Large-Cone Nonnegative Matrix Factorization

Tongliang Liu, Mingming Gong, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Nonnegative matrix factorization (NMF) has been greatly popularized by its parts-based interpretation and the effective multiplicative updating rule for searching local solutions. In this paper, we study the problem of how to obtain an attractive local solution for NMF, which not only fits the given training data well but also generalizes well on the unseen test data. Based on the geometric interpretation of NMF, we introduce two large-cone penalties for NMF and propose large-cone NMF (LCNMF) algorithms. Compared with NMF, LCNMF will obtain bases comprising a larger simplicial cone, and therefore has three advantages. 1) the empirical reconstruction error of LCNMF could mostly be smaller; (2) the generalization ability of the proposed algorithm is much more powerful; and (3) the obtained bases of LCNMF have a low-overlapping property, which enables the bases to be sparse and makes the proposed algorithms very robust. Experiments on synthetic and real-world data sets confirm the efficiency of LCNMF.**

***Index Terms*— Generalization, good local solution, large simplicial cone, nonnegative matrix factorization (NMF), robustness, sparseness.**

I. Introduction

NONNEGATIVE matrix factorization (NMF) allows for the factorization of a nonnegative data matrix X into two nonnegative low-rank matrices W and H. Lee and Seung [27] interpreted NMF as a dimensionality reduction method having the property of learning parts-based representations. Unlike other traditional dimensionality reduction methods [40], [47]–[49], the bases and represented data matrix of NMF are required to be nonnegative. Nonnegativity allows only additive combinations, which forces NMF to learn parts-based representations. Lee and Seung [28] provided an effective multiplicative updating rule (MUR) for optimizing NMF. NMF, therefore, has been greatly popularized and successfully applied to many scientific fields, for example, computer vision [21], [55], machine learning [35], [58], and neuroscience [16].

Since NMF is convex with respect to either W or H but not to both, MUR usually converges at a local solution [29], [32]. In practice, good local solutions, which have relatively small empirical reconstruction errors, are much more preferred. Donoho and Stodden [10] discussed under what conditions NMF has a zero reconstruction error, i.e., $X = WH$. The problem of when $X = WH$ was also studied for a number of special cases of NMF before NMF was formally introduced by Paatero and Tapper [39]. Thomas [51] provided a sufficient and necessary condition for rank-r NMF (where X is of rank r and W has r columns). Kaykobad [23] shows a sufficient condition for symmetric NMF (where $W = H^T$). From their results, we know that in most cases, the solutions to NMF have nonzero reconstruction errors, i.e., $X \neq WH$. Thus, to obtain local solutions having smaller empirical reconstruction errors than those having conventional NMF are both necessary and highly important.

Regarding the generalization analysis in the learning theory, it is always interested in designing algorithms to have good generalization abilities. Many different regularization penalties, such as smoothness and sparseness, are frequently employed for this objective based on the inherent properties of the data X or the prior knowledge about the applications. These popular regularization penalties can certainly help improve the generalization ability of NMF [20], [24]. This improvement can be explained from the perspective of variance and bias [9]. However, based on the geometric interpretation of NMF that it seeks a proper simplical cone to try to contain X, we propose a natural yet meaningful regularization, the large-cone penalty, for NMF to simultaneously obtain a small empirical reconstruction error, and a good generalization ability. We term this new kind of algorithms large-cone NMF (LCNMF).

We propose two approaches to formulate the large-cone penalty, where the formulations are both convex with respect to the bases W. One approach models the volume of the intersection of the parallelotope that is spanned by the bases W and the unit Euclidean ball. The central idea of the approach is that the value of the Gram determinant of the bases W is equal to the square of the volume of the parallelotope. The other method maximizes the volume of the intersection of the simplicial cone and the unit Euclidean ball. It exploits the idea that the volume of the intersection of the simplicial cone and the unit Euclidean ball is an increasing function of the summation of the pairwise angles between the bases.

Compared with NMF, LCNMF is able to obtain better local solutions.[1] When analyzing the objectives of LCNMF and NMF, it seems that the large-cone penalty introduces a tradeoff between the volume of simplicial cone and the empirical reconstruction error of NMF. This tradeoff does indeed not always hold, because, from the geometrical interpretation viewpoint, NMF having a larger simplicial cone interpretation may result in a smaller empirical reconstruction error. We also summarize the generalization bounds for NMF and theoretically prove that LCNMF will lead

Manuscript received August 19, 2015; revised January 3, 2016; accepted May 26, 2016. This work was supported by the Australian Research Council under Project DP-140102164 and Project FT-130101457.

The authors are with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: tliang.liu@gmail.com; mingming.gong@student.uts.edu.au; dacheng.tao@uts.edu.au).

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

Digital Object Identifier 10.1109/TNNLS.2016.2574748

[1] In this paper, comparing the solution with the conventional NMF, a better local solution is always referred to as a local solution having a smaller reconstruction error for the training data, which is defined in (6), and a better generalization ability for the future data.

to a smaller generalization upper bound for NMF under the Vapnik–Chervonenkis learning theory framework [53]. Moreover, experiments on real-world data sets show that the reconstruction error on the training data and test data will be simultaneously decreased because of the large-cone penalty.

In this paper, our contributions are the geometric interpretation of LCNMF and the theoretical and empirical illustrations that LCNMF can surprisingly and simultaneously have the properties of sparseness on the bases of factorization, data clustering, and good local solutions. We note that some existing works have implicitly used the large-cone penalty for NMF, for example, the orthogonal NMF (OrNMF) studied in [54], [59], and [8], which can be interpreted as the large angle NMF, and thus LCNMF. However, the attracting properties for LCNMF have only been partly illustrated. For example, Zheng *et al.* [61] showed that the orthogonality penalty of NMF (the same as the large-angle NMF proposed in this paper) results in sparse bases due to the low-overlapping property; Theis *et al.* [50] and Gillis [15] found that NMF with sparseness constraints may give unique factorizations; Yoo and Choi [59], Ding *et al.* [8], and Ding *et al.* [6] illustrated that the orthogonal NMF can be clearly interpreted for the clustering problems. In particular, Ding *et al.* [6] proved that the row (or column) K-means clustering of data matrix X is equivalent to the bases (or representation) orthogonality constrained NMF. However, to the best of our knowledge, none of them have presented another attracting property of LCNMF: good local solutions. Since detailed theoretical and empirical analyses for the sparseness and clustering properties of LCNMF have been provided in the aforementioned references, we do not repeat them again in this paper. We instead focus on illustrating the good local solution property of LCNMF.

A. Related Work

Interestingly, small-cone constraints on W have also been proposed for NMF [38], [63], [64]. In particular, Zhou *et al.* [63] theoretically and empirically showed that the small-cone constraint on the bases W will impose suitable sparseness on H, which does not contradict our assertion but can be used to explain that the large-cone constraint on W will result in sparseness on W, because there exists a geometric tradeoff between W and H.[2] Notably, Zhou *et al.* [63] also empirically pointed out that the minimum-volume-constrained NMF may have a small reconstruction error on the training data. In this paper, we will provide theoretical analysis for this phenomenon. Moreover, we will theoretically and empirically show that LCNMF will have small reconstruction errors on both the training and test data.

There are numerous results on NMF and its variants. We then briefly present a summary of some related variants instead of a survey on all these results. Based on the models, we divide these variants of NMF into four categories.

