a limiting critical point, or simply a critical point of f if

$$0 \in \partial f.$$

We now turn to prove Theorem 1.

Proof. [Proof of Theorem 1] (P1): First note that for all k, according to our alternating minimization method, we always have $\delta_{\mathbb{U}}(U_k) = \delta_{\mathbb{V}}(V_k) = 0$ and thus $f(W_k) = h(W_k)$.

Since $h(U, V)$ has Lipschitz continuous gradient at $U \in \mathbb{U}, V \in \mathbb{V}$ with Lipschitz gradient L_c and $\lambda > L_c$ as proved in Proposition 1, we define $h_{L_c}(U, U', V)$ as proximal regularization of $h(U, V)$ linearized at U', V:

$$h(U', V) + \langle \nabla_U h(U', V), U - U'\rangle + \frac{L_c}{2}\|U - U'\|_F^2,$$

By the definition of Lipschitz continuous gradient and Taylor expansion, we have

$$h(U, V) \leq h_{L_c}(U, U', V). \tag{4.20}$$

Also by the definition of proximal map, we get:

$$\begin{aligned} U_k = \arg\min_U \delta_{\mathbb{U}}(U) + \frac{\mu}{2}\|U - U_{k-1}\|_F^2 \\ + \langle \nabla_U h(U_{k-1}, V_{k-1}), U - U_{k-1}\rangle. \end{aligned} \tag{4.21}$$

Hence we take $U_k = U$, which implies that

$$\begin{aligned} &\delta_{\mathbb{U}}(U_k) + \frac{\mu}{2}\|U_k - U_{k-1}\|_F^2 \\ &+ \langle \nabla_U h(U_{k-1}, V_{k-1}), U_k - U_{k-1}\rangle \leq \delta_{\mathbb{U}}(U_{k-1}). \end{aligned} \tag{4.22}$$

Combining Eq. (4.20) to Eq. (4.22), we have:

$$\begin{aligned} &h(U_k, V_{k-1}) + \delta_{\mathbb{U}}(U_k) \\ &\leq h(U_{k-1}, V_{k-1}) + \langle \nabla_U h(U_{k-1}, V_{k-1}), U_k - U_{k-1}\rangle \\ &+ \frac{L_c}{2}\|U_k - U_{k-1}\|_F^2 + \delta_{\mathbb{U}}(U_k) \\ &\leq h(U_{k-1}, V_{k-1}) + \frac{L_c}{2}\|U_k - U_{k-1}\|_F^2 \\ &+ \delta_{\mathbb{U}}(U_{k-1}) - \frac{\mu}{2}\|U_k - U_{k-1}\|_F^2 \\ &= h(U_{k-1}, V_{k-1}) + \delta_{\mathbb{U}}(U_{k-1}) - \frac{\mu - L_c}{2}\|U_k - U_{k-1}\|_F^2. \end{aligned} \tag{4.23}$$

Similarly, we have

$$\begin{aligned} &h(U_k, V_k) - h(U_k, V_{k-1}) + \delta_{\mathbb{V}}(V_k) - \delta_{\mathbb{V}}(V_{k-1}) \\ &\leq -\frac{\lambda - L_c}{2}\|V_k - V_{k-1}\|_F^2, \end{aligned} \tag{4.24}$$

which together with the above equation gives Eq. (4.17). Now repeating Eq. (4.17) for all k will give

$$(\min(\lambda,\mu) - L_c)\sum_{k=1}^{\infty}\|W_k - W_{k-1}\|_F^2 \leq f(W_0), \tag{4.25}$$

which gives Eq. (4.18).

REMARK 3. *In our proposed algorithm, since in every update, our solution is closed while satisfying the constraints, thus in fact $\delta_{\mathbb{U}}$ and $\delta_{\mathbb{V}}$ are 0, and ∞ is never achieved.*

(P2) It follows from Eq. (16) that $\{f(W_k)\}_{k\geq 0}$ is a decreasing sequence. Due to the fact that f is lower bounded as $f(W_k) \geq 0$ for all k, we conclude that $\{f(W_k)\}_{k\geq 0}$ is convergent.

