A HOMOTOPY OPTIMIZATION METHOD FOR ORTHOGONAL NON-NEGATIVE MATRIX FACTORIZATION

Ya Liu, Mingjie Shao and Wing-Kin Ma

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR of China

ABSTRACT

Data clustering is a key problem in data science and machine learning. In this paper, we consider orthogonal non-negative matrix factorization (ONMF) for scaled data clustering. The non-convex orthogonality constraint of ONMF raises a great challenge from an optimization viewpoint. We study a convex-constrained transformation of ONMF that allows us to control the approximation accuracy and problem difficulty through a parameter. We then apply a homotopy strategy in which we trace the solution path of a sequence of the aforementioned transformed problems, gradually moving from easy problems to near-ONMF problems. Intuitively, doing so may allow us to avoid local minima. Numerical results show that our homotopy method yields competitive clustering performance in synthetic data experiments and in a real-data hyperspectral clustering experiment.

***Index Terms*—** clustering, orthogonal non-negative matrix factorization, homotopy optimization

1. INTRODUCTION

Clustering data without supervision stands as a key problem in data science and machine learning [1,2]. In document clustering, documents can be classified into different topics [3]; in remote sensing, the pixels of a hyperspectral image can be identified as different materials [4]. Given a corpus of non-negative data points $\boldsymbol{z}_i \in \mathbb{R}_+^M$, $i = 1, \ldots, N$, clustering is to group them into $R < N$ clusters, with high intra-cluster similarity and low inter-cluster similarity. Let $\boldsymbol{u}_r \in \mathbb{R}_+^M$, $r = 1, \ldots, R$, be the centroids of the R clusters. The most widely-used clustering method, K-means clustering [5], aims to find, for each $\boldsymbol{z}_i$, a cluster index $l_i \in \{1, \ldots, R\}$ such that

$$\boldsymbol{z}_i \approx \boldsymbol{u}_{l_i}. \tag{1}$$

K-means clustering also aims at finding the centroids $\boldsymbol{u}_r$'s together with the cluster assignments l_i's.

In this paper, we are interested in a scaled variant of K-means clustering. We want to have

$$\boldsymbol{z}_i \approx \alpha_i \boldsymbol{u}_{l_i}, \tag{2}$$

for some scaling $\alpha_i \geq 0$. This is motivated by the fact that, in many applications, the collected data can be scaled, such as pixels affected by illumination conditions in imaging clustering, word statistics influenced by the document length in document clustering, etc. Clustering scaled data can be formulated as an orthogonal non-negative matrix factorization (ONMF) problem [3], as will be shown in this paper.

In this paper, we are interested in developing an efficient optimization method for tackling ONMF. This is a problem that has received much interest in signal processing, data science, optimization, machine learning, and related areas. The main challenge of ONMF lies in the non-convex orthogonality constraint. Many methods have been proposed to find an approximate solution of ONMF, such as: (1) the multiplicative update (MU) extended from classical non-negative matrix factorization (NMF) [6], wherein the orthogonality constraint is either penalized [3, 7] or implicitly addressed by using the gradient component residing in the tangent space of the Stiefel manifold [8]; (2) manifold algorithms, which penalize the non-negative constraint and use manifold algorithms for handling the orthogonality constraint [9]; (3) direct non-convex methods, such as the non-convex projected gradient method [4], block coordinate descent method [10] and non-convex penalty method [11]; (4) the ε-net approximation method [12], which randomly draws a large number of candidate solutions to form an ε-net of the feasible region. The ε-net approximation method is the only one that can provide global approximation guarantee, but it suffers from high computational complexity. The other methods are computationally more efficient and are usually considered in practice. Also, they can yield reasonable performance in practice.

Our method for tackling ONMF contains several ingredients: a careful reformulation of ONMF, a homotopy optimization method and an efficient first-order algorithm. Our method is a non-convex method, but the use of homotopy optimization makes a difference. The reformulation turns the ONMF problem into a convex-constrained one, and there is a parameter that controls the balance of approximation accuracy and problem difficulty. The homotopy method works by progressively changing the parameter such that, the reformulated problem is gradually handled from an easy (convex) but coarse approximation of ONMF to a hard but accurate approximation of ONMF. In this process, tracking the solution

This work was supported by a General Research Fund (GRF) of Hong Kong Research Grant Council (RGC), under Project ID CUHK 14208819.

path may give us a better chance to avoid bad local minima, as our empirical experience suggests. For more background of homotopy optimization, please refer to [13, 14] and the references therein. The proposed homotopy method is demonstrated to match or outperform many existing ONMF algorithms on both synthetic data experiments and a real data experiment in the application of hyperspectral clustering.