1) NMF using different loss functions: as with the Kullback–Leibler (KL) divergence used in [28], many other divergences have also been exploited for NMF [13]. Sandler and Lindenbaum [44] factorized a nonnegative matrix using the earth mover's distance. To robustly factorize the nonnegative matrix, different robust loss functions have been used, such as the ℓ_1 loss [18], [34] and the ℓ_{21} loss [25].
2) NMF using different regularizations: regularizations will enforce some desirable properties in NMF. The lasso-based sparseness penalty [22] and the orthogonality penalty [30], [61] will help to achieve sparse representations. Cai *et al.* [5] developed a graph-based approach for parts-based representation to consider the geometric structure in the data. This manifold regularization will make the representation of NMF more discriminative for classification problems.
3) NMF with different hard constraints: some hard constraints, such as the orthogonality constraint, are very restrictive to NMF. However, these constraints will provide NMF with interesting interpretations. Ding *et al.* [8] proved that the hard orthogonality constraint leads to a rigorous clustering interpretation of NMF. Ding *et al.* [7] extended NMF to semi-NMF, which allows the data matrix X and the new representations H to have mixed signs and introduces a subtractive property, thereby extending the applicable range of NMF methods.
4) NMF exploiting different structures of data: finding structures in data is important and may lead to fast and efficient NMF. Recht *et al.* [43] proposed a linear programming model to compute NMF for separable data. Arora *et al.* [1] proved that polynomial-time algorithms exist for exact and approximate NMF and also provided an efficient algorithm that runs in time polynomial under the separability condition. Takeuchi *et al.* [46] assumed that multiple nonnegative matrices might share similar sparse nonnegative bases and used auxiliary matrices to better estimate the bases. Ding *et al.* [7] extended NMF to the product of three factor matrices and proposed the convex NMF.

There are, of course, other interesting NMF variants that are not included in the above four categories, for example, the kernel NMF [4], [57] and the partially shared latent factor NMF [33]. The interested reader is referred to further examples in the survey [56].

[2] Some simple algebra shows that adding a large-cone constraint on W is equivalent to adding a small-cone constraint on H. Then, according to the theories in Zhou *et al.* [63], adding a large-cone constraint on W will result in the sparseness on W. In particular, we have

$$\begin{aligned}
&\det((WH)^T(WH)) \\
&\quad = \det((WQQ^{-1}H)^T(WQQ^{-1}H)) \\
&\quad = \det(H^T(Q^{-1})^T Q^T W^T W Q Q^{-1} H) \\
&\quad = \det(H^T \det(Q^{-1})^T)\det(Q^T)\det(W^T W)\det(Q)\det(Q^{-1})\det(H) \\
&\quad = \det(W^T W)\det(Q^T Q) \times \frac{\det(H^T H)}{\det(Q^T Q)}.
\end{aligned}$$

According to the above equations, to find a large-cone (or small-cone) bases $W' = WQ$ implies to find an invertible matrix Q with a large (or small) $\det(Q^T Q)$. Then, the corresponding representation matrix $H' = Q^{-1}H$ will have a small-cone (or large-cone) interpretation, because $\det(H'^T H') = (\det(H^T H)/\det(Q^T Q))$.

II. Background

A. Notations

We denote $X = (x_1, \ldots, x_n) \in \mathbb{R}_+^{m\times n}$ as the nonnegative data matrix consisting of n independent and identically distributed examples drawn from $\mathcal{X} \subset \mathbb{R}^m$ with a Borel measure ρ. We use X_i to denote the ith column of matrix X and X_{ij} to denote the ijth entry of matrix X. Let H be an $r \times n$ matrix. We also use $h \in \mathbb{R}_+^r$ to represent one column of H and $P\{A\}$ to indicate the probability of event A. The notations $\|\cdot\|$, $\langle\cdot,\cdot\rangle$, and $\|\cdot\|_F$ always refer to the Euclidean norm, the inner product, and the Frobenius norm, respectively.

B. Nonnegative Matrix Factorization

NMF seeks a factorization of a nonnegative data matrix $X \in \mathbb{R}_+^{m\times n}$ into two nonnegative low-rank matrices $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{r\times n}$. This can be formulated as

$$\min_{W\in\mathbb{R}_+^{m\times r}, H\in\mathbb{R}_+^{r\times n}} \ell(X, WH)$$

where $\ell(x, y)$ denotes the loss function quantifying the cost of using y to approximate x and the reduced dimensionality $r < \min(m, n)$. The nonnegative constraints lead to parts-based representations, because they only allow additive combinations. The two most popular loss functions used for NMF are the square (Euclidean distance) loss and the KL divergence

$$\min_{W\in\mathbb{R}_+^{m\times r}, H\in\mathbb{R}_+^{r\times n}} \|X - WH\|_F^2 = \sum_{i=1}^{n} (X_i - WH_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \|X_{ij} - (WH)_{ij}\|^2 \quad (1)$$

and

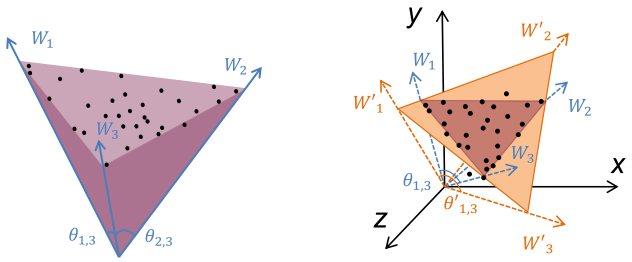
$$\min_{W\in\mathbb{R}_+^{m\times r}, H\in\mathbb{R}_+^{r\times n}} D(X||WH) = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(X_{ij} \log \frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij} \right).$$

As discussed previously, $X \neq WH$ in most cases, thus NMF seeks two nonnegative matrices W and H as a local solution to approximate X. Because NMF is a powerful dimensionality reduction tool, we call the columns of W the bases of NMF (or bases).

C. Geometric Interpretation of NMF

NMF has a geometric interpretation [10], based on which we will propose the large-cone penalty. In this section, we will also place proper constraints (normalization on W) to simplify the setting of NMF.

The equation $X = WH$, where $X \in \mathbb{R}_+^{m\times n}$, $W \in \mathbb{R}_+^{m\times r}$, and $H \in \mathbb{R}_+^{r\times n}$, implies that every column of X can be represented as a nonnegative linear combination of the columns of W. The algebraic characterization can be described using the simplicial cone as follows.



(a) (b)

Fig. 1. Geometric interpretation of NMF. Black points: columns of X (supposing we have scaled them onto the specific hyperplane). (a) NMF seeks a simplicial cone spanned by W to approximately contain X. (b) Simplicial cone spanned by W' is larger than the simplicial cone spanned by W.

Definition 1 (Simplicial Cone): A simplicial cone generated by W is a spanned set

$$C_W = \left\{ x : x = \sum_i^r h_i W_i, h_i \in \mathbb{R}_+ \right\}.$$

The extreme rays of a simplicial cone are defined as follows.

Definition 2 (Extreme Ray): An extreme ray of a simplicial cone C is a ray $R_x = \{ax : a \geq 0\}$, where $x \in C$ cannot be represented as a proper convex combination of any two points x_0 and x_1, which belong to C but not R_x.

The goal of NMF is to find a simplical cone C_W, which can approximately contain $\mathcal{X} \subseteq \mathbb{R}^m$, as shown in Fig. 1(a). Note that such a simplicial cone is contained in the nonnegative orthant.

According to Definition 2, an extreme ray has infinite representations because of the scalar parameter a. This is the same for the bases of NMF, because $WH = WQ^{-1}QH$, where $Q \in \mathbb{R}^{r\times r}$ is an invertible matrix, such that $WQ > 0$ and $Q^{-1}H > 0$. We can normalize the bases W by choosing

$$Q = \begin{pmatrix} \|W_1\| & & & \\ & \|W_2\| & & \\ & & \ddots & \\ & & & \|W_r\| \end{pmatrix}$$

to limit the choice of W without changing any representation ability of the solution to an NMF problem.

Moreover, if the bases W are normalized, the new representations will have the same upper bound as that of the data point. Assume that the data point $x \in \mathcal{X}$ is upper bounded by R. As we have normalized every column of W to be a unit, we have

$$\|x\| \approx \|Wh\| = \sqrt{\sum_{j=1}^{n} \left(\sum_{i=1}^{r} W_{ji} h_i \right)^2} \geq \|h\|.$$

We therefore can imagine that the norm of the column of H should also be no more than R. Under the same setting, Maurer and Pontil [37] proved that $\|h\| \leq R$.