(P3) Since $U_{k'} \in \mathbb{U}, V_{k'} \in \mathbb{V}$ for all k' and both of the sets $\mathbb{U}$ and $\mathbb{V}$ are closed, we have $U^\star \in \mathbb{U}, V^\star \in \mathbb{V}$. Since h is continuous, we have

$$\begin{aligned}\lim_{k'\to\infty} f(W_{k'}) &= \lim_{k'\to\infty} h(U_{k'}, V_{k'}) + \delta_{\mathbb{U}}(U_{k'}) + \delta_{\mathbb{V}}(V_{k'}) \\ &= f(W^\star),\end{aligned}$$

which together with the fact that $\{f(W_k)\}_{k\geq 0}$ is convergent gives Eq. (4.18).

To show $W^\star$ is a critical point, we first consider Eq. (4.21) and the optimality condition yields:

$$\nabla_U h(U_{k-1}, V_{k-1}) + \mu(U_k - U_{k-1}) + \partial\delta_{\mathbb{U}}(U_k) = 0. \tag{4.26}$$

Similarly, we have

$$\nabla_V h(U_k, V_{k-1}) + \lambda(V_k - V_{k-1}) + \partial\delta_{\mathbb{V}}(V_k) = 0. \tag{4.27}$$

Now, define

$$\underbrace{\nabla_U h(U_k, V_k) + \partial\delta_{\mathbb{U}}(U_k)}_{A_k},$$
$$\underbrace{\nabla_V h(U_k, V_k) + \partial\delta_{\mathbb{V}}(V_k)}_{B_k}.$$

Thus, we have

$$A_k \in \partial_U f(U_k, V_k), B_k \in \partial_V f(U_k, V_k). \tag{4.28}$$

It follows from the above that

$$\begin{aligned}&\lim_{k\to\infty} \|A_k\|_F \\ &\leq \lim_{k\to\infty} \|\nabla_U h(U_k, V_k) - \nabla_U h(U_{k-1}, V_{k-1})\|_F \\ &\quad + \mu\|U_k - U_{k-1}\|_F \\ &\leq \lim_{k\to\infty} (L_c + \mu)\|W_k - W_{k-1}\| = 0.\end{aligned} \tag{4.29}$$

Similarly, we have:

$$\lim_{k\to\infty} \|B_k\|_F \leq \lim_{k\to\infty} (L_c + \lambda)\|W_k - W_{k-1}\| = 0. \tag{4.30}$$

Then we have:

$$\operatorname{dist}(0, \partial f(W_k)) \leq (2L_c + \mu + \lambda)\|W_k - W_{k-1}\|. \tag{4.31}$$

Owing to the closedness properties of $\partial f(W_{k'})$, we finally obtain $0 \in \partial f(W^\star)$. Thus, $W^\star$ is a critical point of f. □

THEOREM 2. (SEQUENCE CONVERGENCE) *The sequence $\{W_k\}_{k\geq 0}$ generated by Algorithm 1 with a constant step size $\lambda, \mu > L_c$ is global-sequence convergence.*

Before proving Theorem 2, we give out another important definition.

DEFINITION 2. (**Kurdyka-Lojasiewicz (KL) property**) *[5] We say a proper semi-continuous function $h(\boldsymbol{u})$ satisfies Kurdyka-Lojasiewicz (KL) property, if $\overline{\boldsymbol{u}}$ is a critical point of $h(\boldsymbol{u})$, then there exist $\delta > 0,\ \theta \in [0,1),\ C_1 > 0,\ s.t.$*

$$|h(\boldsymbol{u}) - h(\overline{\boldsymbol{u}})|^\theta \leq C_1 \operatorname{dist}(0, \partial h(\boldsymbol{u})), \quad \forall\ \boldsymbol{u} \in B(\overline{\boldsymbol{u}}, \delta).$$

We mention that the above KL property(also known as KL inequality) states the regularity of $h(u)$ around its critical point u and the KL inequality trivially holds at non-critical point. There are a very large set of functions satisfying the KL inequality including any semi-algebraic functions [3]. Clearly, the objective function f is semi-algebraic as both h, $\delta_{\mathbb{U}}$ and $\delta_{\mathbb{V}}$ are semi-algebraic.