2. ONMF AS SCALED K-MEANS CLUSTERING

Consider the scaled clustering model (2). We argue that the scaled clustering problem can be formulated as an ONMF problem. Re-express (2) as

$$\boldsymbol{Z} \approx \boldsymbol{U}\boldsymbol{X},$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N], \boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R]$,

$$\boldsymbol{X} = [\alpha_1 \boldsymbol{e}_{l_1}, \ldots, \alpha_N \boldsymbol{e}_{l_N}], \tag{3}$$

$\boldsymbol{e}_i \in \mathbb{R}^R$ is the unit vector with the ith entry being 1 and the other entries 0. Since each column $\boldsymbol{x}_i$ of $\boldsymbol{X}$ has only one non-zero element, the rows $\check{\boldsymbol{x}}_r^T$'s of $\boldsymbol{X}$ are orthogonal, i.e., $\check{\boldsymbol{x}}_r^T \check{\boldsymbol{x}}_l = 0$ for $r \neq l$. This motivates us to consider the following ONMF formulation [3] for finding the cluster centroid $\boldsymbol{U}$ and the clustering indicator matrix $\boldsymbol{X}$:

$$\begin{aligned} \min_{\boldsymbol{U} \in \mathbb{R}^{M \times R}, \boldsymbol{X} \in \mathbb{R}^{R \times N}} \ & \|\boldsymbol{Z} - \boldsymbol{U}\boldsymbol{X}\|_F^2 \\ \text{s.t.} \ \ & \boldsymbol{U} \geq \boldsymbol{0}, \ \boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}, \ \boldsymbol{X} \geq \boldsymbol{0}. \end{aligned} \tag{4}$$

One can show that the constraint $\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}, \ \boldsymbol{X} \geq \boldsymbol{0}$ is equivalent to (3) with $\sum_{i:l_i=r} \alpha_i^2 = 1$, and we can assume $\sum_{i:l_i=r} \alpha_i^2 = 1$ without loss of generality.

ONMF as an equivalent formulation of the scaled clustering problem was discussed in the literature [3, 4]. Curiously, a direct explanation such as the one above was not seen.

3. OUR METHOD

The ONMF problem (4) is difficult mainly because the orthogonality constraint $\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}$ is non-convex and on manifold. The method to be presented seeks to use a careful reformulation to tackle the issue.

3.1. Reformulation

To start with, observe that, given any feasible $\boldsymbol{X}$ of problem (4), the optimal solution to problem (4) with respect to $\boldsymbol{U}$ is $\boldsymbol{U} = \boldsymbol{Z}\boldsymbol{X}^T$ when $\boldsymbol{Z} \geq \boldsymbol{0}$. By putting this $\boldsymbol{U}$ into (4), the ONMF problem (4) is simplified to

$$\min_{\boldsymbol{X}} f(\boldsymbol{X}) \triangleq -\|\boldsymbol{Z}\boldsymbol{X}^T\|_F^2 \quad \text{s.t.} \ \boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}, \ \boldsymbol{X} \geq \boldsymbol{0}. \tag{5}$$

To facilitate our subsequent development, let us re-express problem (5) as

$$\min_{\boldsymbol{X} \in \mathcal{O}} f(\boldsymbol{X}) + \mathbb{1}_{\mathbb{R}_+^{M \times N}}(\boldsymbol{X}), \tag{6}$$

where $\mathcal{O} \triangleq \{\boldsymbol{X} \mid \boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}\}$ and

$$\mathbb{1}_{\mathbb{R}_+^{M \times N}}(\boldsymbol{X}) = \begin{cases} 0, & \text{if } \boldsymbol{X} \geq 0 \\ +\infty, & \text{otherwise.} \end{cases}$$