In some cases, for a given data matrix X, there may be many possible simplicial cones containing the data matrix.

As discussed in [10], if C_W is a simplicial cone containing the data matrix X and $C_{W'}$ is another cone containing the first, i.e., $C_W \subset C_{W'}$, then the vectors W' can furnish a representation of the data matrix as well. Based on the geometric interpretation of NMF, we intuitively hope that the simplicial cone generated by W could be as large as possible, because the large cone has the capacity to contain more data points, and thus induces a small empirical reconstruction error and a good generalization.

III. Large-Cone Nonnegative Matrix Factorization

We have discussed that in most cases $X \neq WH$ and that NMF usually converges at local solutions, which are sensitive with respect to the initializations. In these situations, can we place some constraints on the simplicial cone to enable NMF to achieve better performance? The answer is "yes." Because NMF solvers converge at local solutions, larger simplicial cones may lead to better local solutions, which have smaller empirical reconstruction errors. Moreover, larger simplicial cones will intuitively result in better generalizations. We then propose a large-cone penalty for NMF and introduce the LCNMF algorithm as follows:

$$\min_{W\in\mathbb{R}_+^{m\times r},H\in\mathbb{R}_+^{r\times n}} \frac{1}{n}\sum_{i=1}^{n}\ell(X_i, WH_i) + \lambda L(W) \tag{2}$$

where $\lambda > 0$ is the regularization parameter and $L(W)$ is an decreasing function of the volume of the simplicial cone C_W. We call $L(W)$ the large-cone penalty. In the rest of this paper, we will consider the least square loss functions, that is, $\ell(x, y) = (x - y)^2$. Note that the MUR also applies to LCNMF. To enable LCNMF to obtain good local solutions for NMF, we should design the large-cone penalty $L(W)$ to be convex with respect to W.

We propose two approaches to formulate the large-cone penalty. One approach relies on the idea that the value of the Gram determinant of the bases W is equal to the square of the volume of the parallelotope formed by the bases. The other method is based on the idea that the volume of the intersection of the simplicial cone and the unit Euclidean ball is an increasing function of the summation of the pairwise angles of the bases W.

A. Large Volume Nonnegative Matrix Factorization

In this section, we introduce the large-cone penalty that models the volume of the intersection of the parallelotope that is spanned by the bases W and the unit Euclidean ball. We begin with introducing the definition of the Gram determinant.

Definition 3: The Gram determinant of a matrix $X \in \mathbb{R}^{m\times n}$ is defined as

$$\det(X^TX) = \begin{vmatrix} \langle X_1, X_1\rangle & \langle X_1, X_2\rangle & \dots & \langle X_1, X_n\rangle \\ \langle X_2, X_1\rangle & \langle X_2, X_2\rangle & \dots & \langle X_2, X_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle X_n, X_1\rangle & \langle X_n, X_2\rangle & \dots & \langle X_n, X_n\rangle \end{vmatrix}.$$

The following inequality, which shows that the Gram determinant is nonnegative, is well known as the Gram's inequality [11]

$$\det(X^TX) \geq 0 \quad \forall X \in \mathbb{R}^{m\times n}.$$

Geometrically, the Gram determinant of X is equal to the square of the volume of the parallelotope formed by the vectors $\{X_1, \dots, X_n\}$. Therefore, we can formulate the large-cone penalty as

$$L(W) = -\frac{\sqrt{\det(W^TW)}}{\Pi_{i=1}^{r}\|W_i\|}.$$

We have shown that the normalization of W will not change the setting and geometric interpretation of NMF. Let $\|W_i\| = 1, i = 1, \dots, r$. Thus, the large-cone penalty can be written as

$$L(W) = -\sqrt{\det(W^TW)}.$$

However, the derivative of the above function is complicated. For simplicity, we further write the penalty as the following log-determinant function:

$$L(W) = -\log(\det(W^TW)) \tag{3}$$

whose derivative is $-2(W^+)^T$, where W^+ denotes the pseudoinverse matrix of W. Moreover, the NMF problems where $X \neq WH$ implies that the learned bases for NMF should be of full column rank. Then, the log-determinant function is convex with respect to the bases W [12], [62]. Let the loss function ℓ be the square loss function. The LCNMF algorithm (2) can therefore be specialized as

$$\begin{aligned} &\min_{W,H} \frac{1}{n}\|X - WH\|_F^2 - \lambda\log(\det(W^TW)) \\ &\text{s.t. } \|W_i\| = 1, \quad i = 1, \dots, r \\ &\qquad W \in \mathbb{R}_+^{m\times r}, \quad H \in \mathbb{R}_+^{r\times n}. \end{aligned} \tag{4}$$

Since minimizing $L(W)$ in (3) is equal to maximizing the volume of the intersection of the parallelotope that is spanned by the bases W and the unit Euclidean ball, we call algorithm (4) the large volume NMF (LVNMF).

B. Large Angle Nonnegative Matrix Factorization

In this section, we introduce the large-cone penalty that can maximize the volume of the intersection of the simplicial cone and the unit Euclidean ball.

As shown in Fig. 1, to make the simplicial cone that is spanned by the bases W as large as possible, intuitively, the pairwise angles between the bases should be as large as possible, where the angles are denoted by $\theta_{i,j}, i \neq j \in \{1, \dots, r\}$. Fig. 1(b) shows our purpose. Let $L(\theta)$ be a decreasing function of $\theta_{i,j}$. The objective function of LCNMF is, therefore, formulated as

$$\min_{W\in\mathbb{R}_+^{m\times r},H\in\mathbb{R}_+^{r\times n}} \frac{1}{n}\sum_{i=1}^{n}\ell(X_i, WH_i) + \lambda L(\theta).$$

We have that $\cos\theta_{i,j} = (\langle W_i, W_j\rangle/\|W_i\|\|W_j\|)$, which is a decreasing function of $\theta_{i,j}$, because W is nonnegative.

Since the normalization of W will not change the setting and geometric interpretation of NMF, we let $\|W_i\| = 1$, $i = 1, \ldots, r$. Then, $\cos\theta_{i,j} = \langle W_i, W_j \rangle$. Let

$$L(\theta) = \sum_{i \neq j} \beta_{i,j} \cos^2 \theta_{i,j}, \quad \beta_{i,j} \in \mathbb{R}_+$$

where $\beta_{i,j}$ are the regularization parameters for $\cos^2\theta_{i,j}$, $i, j = 1, \ldots, r$. We then have

$$L(\theta) = \|B \circ (W^T W - I)\|_F^2$$

where B is a symmetric matrix with the ijth entry of $\sqrt{\beta_{i,j}}$, I denotes the $r \times r$ identity matrix, and $X \circ Y$ denotes the Hadamard product of X and Y.

Let the loss function ℓ be the square loss function and take B as a matrix with all entries being 1. The LCNMF algorithm (2) can, therefore, be specialized as

$$\begin{aligned} &\min_{W,H} \ \frac{1}{n}\|X - WH\|_F^2 + \lambda \|W^T W - I\|_F^2 \\ &\text{s.t. } \|W_i\| = 1, \quad i = 1, \ldots, r \\ &\qquad W \in \mathbb{R}_+^{m \times r}, \quad H \in \mathbb{R}_+^{r \times n}. \end{aligned} \tag{5}$$

We call the algorithm (5) the large angle NMF (LANMF).

Note that $L(\theta) = \|W^T W - I\|_F^2$ is convex with respect to W. Let $f(W) = W^T W$ and $g(Y) = \|Y - I\|_F^2$. Then, $L(\theta) = g(f(W))$. We know that f and g are convex functions. It can be concluded that $L(\theta)$ is convex with respect to W if we could prove that $g(Y)$ is nondecreasing with respect to Y. The derivative of $g(Y)$ is $2(Y - I)$. Since we have normalized W, the diagonal entries of $Y = W^T W$ are all equal to 1. Then, $2(Y - I)$ is nonnegative, which means that $g(Y)$ is nondecreasing with respect to Y.