LEMMA 2. (UNIFORM KL PROPERTY) *There exist $\delta_0 > 0,\ \theta_{KL} \in [0,1),\ C_{KL} > 0$ such that for all W s.t. $\operatorname{dist}((W), \mathbb{C}(W_0)) \leq \delta_0$:*

$$\left|f(W) - \overline{f}\right|^{\theta_{KL}} \leq C_{KL} \operatorname{dist}(0, \partial f(W)) \tag{4.32}$$

with $\overline{f}$ denoting the limiting function value defined in P (2) of Theorem 1.

Proof. First we recognize the union $\bigcup_i B(W_i^\star, \delta_i)$ forms an open cover of $\mathbb{C}(W_0)$ with $W_i^\star$ representing all points in $\mathbb{C}(W_0)$ and δ_i to be chosen so that the the following KL property of f at $W_i^\star \in \mathbb{C}(W_0)$ holds:

$$\left|f(W) - \overline{f}\right|^{\theta_i} \leq C_i \operatorname{dist}(0, \partial f(W)) \ \ \forall\ (W) \in B(W_i^\star, \delta_i)$$

where we have used all $f(W_i^\star) = \overline{f}$ by assertion (P3) of Theorem 1. Then due to the compactness of the set $\mathbb{C}(W_0)$, it has a finite subcover $\bigcup_{i=1}^p B(W_{k_i}^\star, \delta_{k_i})$ for some positive integer p. Now combining all, we have for all $W \in \bigcup_{i=1}^p B(W_{k_i}^\star, \delta_{k_i})$,

$$\left|f(W) - \overline{f}\right|^{\theta_{KL}} \leq C_{KL} \operatorname{dist}(0, \partial f(W)) \tag{4.33}$$

with $\theta_{KL} = \max_{i=1}^p\{\theta_{k_i}\}$ and $C_{KL} = \max_{i=1}^p\{C_{k_i}\}$. Finally, since $\bigcup_{i=1}^p B(W_{k_i}^\star, \delta_{k_i})$ is an open cover of $\mathbb{C}(W_0)$, there exists a sufficiently small number δ_0 so that

$$\{(W) : \operatorname{dist}(W, \mathbb{C}(W_0)) \leq \delta_0\} \subset \bigcup_{i=1}^p B(W_i^\star, \delta_{k_i}).$$

Therefore, eq. (4.33) holds whenever $\operatorname{dist}(W, \mathbb{C}(W_0)) \leq \delta_0$. □

We now turn to prove Theorem 2.

Proof. [Proof of Theorem 2] According to Definition 2, there exists a sufficiently large k_0 satisfying:

$$[f(W_k) - f(W^\star)]^\theta \le C_2 \,\mathrm{dist}(0, \partial f(W_k)), \quad \forall k \ge k_0. \tag{4.34}$$

In the subsequent analysis, we restrict to $k \ge k_0$. Construct a concave function $x^{1-\theta}$ for some $\theta \in [0,1)$ with domain $x > 0$. Obviously, by the concavity, we have

$$x_2^{1-\theta} - x_1^{1-\theta} \ge (1-\theta) x_2^{-\theta}(x_2 - x_1), \forall x_1 > 0, x_2 > 0$$

. Replacing x_1 by $f(W_{k+1}) - f(W^\star)$ and x_2 by $f(W_k) - f(W^\star)$ and using the sufficient decrease property, we have

$$\begin{aligned}
&[f(W_k) - f(W^\star)]^{1-\theta} - [f(W_{k+1}) - f(W^\star)]^{1-\theta} \\
&\ge (1-\theta) \frac{f(W_k) - f(W_{k+1})}{[f(W_k) - f(W^\star)]^\theta} \\
&\ge \frac{\lambda(1-\theta)}{2C_2} \frac{\|W_k - W_{k+1}\|_F^2}{\mathrm{dist}(0, \partial f(W_k))}, \\
&\ge \frac{\lambda(1-\theta)}{2C_2 C_3} \frac{\|W_k - W_{k+1}\|_F^2}{\|W_k - W_{k-1}\|_F} \\
&= \kappa\Big(\frac{\|W_k - W_{k+1}\|_F^2}{\|W_k - W_{k-1}\|_F} + \|W_k - W_{k-1}\|_F\Big) \\
&- \kappa \|W_k - W_{k-1}\|_F \\
&\ge \kappa \left(2\|W_k - W_{k+1}\|_F - \|W_k - W_{k-1}\|_F\right).
\end{aligned}$$