Next, we consider the orthogonality constraint $\boldsymbol{X} \in \mathcal{O}$. As a folklore result, we have the following equivalence:

$$\boldsymbol{X} \in \mathcal{O} \Leftrightarrow \{\boldsymbol{X} \mid \boldsymbol{X} \in \mathcal{S}, \|\check{\boldsymbol{x}}_r\|_2 = 1, r = 1, \ldots, R\}, \tag{7}$$

where $\mathcal{S} \triangleq \{\boldsymbol{X} | \|\boldsymbol{X}\|_2 \leq 1\}$, and $\|\boldsymbol{X}\|_2 \triangleq \sigma_{\max}(\boldsymbol{X})$ denotes the spectral norm of $\boldsymbol{X}$, with $\sigma_{\max}(\boldsymbol{X})$ being the largest singular value of $\boldsymbol{X}$. In the Appendix, we provide the proof of (7). Taking insight from (7), we propose to approximate problem (6) by

$$\min_{\boldsymbol{X} \in \mathcal{S}} f(\boldsymbol{X}) + \mathbb{1}_{\mathbb{R}_+^{M \times N}}(\boldsymbol{X}) - \lambda \|\boldsymbol{X}\|_F^2 \tag{8}$$

for some λ. Intuitively, the penalty term $-\lambda\|\boldsymbol{X}\|_F^2$ with a large $\lambda > 0$ promotes large row length $\|\check{\boldsymbol{x}}_r\|_2^2$ for all r; on the other hand, one can show that the constraint $\boldsymbol{X} \in \mathcal{S}$ restricts $\|\check{\boldsymbol{x}}_r\|_2^2 \leq 1$ for all r. As a result, by applying a sufficiently large $\lambda > 0$, the optimal solution to problem (8) should satisfy $\|\check{\boldsymbol{x}}_r\|_2^2 = 1$ for all r. This idea is inspired by our recent work on binary optimization [15]; here, we extend the idea to deal with the orthogonality constraint.

3.2. Homotopy Optimization

Problem (8) admits a convex constraint, but it is still non-convex for a general λ. This motivates us to consider the following homotopy optimization method. Note that $f(\boldsymbol{X})$ is ρ-weakly convex [16] for $\rho \geq 2\sigma_{\max}^2(\boldsymbol{Z})$; that is, $f(\boldsymbol{X}) + \frac{\rho}{2}\|\boldsymbol{X}\|_F^2$ is convex. Thus, for $\lambda = -\rho/2$, problem (8) is a convex approximation of problem (6), an "easy" problem; for a large $\lambda > 0$, problem (8) should approach the original problem (6). Homotopy optimization takes the following strategy, also shown in Algorithm 1: Suppose λ is gradually changed from $\lambda = -\rho/2$ to a large $\lambda > 0$. The landscape of problem (8) should slightly change between two successive λ's. Starting from a small $\lambda_1 = -\rho/2$, we solve problem (8), which is convex, and obtain an optimal solution $\boldsymbol{X}^1$; next, λ_1 is slightly increased to λ_2 and we expect that a solver warm-started by $\boldsymbol{X}^1$ should get to an optimal solution $\boldsymbol{X}^2$ to the new problem. By continuing the above process and tracing the solution path of $\boldsymbol{X}^k$ for increasing λ, we argue that the homotopy method may stand a good chance to avoid bad local minima and to find a globally optimal solution.

3.3. An Approximation

We need a solver for handling problem (8) for a fixed λ. In this regard, the non-smooth indicator function $\mathbb{1}_{\mathbb{R}_+^{M \times N}}(\boldsymbol{X})$ hinders efficient algorithms. We approximate it by the smooth square distance function

$$\mathbb{1}_{\mathbb{R}_+^{M \times N}}(\boldsymbol{X}) \approx \mu \cdot \text{dist}(\boldsymbol{X}, \mathbb{R}_+^{M \times N})^2, \tag{9}$$

Algorithm 1 A homotopy strategy for tackling problem (8)

given initialization $\boldsymbol{X}^0$, $\lambda_1 = -\rho/2$, iteration index $k = 0$.
repeat
  $k = k + 1$
  warm-start from $\boldsymbol{X}^{k-1}$, run a solver to compute a solution to problem (8), set the solution as $\boldsymbol{X}^k$
  set λ_{k+1} as an increased version of λ_k
until some stopping criterion is satisfied.