IV. Theoretical Analysis of LCNMF

In this section, we first summarize the generalization bounds for NMF, which helps us better understand it. Then, we prove that LCNMF has a smaller upper generalization bound than NMF under the Vapnik–Chervonenkis learning theory framework. The main result is presented in Theorems 6 and 7.

Let the empirical reconstruction error of NMF be defined as follows:

$$R_n(W) = \frac{1}{n} \sum_{i=1}^{n} \min_{h \in \mathbb{R}_+^r} \|x_i - Wh\|^2. \tag{6}$$

And define the expected reconstruction error as

$$R(W) = E_x R_n(W).$$

The following theorem [3], proven using the Rademacher complexity and the Hoeffding's inequality, plays an important role in proving the generalization error bounds.

Theorem 1 [3]: Let F be an $[a, b]$-valued function class on $\mathcal{X}$, and $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ be independent and identically distributed examples. For any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\sup_{f \in F} \left(E_x f(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right) \leq \mathfrak{R}(F) + (b - a)\sqrt{\frac{\ln(1/\delta)}{2n}}$$

where the Rademacher complexity is defined as

$$\mathfrak{R}(F) = E_{\sigma,x} \sup_{f \in F} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i)$$

and $\sigma_1, \ldots, \sigma_n$ are independent Rademacher variables.

Maurer and Pontil [37] proved a dimensionality independent generalization bound for NMF by employing Theorem 1.

Theorem 2 [37]: For NMF problems, assume that x is upper bounded by 1. For any learned normalized W and any $\delta > 0$, with probability at least $1 - \delta$, we have

$$|R(W) - R_n(W)| \leq \frac{14r\sqrt{r}}{\sqrt{n}} + \sqrt{\frac{r^2 \ln(16nr)}{4n}} + \sqrt{\frac{\ln 2/\delta}{2n}}.$$

The following theorem [41], proven utilizing the covering number and Hoeffding's inequality, is also frequently used to derive generalization error bounds.

Theorem 3 [41]: Let F be an $[a, b]$-valued function class on $\mathcal{X}$, and $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ be independent and identically distributed examples. For any $\delta > 0$ and $\xi > 0$, with probability at least $1 - \delta$, we have

$$\begin{aligned} &\mathrm{P}\left\{ \sup_{f \in F} E_x f(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \geq \xi \right\} \\ &\qquad \leq 8E\mathcal{N}_1(F, \xi/8, X) \exp\left(-\frac{n\xi^2}{32(b-a)^2} \right) \end{aligned}$$

where $\mathcal{N}_1(F, \xi, X)$ is the covering number of $F(X) = \{f(x_1), \ldots, f(x_n) : f \in F\}$ at radius ξ with respect to the ℓ_1 norm metric.

The definition of covering number $\mathcal{N}_p(F, \epsilon, X)$ [60] is as follows.

Definition 4 [60]: Let B be a metric space with metric d. Given observations $X = (x_1, \ldots, x_n)$ and vectors $f(X) = (f(x_1), \ldots, f(x_n)) \in B^n$, the covering number in p-norm, denoted as $\mathcal{N}_p(F, \epsilon, X)$, is the minimum number m of a collection of vectors $v_1, \ldots, v_m \in B^n$, such that $\forall f \in F, \exists v_j$

$$\|d(f(X), v_j)\|_p = \left[\sum_{i=1}^{n} d\big(f(x_i), v_j^i\big)^p \right]^{1/p} \leq n^{1/p}\epsilon$$

where v_j^i is the ith component of vector v_j. We also define $\mathcal{N}_p(F, \epsilon, n) = \sup_X \mathcal{N}_p(F, \epsilon, X)$.

Based on Theorem 3, we provide a dimensionality dependent generalization bound for NMF [36], different from those in [17] and [52].

Theorem 4 [36]: For NMF problems, assume that x is upper bounded by 1. For any learned normalized W and any $\delta > 0$, with probability at least $1 - \delta$, we have

$$|R(W) - R_n(W)| \leq \frac{2}{n} + \sqrt{\frac{mr \ln(4(1 + r)\sqrt{m}rn) - \ln\frac{\delta}{2}}{2n}}.$$

A detailed proof of Theorem 4 is presented in Appendix A for completeness.

Theorems 2 and 4 are both proved under the Vapnik–Chervonenkis learning theory. Theorem 2 exploits the Rademacher complexity to measure the complexity of the whole predefined hypothesis class and Theorem 4 employs

the covering number likewise. Using the same methods, we can easily prove the following lemma.

Lemma 1: The induced Rademacher complexity and covering number of LCNMF are smaller than those of NMF.

Proof: The soft large-cone regularization shrinks the search space for optimizing the bases W. According the definitions of the Rademacher complexity and covering number, Lemma 1 is easily proved (see [2, Th. 12]). ■

For LCNMF problems, if the bases W are normalized, the new representations will also have the same upper bound as that of the data point.

Theorem 5 [36]: For LCNMF problems, if x is upper bounded by R, then every column of H is upper bounded by R, that is, $\|h\| \leq R$.

A detailed proof of Theorem 5 is presented in Appendix B for completeness.

Lemma 1 implies that the generalization bound of LCNMF is potentially smaller than that of NMF. However, deriving an explicit representation for the smaller upper generalization bound of LCNMF is quite complicated, so we instead present the follow theorem, which can be easily proved using Lemma 1 and Theorem 5, to give a slightly weaker but meaningful generalization bound for LCNMF.

Theorem 6: For LCNMF problems, assume that x is upper bounded by 1. For any learned normalized W and any $\delta > 0$, with probability at least $1 - \delta$, we have

$$|R(W) - R_n(W)| \leq \min \left\{ \frac{14r\sqrt{r}}{\sqrt{n}} + \sqrt{\frac{r^2 \ln(16nr)}{4n}} + \sqrt{\frac{\ln 2/\delta}{2n}}, \frac{2}{n} + \sqrt{\frac{mr \ln(4(1+r)\sqrt{m}rn) - \ln\frac{\delta}{2}}{2n}} \right\}.$$

To clarify that the generalization bound of LCNMF is smaller than that of NMF, we derive a dimensionality independent generalization bound for the orthogonal NMF, which is a special case of LANMF.

Theorem 7: For orthogonal NMF problems, assume that x is upper bounded by 1. For any learned normalized W and any $\delta > 0$, with probability at least $1 - \delta$, we have

$$|R(W) - R_n(W)| \leq 6r\sqrt{\frac{\pi}{n}} + \sqrt{\frac{\ln 2/\delta}{2n}}.$$

A detailed proof of Theorem 7 is presented in Appendix C.

Comparing the dimensionality independent generalization bounds in Theorems 2 and 7, we can clearly see that LCNMF will lead to a smaller generalization bound for NMF under the Vapnik–Chervonenkis learning theory framework.

However, a smaller generalization bound is not a sufficient condition for better performance. In Sections V and VI, we will empirically show that LCNMF gives better local solutions and generalizations than NMF.

V. Optimization of LCNMF

We provide optimization algorithms for LCNMF in this section before presenting the empirical results in Section VI.

Lee and Seung [28] developed MUR for optimizing NMF. We present the updating rule for the NMF algorithm (1) as follows:

$$H_{t+1} = H_t \circ \frac{W_t^T X}{W_t^T W_t H_t}$$
$$W_{t+1} = W_t \circ \frac{X H_{t+1}^T}{W_t H_{t+1} H_{t+1}^T}.$$

MUR employed for LCNMF is similar to that for NMF. However, the normalization constraint on W may introduce difficulty in proving the convergence of the algorithm. For simplicity, we use the block coordinate descent method by alternatively solving

$$H_{t+1} = \arg \min_{H \in \mathbb{R}_+^{r \times n}} \|X - W_t H\|_F^2$$

and

$$W_{t+1} = \arg \min_{W \in \mathfrak{W}} \frac{1}{n} \|X - W H_{t+1}\|_F^2 + \lambda L$$

where $\mathfrak{W} = \{W : \|W_i\| = 1, i = 1, \ldots, r, W \in \mathbb{R}_+^{m \times r}\}$ to optimize LCNMF. Most existing NMF solvers are special implementations of the block coordinate descent method [19], [31], [42].