And accordingly, we have:

$$\begin{aligned}
&2\|W_k - W_{k+1}\|_F - \|W_k - W_{k-1}\|_F \\
&\le \beta\left([f(W_k) - f(W^\star)]^{1-\theta} - [f(W_{k+1}) - f(W^\star)]^{1-\theta}\right),
\end{aligned} \tag{4.35}$$

with $C_3 := 2L_c + \mu + \lambda, \kappa := \frac{\lambda(1-\theta)}{2C_2C_3}, \beta := \left(\frac{\lambda(1-\theta)}{2C_2C_3}\right)^{-1}$.

Summing the above inequalities up from some $\widetilde{k} > k_0$ to infinity yields

$$\begin{aligned}
&\sum_{k=\widetilde{k}}^{\infty} \|W_k - W_{k+1}\|_F \\
&\le \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta}
\end{aligned} \tag{4.36}$$

implying

$$\sum_{k=\widetilde{k}}^{\infty} \|W_k - W_{k+1}\|_F < \infty.$$

Following some standard arguments one can see that

$$\limsup_{t\to\infty, t_1, t_2 \ge t} \|W_{t_1} - W_{t_2}\|_F = 0,$$

which implies that the sequence $\{W_k\}$ is Cauchy, and hence convergent. Hence, the limit point set $\mathcal{C}(W_0)$ is singleton $W^\star$. □

THEOREM 3. (CONVERGENCE RATE) *The convergence rate is at least sub-linear.*

Towards that end, we first know from the above argument that $\{W_k\}$ converges to some point $W^\star$, i.e., $\lim_{k\to\infty} W^k = W^\star$. Then using Equation (4.36) and the triangle inequality, we obtain

$$\begin{aligned}
\|W_{\widetilde{k}} - W^\star\|_F &\le \sum_{k=\widetilde{k}}^{\infty} \|W_k - W_{k+1}\|_F \\
&\le \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta},
\end{aligned} \tag{4.37}$$

which indicates the convergence rate of $W_{\widetilde{k}} \to W^\star$ is at least as fast as the rate that $\|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta}$ converges to 0. In particular, the second term $\beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta}$ can be controlled:

$$\begin{aligned}
&\beta[f(W_{\widetilde{k}}) - f(W^\star)]^\theta \le \beta C_2 \,\mathrm{dist}(0, \partial f(W_{\widetilde{k}})) \\
&\le \underbrace{\beta C_2(2B_0 + \lambda + \|\boldsymbol{X}\|_F)}_{:=\alpha} \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F.
\end{aligned} \tag{4.38}$$

Plugging (4.38) back to (4.37), we then have

$$\sum_{k=\widetilde{k}}^{\infty} \|W_k - W_{k+1}\|_F \le \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \alpha \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F^{\frac{1-\theta}{\theta}}.$$

We divide the following analysis into two cases based on the value of the KL exponent θ.

- *Case I*: If $\theta = 0$, we set $Q := \{k \in \mathbb{N} : W_{k+1} \ne W_k\}$ and take k in Q. When k is sufficiently large, then we have:

  $$\|W_{k+1} - W_k\|_F^2 := C_4 > 0. \tag{4.39}$$

  On the other hand,

  $$\begin{aligned}
  f(W_{k+1}) - f(W_k) &\ge \frac{\min(\lambda,\mu) - L_c}{2} \|W_{k+1} - W_k\|_F^2 \\
  &= \frac{\min(\lambda,\mu) - L_c}{2} C_4.
  \end{aligned} \tag{4.40}$$

  Since $f(W_k)$ is known to be converged to 0, Eq. (4.40) implies that Q is finite and sequence W_k converges in a finite number of steps.