for some $\mu > 0$, where

$$\text{dist}(\boldsymbol{X}, \mathbb{R}_+^{M\times N}) \triangleq \min_{\boldsymbol{Y}\in\mathbb{R}_+^{M\times N}} \|\boldsymbol{X} - \boldsymbol{Y}\|_F = \|\min(\boldsymbol{X}, \boldsymbol{0})\|_F$$

is the distance function from a point $\boldsymbol{X}$ to the set $\mathbb{R}_+^{M\times N}$. This leads to an approximation of problem (8)

$$\min_{\boldsymbol{X}\in\mathcal{S}} G_\mu(\boldsymbol{X}) - \lambda\|\boldsymbol{X}\|_F^2, \tag{10}$$

where $G_\mu(\boldsymbol{X}) \triangleq f(\boldsymbol{X}) + \mu\text{dist}(\boldsymbol{X}, \mathbb{R}_+^{M\times N})^2$. With (9), the approximation errors between problems (10) and (8), and between problem (6) and its approximation

$$\min_{\boldsymbol{X}\in\mathcal{O}} G_\mu(\boldsymbol{X}) \tag{11}$$

can be small given a large μ. Consider the following fact, of which the proof is omitted here due to the limited space and will be provided in the extended version of this work.

Fact 1 *Any optimal solution $\boldsymbol{X}^\star$ to problem* (10) *or* (11) *is $\mathcal{O}(1/\sqrt{\mu})$ non-negative, i.e., $\|\min(\boldsymbol{X}^\star, 0)\|_F^2 = \mathcal{O}(1/\mu)$.*

Previously, we provide the intuition why problem (8) may work. But we did not provide theoretical justification. As it turns out, the technical difficulty is also $\mathbb{1}_{\mathbb{R}_+^{M\times N}}(\boldsymbol{X})$. With the approximation in (9), we can analyze the effect of λ on enforcing orthogonality in problem (10).

Theorem 1 *Let L be a Lipschitz constant of $G_\mu(\boldsymbol{X})$ on $\mathcal{S}$, which must exist. For any $\lambda > L$, any globally optimal solution to problem* (10) *is also a globally optimal solution to problem* (11)*. Also, any globally optimal solution to problem* (11) *is a globally optimal solution to problem* (10)*.*

Theorem 1 is an extension of [15, Theorem 1], and we omit the details here due to space limitation.

3.4. An Efficient First-Order Algorithm

We complete our method by presenting a solver for tackling problem (10) for a fixed λ. We choose the majorization-minimization (MM) method. By utilizing the inequality $-\|\boldsymbol{X}\|_F^2 \le -\|\bar{\boldsymbol{X}}\|_F^2 - 2\langle\bar{\boldsymbol{X}}, \boldsymbol{X}\rangle$ for any $\bar{\boldsymbol{X}}$, we can construct a surrogate problem of problem (10) at the $(k+1)$th MM iteration as follows:

$$\boldsymbol{X}^{k+1} \in \arg\min_{\boldsymbol{X}\in\mathcal{S}} H_{\lambda,\mu}(\boldsymbol{X};\boldsymbol{X}^k) \triangleq G_\mu(\boldsymbol{X}) - 2\lambda\langle\boldsymbol{X}^k, \boldsymbol{X}\rangle, \tag{12}$$

where $\boldsymbol{X}^k$ is the solution to the kth MM iteration. Then, we can apply the projected gradient method to handle problem (12); there, the key operation is the projection onto $\mathcal{S}$, which can be obtained by singular value thresholding, i.e., $\Pi_{\mathcal{S}}(\boldsymbol{X}) \triangleq \arg\min_{\boldsymbol{Y}\in\mathcal{S}} \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 = \boldsymbol{W}\min(\boldsymbol{\Sigma}, 1)\boldsymbol{Q}^T$, where $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{Q}^T$ is the singular value decomposition (SVD) of $\boldsymbol{X}$.

In fact, what we adopt is an advanced version of MM, namely, gradient-extrapolated MM (GEMM) [15]. In GEMM, the MM subproblem (12) is updated by a one-step accelerated projected gradient (APG). GEMM is computationally light due to the inexact APG update. Also, GEMM is guaranteed to converge to a stationary point of problem (10) [15].

Finally, we assemble all the building blocks and show the pseudo code of the overall algorithm in Algorithm 2: $1/\beta_k$ is the step size; λ and μ are updated by the subgradient update rule in [14]. Note that we also progressively increase the value of the non-negativity penalty parameter μ.