When H is fixed, the objective function for optimizing W is convex. The normalization constraint on W may slow down the convergence rate, but it does not ultimately hinder the search for the optimal solution.

Let the objective function of LVNMF be

$$\text{Obj}_1(W, H) = \frac{1}{n} \|X - WH\|_F^2 - \lambda \log(\det(W^T W)).$$

The optimization algorithm for LVNMF is summarized in Algorithm 1. Let the objective function of LANMF be

$$\text{Obj}_2(W, H) = \frac{1}{n} \|X - WH\|_F^2 + \lambda \|W^T W - I\|_F^2.$$

The optimization algorithm for LANMF is summarized in Algorithm 2.

VI. Experiments

In this section, we detail experiments conducted on both synthetic and real data sets (including the extended YaleB data set [14], Multi-PIE data set [45], and MNIST [26]) to illustrate the performance of the proposed LCNMF. The results verify that LCNMF not only fits the given training data well but also generalizes well on the unseen test data. We also illustrate that LCNMF will converge at good local solutions even though the initializations are poor. The parameter λ was tuned among several values $\{10^{-6}, 10^{-5}, \ldots, 10^3\}$[3] and the λ having the smallest empirical reconstruction errors on the training data was finally chosen. All the averaged empirical reconstruction errors are based on ten trials.

[3]To make it easy to tune the tradeoff parameter λ, we have preprocessed all the data points to be on the unit ball.

Algorithm 1 LVNMF

Input: $X \in \mathbb{R}_+^{m\times n}$, $1 \le r \le \min(m,n)$ and λ
Output: $W \in \mathbb{R}_+^{m\times r}$, $H \in \mathbb{R}_+^{r\times n}$

Initialize $W_0 \in \mathbb{R}_+^{m\times r}$, $H_0 \in \mathbb{R}_+^{r\times n}$, $t = 0, \epsilon, \eta$
repeat
  repeat
    $k = 0$; $H_k = H_t$;
    % Gradient descent;
    $H_{\text{temp}} = H_k - \epsilon(W_t^T W_t H_k - W_t^T X)$;
    % Constraint satisfaction;
    $H_{k+1} = \max(0, H_{\text{temp}})$;
    $k = k + 1$;
  until $\frac{\text{abs}(\text{Obj}_1(W_t,H_{k-1}) - \text{Obj}_1(W_t,H_k))}{\text{Obj}_1(W_t,H_{k-1})} \le \eta$
  $H_{t+1} = H_k$;
  repeat
    $k = 0$; $W_k = W_t$;
    % Gradient descent;
    $W_{\text{temp}} = W_k - \epsilon(W_k H_{t+1} H_{t+1}^T - X H_{t+1}^T - \lambda n(\text{pinv}(W_k))^T$;
    % Constraint satisfaction;
    $W_{\text{temp}} = \max(0, W_{\text{temp}})$;
    % Constraint satisfaction;
    $W_{k+1} = \text{normc}(W_{\text{temp}})$;
    $k = k + 1$;
  until $\frac{\text{abs}(\text{Obj}_1(W_{k-1},H_{t+1}) - \text{Obj}_1(W_k,H_{t+1}))}{\text{Obj}_1(W_{k-1},H_{t+1})} \le \eta$
  $W_{t+1} = W_k$;
until $\frac{\text{abs}(\text{Obj}_1(W_t,H_t) - \text{Obj}_1(W_{t+1},H_{t+1}))}{\text{Obj}_1(W_t,H_t)} \le \eta$
$W = W_{t+1}$; $H = H_{t+1}$;

Algorithm 2 LANMF

Input: $X \in \mathbb{R}_+^{m\times n}$, $1 \le r \le \min(m,n)$ and λ
Output: $W \in \mathbb{R}_+^{m\times r}$, $H \in \mathbb{R}_+^{r\times n}$

Initialize $W_0 \in \mathbb{R}_+^{m\times r}$, $H_0 \in \mathbb{R}_+^{r\times n}$, $t = 0, \epsilon, \eta$
repeat
  repeat
    $k = 0$; $H_k = H_t$;
    % Gradient descent;
    $H_{\text{temp}} = H_k - \epsilon(W_t^T W_t H_k - W_t^T X)$;
    % Constraint satisfaction;
    $H_{k+1} = \max(0, H_{\text{temp}})$;
    $k = k + 1$;
  until $\frac{\text{abs}(\text{Obj}_2(W_t,H_{k-1}) - \text{Obj}_2(W_t,H_k))}{\text{Obj}_2(W_t,H_{k-1})} \le \eta$
  $H_{t+1} = H_k$;
  repeat
    $k = 0$; $W_k = W_t$;
    % Gradient descent;
    $W_{\text{temp}} = W_k - \epsilon(W_k H_{t+1} H_{t+1}^T - X H_{t+1}^T + \lambda n(W_k W_k^T W_k + W_k)$;
    % Constraint satisfaction;
    $W_{\text{temp}} = \max(0, W_{\text{temp}})$;
    % Constraint satisfaction;
    $W_{k+1} = \text{normc}(W_{\text{temp}})$;
    $k = k + 1$;
  until $\frac{\text{abs}(\text{Obj}_2(W_{k-1},H_{t+1}) - \text{Obj}_2(W_k,H_{t+1}))}{\text{Obj}_2(W_{k-1},H_{t+1})} \le \eta$
  $W_{t+1} = W_k$;
until $\frac{\text{abs}(\text{Obj}_2(W_t,H_t) - \text{Obj}_2(W_{t+1},H_{t+1}))}{\text{Obj}_2(W_t,H_t)} \le \eta$
$W = W_{t+1}$; $H = H_{t+1}$;

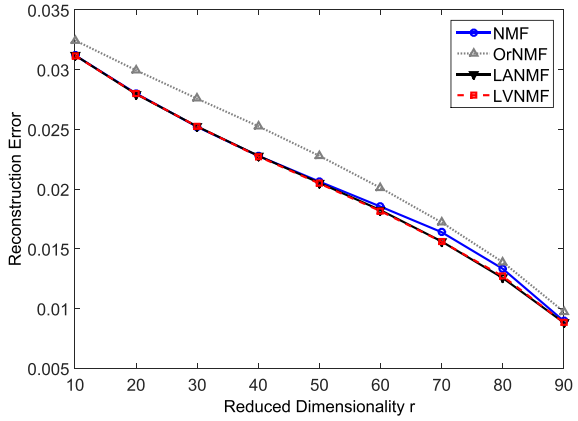


Fig. 2. Average empirical reconstruction errors on the synthetic data. For a large range of the reduced dimensionality r, the empirical reconstruction errors of LANMF and LVNMF are much smaller than those of NMF and OrNMF.

A. Experiments on Synthetic Data

We randomly generated a synthetic data matrix by $X = \text{rand}(100, 200)$ and compared the empirical reconstruction errors between NMF, OrNMF, LANMF, and LVNMF on the synthetic data set in Fig. 2. It shows that for a large range of the reduced dimensionality r, the empirical reconstruction errors of LVNMF and LANMF are smaller than those of NMF and OrNMF because of the proposed large-cone penalty.

Comparing the models of NMF (1), LANMF (5), and LVNMF (4), we find that NMF minimizes the empirical reconstruction error directly, while LANMF and LVNMF minimize the summation of the empirical reconstruction error and the large-cone penalty. However, LANMF and LVNMF can still achieve smaller empirical reconstruction errors because of the large-cone interpretation.