- *Case II*: $\theta \in (0, \frac{1}{2}]$. This case means $\frac{1-\theta}{\theta} \ge 1$. We define $P_{\widetilde{k}} = \sum_{i=\widetilde{k}}^{\infty} \|W_{i+1} - W_i\|_F$,

  $$P_{\widetilde{k}} \le P_{\widetilde{k}-1} - P_{\widetilde{k}} + \alpha \left[P_{\widetilde{k}-1} - P_{\widetilde{k}}\right]^{\frac{1-\theta}{\theta}}. \tag{4.41}$$

Since $P_{\widetilde{k}-1} - P_{\widetilde{k}} \to 0$, there exists a positive integer k_1 such that $P_{\widetilde{k}-1} - P_{\widetilde{k}} < 1$, $\forall\, \widetilde{k} \geq k_1$. Thus,

$$P_{\widetilde{k}} \leq (1+\alpha)\,(P_{\widetilde{k}-1} - P_{\widetilde{k}}), \quad \forall\, \widetilde{k} \geq \max\{k_0, k_1\},$$

which implies that

$$P_{\widetilde{k}} \leq \rho \cdot P_{\widetilde{k}-1}, \quad \forall\, \widetilde{k} \geq \max\{k_0, k_1\}, \tag{4.42}$$

where $\rho = \frac{1+\alpha}{2+\alpha} \in (0,1)$. This together with (4.37) gives the linear convergence rate

$$\|W_k - W^\star\|_F \leq \mathcal{O}(\rho^{k-\overline{k}}),\ \forall\, k \geq \overline{k}. \tag{4.43}$$

where $\overline{k} = \max\{k_0, k_1\}$.

- *Case III*: $\theta \in (1/2, 1)$. This case means $\frac{1-\theta}{\theta} \leq 1$. Based on the former results, we have

$$P_{\widetilde{k}} \leq (1+\alpha)\left[P_{\widetilde{k}-1} - P_{\widetilde{k}}\right]^{\frac{1-\theta}{\theta}}, \quad \forall\, \widetilde{k} \geq \max\{k_0, k_1\}.$$

We now run into the same situation as in [2]. Hence following a similar argument gives

$$P_{\widetilde{k}}^{\frac{1-2\theta}{1-\theta}} - P_{\widetilde{k}-1}^{\frac{1-2\theta}{1-\theta}} \geq \zeta,\ \forall\, k \geq \overline{k}$$

for some $\zeta > 0$. Then repeating and summing up the above inequality from $\overline{k} = \max\{k_0, k_1\}$ to any $k > \overline{k}$, we can conclude

$$P_{\widetilde{k}} \leq \left[P_{\widetilde{k}-1}^{\frac{1-2\theta}{1-\theta}} + \zeta(\widetilde{k} - \overline{k})\right]^{-\frac{1-\theta}{2\theta-1}} = \mathcal{O}\left((\widetilde{k} - \overline{k})^{-\frac{1-\theta}{2\theta-1}}\right).$$

Finally, the following sublinear convergence holds

$$\|W_k - W^\star\|_F \leq \mathcal{O}\left((k - \overline{k})^{-\frac{1-\theta}{2\theta-1}}\right),\ \forall\, k > \overline{k}. \tag{4.44}$$

We end this proof by commenting that both linear and sublinear convergence rate are closely related to the KL exponent θ at the critical point $W^\star$.

5 Experiments

In this section, we are going to apply our proposed spherical PCA to both synthetic data and real-world data sets to test the performance of our proposed method. The experiment on synthetic data will be performed first, followed by experiments on real-world data sets.

5.1 Synthetic Data Experiment We first generate 200 data, half of which is distributed within the region between $X = Z$ and Z axis (denoted as blue dots in the top part of Fig. 2), while another group is generated within the region between $Y = Z$ and Z axis (denoted as the red dots). These two clusters of data are generated through different angles. Thus when we do clustering, it should be the angle distance rather than the Euclidean distance to determine the clustering result. For our method, we learn a projection matrix $U \in \mathbb{R}^{3\times 2}$ and plot the component matrix $V \in \mathbb{R}^{2\times 200}$ as the bottom part illustrates. We see that, the Euclidean distance-based method (such as K-means) will yield poor clustering result (middle part), while spherical PCA will obtain good clustering result.

Also, we show the convergence of $\{W_k\}_{k\geq 0} = \{(U_k, V_k)\}_{k\geq 0}$ generated by our method. As Fig. 3 shows, after short iterations, the generated sequences will be stable, which is in accordance with the convergence proof. It also illustrates the objective with update. We see that it converges fast with a sublinear rate, which validates our convergence rate analysis.