Algorithm 2 Homotopy Method and GEMM for ONMF

given initialization $\lambda_1 = -\rho/2$, μ_1, $\boldsymbol{X}^0$, $\delta_0 = 0$, c_λ, c_μ, $\beta_0 = 2\sigma_{\max}^2(\boldsymbol{Z})$, $k = 0$.
repeat
  $k = k + 1$
  $\beta_k = \beta_0 + 2\mu_k$
  $j = 0$, warm-start with $\hat{\boldsymbol{X}}^1 = \hat{\boldsymbol{X}}^0 = \boldsymbol{X}^{k-1}$
  repeat
    $j = j + 1$
    $\alpha_j = \frac{\delta_{j-1}-1}{\delta_j}$, $\delta_j = \frac{\sqrt{1+4\delta_{j-1}^2}}{2}$
    $\boldsymbol{V}^j = \hat{\boldsymbol{X}}^j + \alpha_j(\hat{\boldsymbol{X}}^j - \hat{\boldsymbol{X}}^{j-1})$
    $\hat{\boldsymbol{X}}^{j+1} = \Pi_{\mathcal{S}}(\boldsymbol{V}^j - 1/\beta_k\nabla H_{\lambda,\mu}(\boldsymbol{V}^j; \hat{\boldsymbol{X}}^j))$
  until convergence, set the output $\hat{\boldsymbol{X}}^{j+1}$ as $\boldsymbol{X}^k$
  $\lambda_k = \lambda_{k-1} + \frac{c_\lambda}{\sqrt{k+1}}(R - \|\boldsymbol{X}^k\|_F^2)$
  $\mu_k = \mu_{k-1} + \frac{c_\mu}{\sqrt{k+1}}\|\min(\boldsymbol{X}^k, 0)\|_F^2$
until convergence.

4. SIMULATION RESULTS AND CONCLUSION

In this section, we examine the performance of our proposed homotopy method on both synthetic and real-world experiments, particularly, under the application of hyperspectral clustering. For benchmarking, we consider K-means clustering and five state-of-the-art ONMF algorithms, namely, DTPP [3], ONMF-S [8], HALS [10], ONPMF [4] and NCP [11]. The first two methods are variants of the NMF multiplicative update (MU) [6], while the last three methods are direct non-convex methods. We specifically choose the initialization strategies suitable for each algorithm. K-means clustering is initialized by randomly selected R data points as centroids; DTPP, ONMF-S, HALS and NCP by the Nonnegative Double Singular Value Decomposition (NNDSVD) [17];

ONPMF and our homotopy method by SVD initialization [4]. We evaluate the performance by clustering accuracy, i.e., the proportion of correctly clustered data points.

In Algorithm 2, we set $\lambda_1 = 1$, $\mu_1 = 0.1$, $c_\lambda = 1.5$ and $c_\mu = 1.2$. Within each iteration k, we run GEMM until the distance of successive iterates satisfies $\|\hat{\boldsymbol{X}}^{j+1} - \hat{\boldsymbol{X}}^{j}\|_F \leq 10^{-5}$ or the maximum iteration number $10,000$ is reached. The whole homotopy algorithm terminates when the distance of successive iterates satisfies $\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^{k}\|_F \leq 10^{-6}$ or the maximum iteration number $40,000$ is reached.

4.1. Synthetic Data Experiments

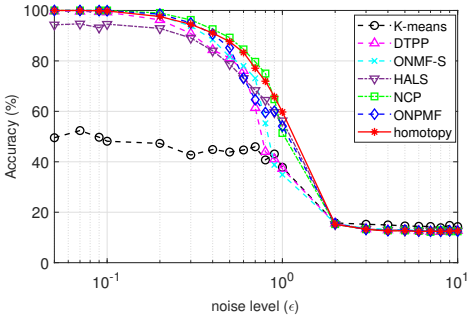
In synthetic data experiments, we generate $\boldsymbol{X}$ by (3) where α_i is randomly generated over $[0.1, 1]$. We consider imbalanced cluster sizes, i.e., each row $\check{\boldsymbol{x}}_i$ in $\boldsymbol{X}$ has $500 - 50(i-1)$ nonzero elements, $i = 1, \ldots, R$. The factor $\boldsymbol{U}$ is generated in two ways: 1) $\boldsymbol{U} \in \mathbb{R}_+^{300\times 10}$, with each element $u_{i,j}$ independent and identically distributed (i.i.d.) and uniformly distributed on $[0, 1]$; 2) randomly selected $R = 5$ hyperspectral signature vectors of length $M = 224$ from a subset of the USGS library [18]. Note that the first simulates a generic instance, while the second a hyperspectral clustering problem. Then, the data matrix $\boldsymbol{Z}$ is generated by $\boldsymbol{Z} = \max\{\boldsymbol{U}\boldsymbol{X} + \boldsymbol{N}, \boldsymbol{0}\}$, where each element $n_{i,j}$ of $\boldsymbol{N}$ is i.i.d. white Gaussian, specifically, $n_{i,j} \sim \mathcal{N}(0, \epsilon^2)$. A number of 10 Monte-Carlo trials were used to assess the performance of the various clustering algorithms.