Note that NMF mostly converges at local solutions that are easily affected by the initializations. We further empirically illustrate that LANMF and LVNMF can converge at good local solutions even though the initializations are poor. As shown in Fig. 6(a), we initialized NMF, LANMF, and LVNMF with randomly generated initializations $W0$ and $H0$. LANMF and LVNMF converged at local solutions having smaller empirical reconstruction errors. If we initialize NMF by the outputs of LANMF and LVNMF, the empirical reconstruction errors of NMF will start to decrease from very small values. This directly shows that the outputs of LANMF and LVNMF are around the good local solutions for NMF and that they can serve as good initializations for NMF. Thus, we can conclude

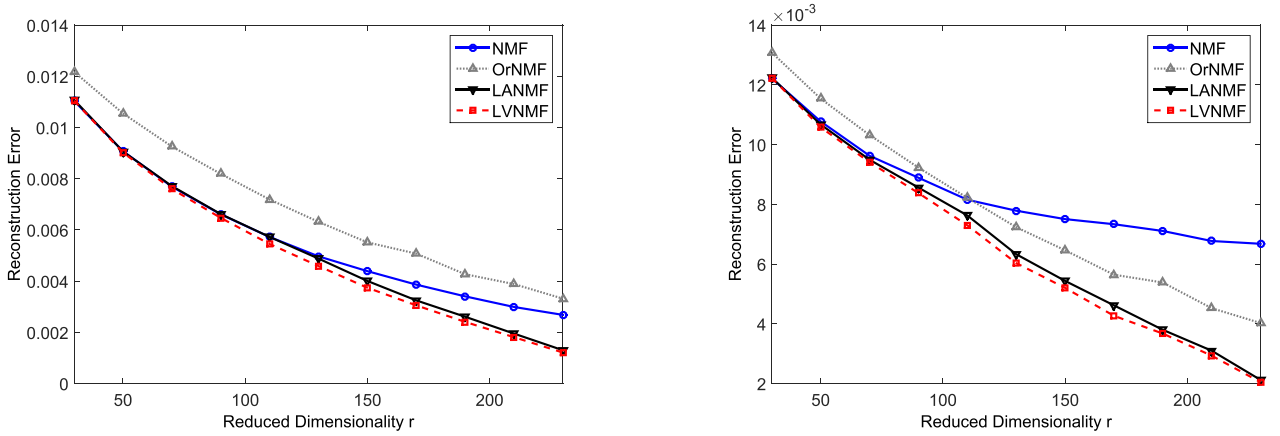


Fig. 3. Average empirical reconstruction errors on the extended YaleB data set. (a) Comparison of the reconstruction errors of NMF, OrNMF, LANMF, and LVNMF on the training data. For a large range of the reduced dimensionality r, the empirical reconstruction errors of LANMF and LVNMF are much smaller than those of NMF and OrNMF. (b) Comparison of the reconstruction errors on the test data. The empirical reconstruction errors of LANMF and LVNMF are always much smaller than those of NMF and OrNMF.

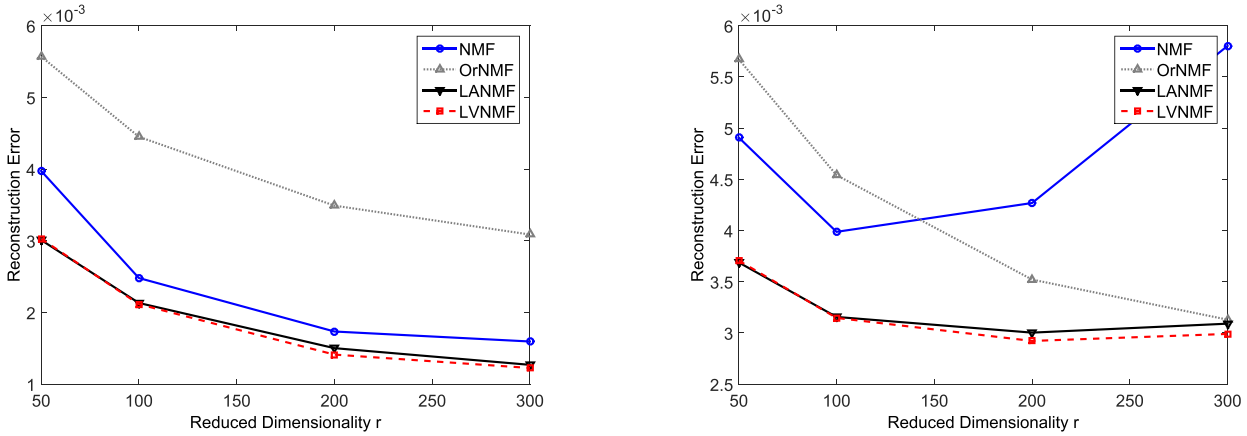


Fig. 4. Average empirical reconstruction errors on the Multi-PIE data set. (a) Reconstruction errors on the training set. (b) Reconstruction error on the test set. (a) and (b) show that the empirical reconstruction errors of LANMF and LVNMF are always much smaller than those of NMF and OrNMF.

that LANMF and LVNMF could find local solutions having smaller empirical reconstruction errors than NMF on the synthetic data.

B. Experiments on the Extended YaleB Data Set

The extended Yale Face Data Set B contains 16 128 images of 28 human subjects under 9 poses and 64 illumination conditions. We used the front pose face image and downsized the faces to 16×16 matrices, and 1000 examples were randomly chosen and then divided into independent training and test sets of 500 examples for each.

The comparison between the reconstruction errors of NMF, OrNMF, LVNMF, and LANMF on the training data is shown in Fig. 3(a), which demonstrates that the empirical reconstruction errors of LVNMF and LANMF are mostly smaller than those of NMF. Though the conventional NMF minimizes the empirical reconstruction error directly while LCNMF and LANMF do not, we show in Fig. 6(b) that LANMF and LVNMF can obtain smaller empirical reconstruction errors and converge to better local solutions with poor initializations. These imply that the large-cone penalties are powerful in helping find local solutions with smaller empirical reconstruction errors.

The comparison of the empirical reconstruction errors of NMF, OrNMF, LANMF, and LVNMF on the test data is shown in Fig. 3(b), where the errors were based on the bases learned from the training data. It shows that the large-cone penalty can significantly reduce the reconstruction errors on the test data. We thus conclude that the large-cone penalty contributes to obtaining much better generalizations.

Comparing Fig. 3(a) and (b), we find that the empirical reconstruction errors of LANMF and LVNMF on the training and test data have not changed much, and that for some dimensionality r, the empirical reconstruction errors of LANMF and LVNMF on the test data are even smaller than those of NMF

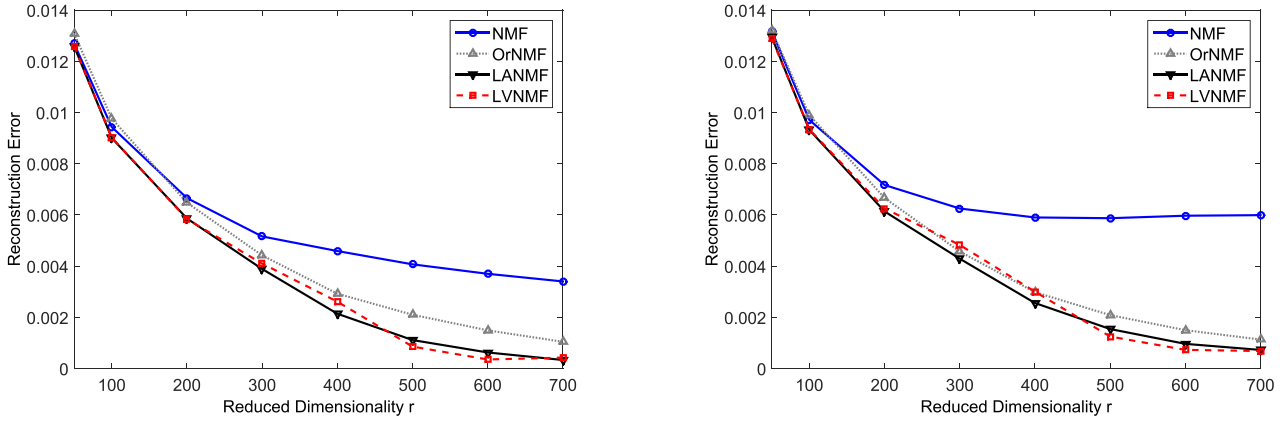


Fig. 5. Average empirical reconstruction errors on the MNIST data set. For a large range of the reduced dimensionality r, the empirical reconstruction errors of LANMF and LVNMF are much smaller than those of NMF and OrNMF. (a) Comparison of the reconstruction errors on the training data. (b) Comparison of the reconstruction errors on the test data.