5.2 Real-world Datasets Experiment It is known that in information retrieval, similarities or dissimilarities (proximities) between objects are more critical than Euclidean distance. In this subsection, we will test our proposed method on the widely-used 20-newsgroup dataset for clustering. We have different newsgroups such as: *comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast, etc.*. 200 documents are randomly sampled from each newsgroup. The word-document matrix X is constructed with 500 words selected according to the mutual information between words and documents. *Tf.idf* term weighting is used before normalization. Clustering accuracy are computed using the known class labels. Results will be compared including clustering accuracy (Acc.) and Normalized Mutual Information (NMI) [23].

Different clustering algorithms will be compared including:

1. **R1-PCA**, which proposes a rotational invariant ℓ_1-norm PCA, where a robust covariance matrix will soften the effects of outliers [7];

2. **K-SVD**, which is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data [1];

3. **PCA**, *i.e.*the vanilla PCA method in Eq. (1.2) without the constraint on G, which will be Euclidean distance-based by default;

4. **NMF** Matrix Factorization proposed by [11,13,14, 20] where U and V are obtained by Multiplicative Updating Algorithm with nonnegative constraint

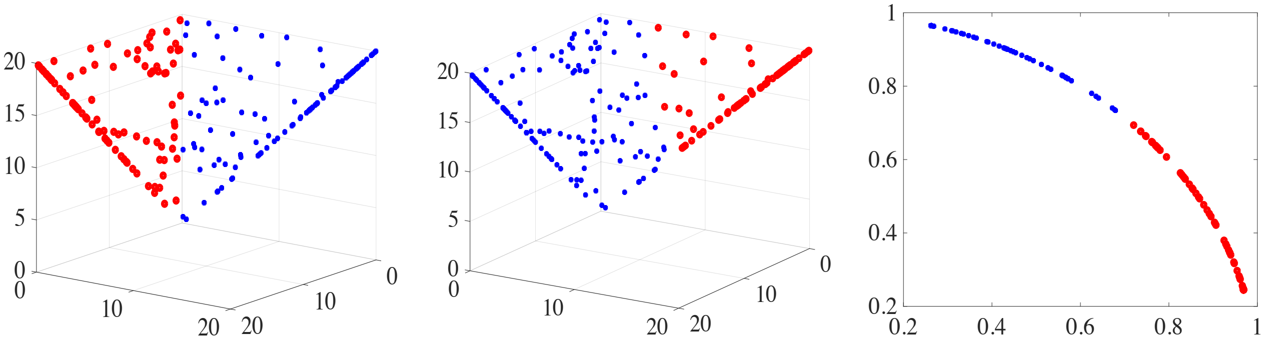

Figure 2: **Top**: two groups of data generated from two angles. **Middle**: clustering result with distance -based method K-means. **Bottom**: clustering result with our method. Blue and red represent different clusters.

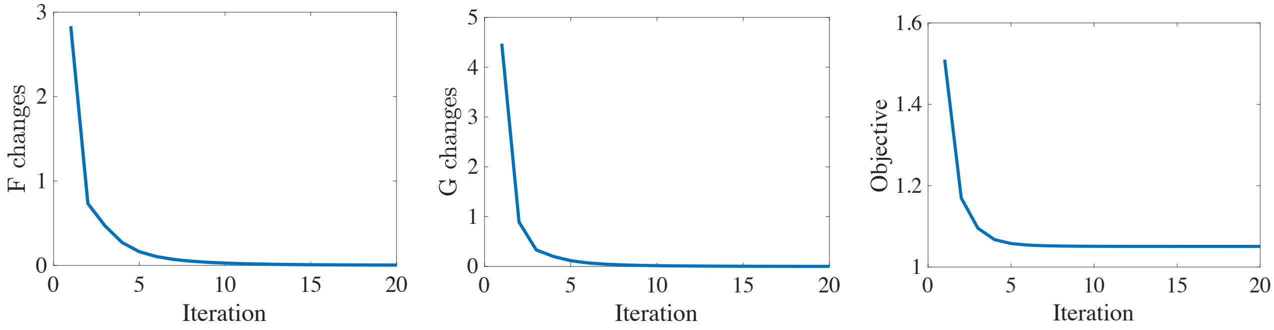

Figure 3: **Left**: $\|U_{k+1} - U_k\|_F$ with updates. **Center**: $\|V_{k+1} - V_k\|_F$ with updates. Both converge to 0 after several iterations. **Right**: Objective converges at sub-linear rate. All validate our analysis.