Fig. 1 shows the clustering accuracies under different noise power levels. It is seen that the ONMF-based methods achieve much higher clustering accuracies than K-means clustering when the data points are scaled. In Fig. 1(a), except for K-means clustering, the other methods achieve comparably good clustering performance, with DTPP and ONMF-S performing slightly worse at moderate noise levels and with HALS performing slightly worse at low noise levels. In Fig. 1(b), it is seen that the proposed homotopy method achieves a higher clustering accuracy than the other methods. For the tested simulations in Fig. 1, the runtime of the proposed homotopy method is comparable to those of NCP and ONPMF; the other methods cost less time.

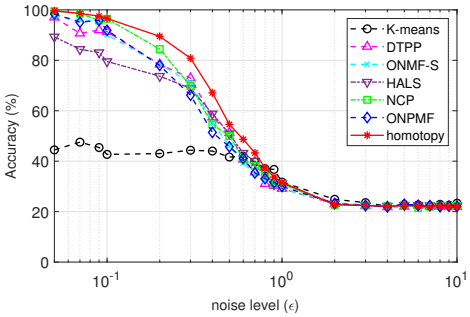
4.2. A Real Data Experiment: Hyperspectral Clustering

We consider a clustering problem arising from hyperspectral imaging. A hyperspectral image contains a collection of images taken at different wavelengths and at the same scene. Different materials admit different spectral responses. Hyperspectral clustering is to identify the main material contained in each pixel.

The real data we test is the HYDIC Urban hyperspectral image [19,20], which contains $M = 162$ clean spectral bands and $N = 307 \times 307$ pixels. There are mainly six materials in the data: roof, dirt, metal, asphalt road, grass and tree;



(a) Randomly generated $\boldsymbol{U}$ from $[0, 1]^{300\times 10}$.



(b) Hyperspectral $\boldsymbol{U}$ from the USGS library.

Fig. 1: Accuracy results for different ONMF algorithms.

the labels are accessed from [20]. Fig. 2 shows the clustering results obtained by different algorithms together with the clustering accuracies; different colors are used to represent different materials. It is seen that ONPMF and our homotopy method are the best two in this task; both can extract most of the materials correctly, with the homotopy method achieving a higher clustering accuracy. ONPMF performs not so well in distinguishing road and dirt, while our homotopy method is less satisfactory in discovering metal.

In this work, we have proposed a homotopy optimization method for handling the ONMF problem in data clustering. The homotopy solution-path tracking idea yields promising results, as our numerical experiments showed.

5. REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[2] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, 2005.

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

[4] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur, "Two algorithms for orthogonal nonnegative matrix factoriza-