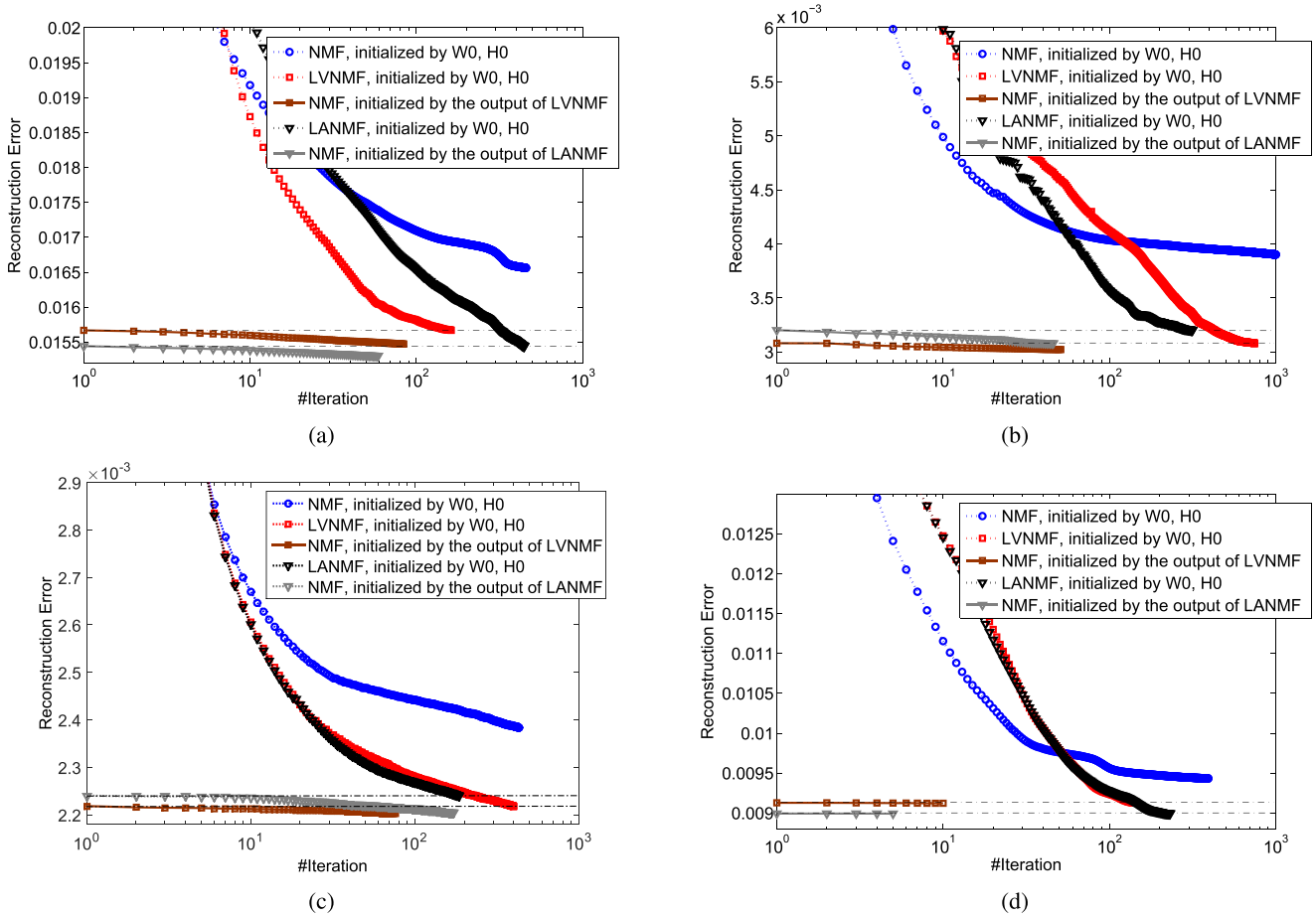


Fig. 6. Empirical reconstruction errors versus iteration numbers. The figures shows that using the same randomly generated inputs, LANMF and LVNMF converged with much smaller empirical reconstruction errors than NMF. (a) Synthetic data set and $r = 70$. (b) Extended YaleB data set and $r = 170$. (c) Multi-PIE data set and $r = 100$. (d) MNIST data set and $r = 100$.

on the training data, which shows that LANMF and LVNMF are not only very efficient for obtaining good local solutions but also very robust.

C. Experiments on the Multi-PIE Data Set

The Multi-PIE data set [45] contains more than 750 000 images of 337 people recorded in up to four sessions over the span of five months. Subjects were imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions. We randomly chose 1200 examples from session 1 and downsized the faces to 32×24 matrices, and then divided them into independent training and test sets of 600 examples for each.

The comparisons between the reconstruction errors of NMF, OrNMF, LVNMF, and LANMF on the training and test data are shown in Fig. 4(a), which demonstrates that the empirical reconstruction errors of LVNMF and LANMF are smaller than those of NMF and OrNMF and that LCNMF can simultaneously have smaller reconstruction errors on both the training and test sets.

In Fig. 6(c), we show that LANMF and LVNMF can obtain better local solutions than NMF when they are initialized by the same randomly generated matrices. This means that LCNMF is powerful in finding good local solutions to NMF on the Multi-PIE data set.

D. Experiments on the MNIST Data Set

The MNIST is a data set of handwritten digits. The digits have been size-normalized and centered in the matrices of size 28×28. There are 10 different handwritten digit numbers and 70 000 examples, which is of rank 713. We randomly chose 2000 examples and divided them into independent training and test sets of 1000 examples for each.

Fig. 5(a) and (b) compares the empirical reconstruction errors of NMF, OrNMF, LANMF, and LVNMF on the training and test data, respectively. They show that the empirical reconstruction errors of LANMF and LVNMF on both the training and test data are much smaller than those of NMF. In Fig. 6(d), we also show that LANMF and LVNMF can obtain good local solutions even though they have been initialized around bad local solutions, which illustrates that the proposed large-cone penalties are very powerful in finding better local solutions for NMF.

Comparing Fig. 5(a) and (b), we can see that the empirical reconstruction errors of NMF on the test data are much larger than those on the training data. However, the empirical reconstruction errors of LANMF, LVNMF, and OrNMF do not change much on training and test sets. This means that OrNMF, LANMF, and LVNMF are very robust algorithms.

Fig. 5 also shows that on the MNIST data set, our proposed methods perform slightly better than OrNMF. This is because the MNIST data set has a good clustering property (it can be easily clustered into ten clusters). In this case, an LCNMF may perform similar to an OrNMF.

From the results of the overall experiments on the synthetic and real-world data sets, we can empirically conclude that the proposed large-cone penalties can help not only fit the given training data well but also generalize well on the unseen test data.