Table 1: Clustering performance of different algorithms on 20-newsgroup dataset

Methods	K-means		MUA		PCA		R1-PCA		K-SVD		**Spherical PCA**	
#Groups	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
5	0.651	0.621	0.674	0.614	0.703	0.628	0.745	0.647	0.789	0.673	**0.838**	**0.695**
10	0.487	0.316	0.478	0.320	0.502	0.383	0.535	0.398	0.527	0.394	**0.588**	**0.401**
15	0.398	0.307	0.387	0.301	0.412	0.319	0.423	0.320	0.461	0.377	**0.486**	**0.385**
20	0.315	0.242	0.314	0.221	0.362	0.248	0.394	0.260	0.412	0.280	**0.431**	**0.294**

Table 2: Clustering performance of different algorithms on four UCI data sets

Methods	K-means		MUA		PCA		R1-PCA		K-SVD		**Spherical PCA**	
Data (#class)	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
glass (6)	0.687	0.566	0.692	0.574	0.732	0.608	0.769	0.626	**0.801**	**0.648**	0.788	0.635
diabetes (2)	0.775	0.632	0.788	0.654	0.761	0.613	0.808	0.631	0.827	0.672	**0.832**	**0.680**
mfeat (10)	0.365	0.223	0.358	0.211	0.371	0.225	**0.431**	**0.342**	0.412	0.328	0.425	0.330
isolet (26)	0.267	0.198	0.253	0.181	0.262	0.182	0.324	0.201	0.357	0.246	**0.373**	**0.250**

5. ***K*-means** [9].

We vary the number of clusters from 5 to $10, 15$ and 20. In each newsgroup, 200 documents are randomly sampled, and we repeat for 10 times by taking the average and report the clustering result as Table 1 demonstrates.

We see that our proposed method Spherical PCA can always achieve both higher clustering accuracy and normalized mutual information in text analysis.

We also compare our method with other methods on UCI data sets including: *glass, diabetes, mfeat* and *isolet*. Table 2 illustrates the results. We see that though our method doesn't show the absolute advantage as on text, still the result is considerably good.

All the experiments indicate that our method can achieve good performance on both text and non-text data sets, showing its potential for broader application.

6 Conclusion

In this paper, we study spherical PCA where the direction matrix is orthonormal and the component vectors are assumed to lie in the unitary sphere. The benefit is obvious that it can make the angle distance equivalent to Euclidean distance. Due to the nonconvexity of objective function and constraints on the factors which are difficult to tackle, we propose an alternating linearized minimization method to derive the solution, which is proved to be sequence convergent. Moreover, we analyze the convergence rate which is validated by our experiments. The results on real-world datasets and synthetic data illustrate the superiority of our method.

References

[1] Michal Aharon, Michael Elad, Alfred Bruckstein, et al. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311, 2006.

[2] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

[3] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[5] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[6] Patrick L Combettes and Teemu Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

[7] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288. ACM, 2006.

[8] Hsieh Hou. A fast recursive algorithm for computing the discrete cosine transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10):1455–1461, 1987.

[9] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[10] Alexander Kaplan and Rainer Tichatschke. Proximal point methods and nonconvex optimization. *Journal of global Optimization*, 13(4):389–406, 1998.

[11] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2001.

[12] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*.

[13] Kai Liu and Hua Wang. Robust multi-relational clustering via l1-norm symmetric nonnegative matrix factorization. In *ACL*, volume 2, pages 397–401, 2015.

[14] Kai Liu and Hua Wang. High-order co-clustering via strictly orthogonal and symmetric ℓ_1-norm nonnegative matrix tri-factorization. In *IJCAI*, pages 2454–2460, 2018.

[15] Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pages 709–720. Springer, 2010.

[16] R Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976.

[17] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[18] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[19] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[20] Hua Wang, Heng Huang, and Chris Ding. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 279–284. ACM, 2011.

[21] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *International Conference on Machine Learning (ICML)*, pages 352–360, 2013.

[22] Hua Wang, Feiping Nie, Heng Huang, and Fillia Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *IJCAI*, 2011.

[23] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[24] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.