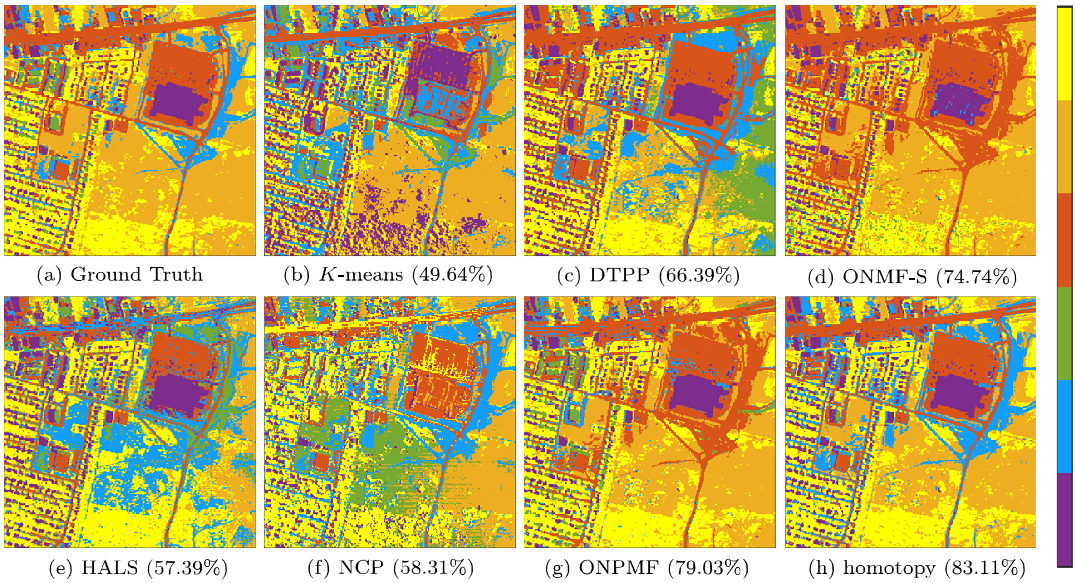


Fig. 2: Hyperspectral clustering for Urban data. Different colors in the colorbar represent different materials, from top to down: tree, grass, asphalt road, metal, dirt, roof.

tion with application to clustering," *Neurocomput.*, vol. 141, pp. 15–25, 2014.

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

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] Z. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 734–749, 2010.

[8] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. Int. Joint Conf. Neural Netw.* IEEE, 2008, pp. 1828–1832.

[9] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1, pp. 397–434, 2013.

[10] K. Kimura, Y. Tanaka, and M. Kudo, "A fast hierarchical alternating least squares algorithm for orthogonal nonnegative matrix factorization," in *Asian Conf. Machine Learn.*, 2015, pp. 129–141.

[11] S. Wang, T. Chang, Y. Cui, and J. Pang, "Clustering by orthogonal non-negative matrix factorization: A sequential non-convex penalty approach," in *Proc. Int. Conf. Acoust, Speech, Signal Process. (ICASSP)*. IEEE, 2019, pp. 5576–5580.

[12] M. Asteris, D. Papailiopoulos, and A. G. Dimakis, "Orthogonal nmf through subspace exploration," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 343–351.

[13] D. M. Dunlavy and D. P. O'Leary, "Homotopy optimization methods for global optimization," *Report SAND2005-7495, Sandia National Laboratories*, 2005.

[14] M. Shao and W.-K. Ma, "Binary MIMO detection via homotopy optimization and its deep adaptation," *IEEE Trans. Signal Process.*, vol. 69, pp. 781–796, 2021.

[15] M. Shao, Q. Li, W.-K. Ma, and A. M.-C. So, "A framework for one-bit and constant-envelope precoding over multiuser massive MISO channels," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5309–5324, 2019.

[16] E. Nurminskii, "The quasigradient method for the solving of the nonlinear programming problems," *Cybernetics*, vol. 9, no. 1, pp. 145–150, 1973.

[17] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[18] R. F. Kokaly *et al.*, "USGS spectral library version 7," US Geological Survey, Reston, VA, USA, USGS Numbered Series 1035, Tech. Rep., 2017.

[19] "US Army Corps of Engineers," http://www.tec.army.mil/Hypercurbe.

[20] "Remote Sensing Laboratory," https://rslab.ut.ac.ir/data, accessed on Nov. 18, 2020.

6. APPENDIX

If $\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}$, then we have $\sigma_r(\boldsymbol{X}) = 1$ for all r, where $\sigma_r(\boldsymbol{X})$ denotes the rth largest singular value of matrix $\boldsymbol{X}$; and $\|\check{\boldsymbol{x}}_r\|_2 = 1$ for all r. Thus, the right-hand side (RHS) of (7) holds. Conversely, if the RHS of (7) holds, we have $\|\boldsymbol{X}\|_F^2 = \sum_{r=1}^R \|\check{\boldsymbol{x}}_r\|_2^2 = R$. Also, note that $\|\boldsymbol{X}\|_F^2 = \sum_{r=1}^R \sigma_r^2(\boldsymbol{X})$. As $\|\boldsymbol{X}\|_2 \leq 1$ implies $\sigma_r(\boldsymbol{X}) \leq 1$ for all r, we have $\sigma_r(\boldsymbol{X}) = 1$ for all r, which arrives at the orthogonality of $\boldsymbol{X}$ on the left-hand side of (7).