VII. Conclusion

In this paper, we studied the problem of how to obtain attractive local solutions for NMF. We presented a large-cone penalty framework for NMF for this purpose. We first explained that the large-cone penalty could intuitively lead to good local solutions from the perspective of geometric interpretation. We then theoretically proved that LCNMF has a smaller upper generalization bound than NMF under the Vapnik–Chervonenkis learning theory framework, which was proved by showing that the induced Rademacher complexity and covering number of LCNMF are smaller than those of NMF. Finally, we empirically showed that the proposed LANMF and LVNMF could achieve local solutions that simultaneously have smaller reconstruction errors on the given training data and better generalization abilities for the future data on both synthetic and real-world data sets. The proposed large-cone penalty is convex and can be deployed to most existing NMF models.

Appendix

A. Proof of Theorem 4

The proof method is the same as that in [36]. We provide the proof here for completeness. Given the bases W, the representatives H are fixed because of convexity. Thus, the induced covering number can be determined by W. Let us define the induced hypothesis class as

$$F_{\mathcal{W}} = \left\{ f_W(x) = \min_{h \in \mathbb{R}_+^r} \left\| x - \sum_{i=1}^{r} W_i h_i \right\|^2 : W \in \mathbb{R}_+^{m \times r} \right\}.$$

We first give an upper bound for the induced covering number $\mathcal{N}_1(F_{\mathcal{W}}, \xi, n)$ of NMF.

Lemma 2: Let $F_{\mathcal{W}}$ be the loss function class induced by the reconstruction error with the bases class $\mathcal{W} = \{W : W \in \mathbb{R}_+^{m \times r}\}$. Then

$$\ln \mathcal{N}_1(F_{\mathcal{W}}, \xi, n) \leq mr \ln \left(\frac{4(R+r)\sqrt{mr}}{\xi} \right).$$

Proof: We will bound the covering number of the loss function class F_W by bounding the covering number of the bases class $\mathcal{W}$. Cutting the subspace $[-1, 1]^m \subset \mathbb{R}^m$ into small m-dimensional regular solids with width ξ, there are a total of

$$\left\lceil \frac{2}{\xi} \right\rceil^m \leq \left(\frac{2}{\xi} + 1 \right)^m \leq \left(\frac{4}{\xi} \right)^m$$

such regular solids. If we pick out the centers of these regular solids and use them to make up a W, there are

$$\left\lceil \frac{2}{\xi} \right\rceil^{mr} \leq \left(\frac{4}{\xi} \right)^{mr}$$

choices that are denoted by $\mathcal{S}$. Then, $|\mathcal{S}|$ is the upper bound of the ξ-cover of the bases class $\mathcal{W}$.

We will prove that for every W, there exists a $W' \in \mathcal{S}$, such that $|f_W - f_{W'}| \leq \xi'$, where $\xi' = (R+r)\sqrt{mr}\xi$

$$\begin{aligned}
|f_W - f_{W'}| &= \left| \min_h \|x - Wh\|^2 - \min_h \|x - W'h\|^2 \right| \\
&= \left| \min_h \|x - Wh\|^2 + \max_h \left(-\|x - W'h\|^2 \right) \right| \\
&\leq \left| \max_h \left(\|x - Wh\|^2 - \|x - W'h\|^2 \right) \right| \\
&\leq \left| \max_h 2x^T Wh - 2x^T W'h \right| \\
&\quad + \left| \max_h \|Wh\|^2 - \|W'h\|^2 \right| \\
&= \left| \max_h \sum_{i=1}^{r} h_i \langle 2x, (W - W')e_i \rangle \right|
\end{aligned}$$

$$
\begin{aligned}
&+\left|\max_h \sum_{i,j}^{r} h_i h_j \langle (W+W')e_i, (W-W')e_j\rangle\right| \\
&\quad \text{(Using Hölder's inequality)} \\
&\le \left|\sum_{i=1}^{r} |\langle 2x, (W-W')e_i\rangle|\right| \\
&\quad + \left|\sum_{i,j}^{r} |\langle (W+W')e_i, (W-W')e_j\rangle|\right| \\
&\quad \text{(Using Cauchy-Schwarz inequality)} \\
&\le \left|\sum_{i=1}^{r} \|2x\|\|(W-W')e_i\|\right| \\
&\quad + \left|\sum_{i,j}^{r} \|(W+W')e_i\| \left\|(W-W')e_j\right\|\right| \\
&\le \left|\sum_{i=1}^{r} \|2x\| \left\|\frac{\xi}{2}\mathbf{1}\right\|\right| + \left|\sum_{i,j}^{r} \|(W+W')e_i\| \left\|\frac{\xi}{2}\mathbf{1}\right\|\right| \\
&\le \sqrt{m}Rr\xi + \sqrt{m}r^2\xi \\
&= (R+r)\sqrt{m}r\xi = \xi'.
\end{aligned}
$$

The sixth inequality holds because of the triangle inequality. We have

$$
\sum_{i,j}^{r} \|(W+W')e_i\| \le \sum_{i,j}^{r} (\|We_i\| + \|W'e_i\|) \le \sum_{i,j}^{r} 2 = 2r^2.
$$

Let the metric d be the absolute difference metric. According to definition of covering number, for $\forall f_W \in F_{\mathcal{W}}$, there is a $W' \in \mathcal{S}$, such that

$$
\|d(f_W(X), f_{W'}(X))\|_1 = \left[\sum_{i=1}^{n} d(f_W(x_i), f_{W'}(x_i))\right] \le n\xi'.
$$

Thus

$$
\mathcal{N}_1(F_{\mathcal{W}}, \xi', n) \le |\mathcal{S}| \le \left(\frac{4}{\xi}\right)^{mr} = \left(\frac{4(R+r)\sqrt{mr}}{\xi'}\right)^{mr}.
$$

Taking log on both the sides, we have

$$
\ln \mathcal{N}_1(F_{\mathcal{W}}, \xi', n) \le mr \ln\left(\frac{4(R+r)\sqrt{mr}}{\xi'}\right).
$$

■

By Theorem 3, we have:

Lemma 3: Assume that $F_{\mathcal{W}}$ has the range $[0, b]$. For any $W \in \mathcal{W}$, any $\xi > 0$, and $n \ge \frac{8b^2}{\xi^2}$, we have

$$
\begin{aligned}
&P\{|R(W) - R_n(W)| \ge \xi\} \\
&\qquad \le 8\left(\frac{32(R+r)\sqrt{mr}}{\xi}\right)^{mr} \exp\left(-\frac{n\xi^2}{32b^2}\right)
\end{aligned}
$$

and for any $\epsilon > 0$ and $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$
|R(W) - R_n(W)| \le 2\epsilon + b\sqrt{\frac{mr \ln\left(\frac{4(R+r)\sqrt{mr}}{\epsilon}\right) - \ln\frac{\delta}{2}}{2n}}.
$$

Let $\epsilon = \frac{1}{n}$. For any $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$
|R(W) - R_n(W)| \le \frac{2}{n} + b\sqrt{\frac{mr \ln\left(4(R+r)\sqrt{mr}n\right) - \ln\frac{\delta}{2}}{2n}}.
$$

Proof: The first part can be easily obtained according to Theorem 3.

To prove the second part, let $F_{\mathcal{W},\epsilon}$ be a minimal ϵ-cover of $F_{\mathcal{W}}$. It can be easily verified that

$$
\sup_{f_W \in F_{\mathcal{W}}} |R(W) - R_n(W)| \le 2\epsilon + \sup_{f_W \in F_{\mathcal{W},\epsilon}} |R(W) - R_n(W)|. \tag{7}
$$

Using Hoeffding's inequality and the union bound property, we have

$$
P\left\{\sup_{f_W \in F_{\mathcal{W},\epsilon}} |R(W) - R_n(W)| \ge \xi\right\} \le 2|F_{\mathcal{W},\epsilon}| \exp\left(-\frac{2n\xi^2}{b^2}\right).
$$

We have proved in Lemma 2 that $|F_{\mathcal{W},\epsilon}| = \mathcal{N}_1(F_{\mathcal{W}}, \epsilon, n) \le (4(R+r)\sqrt{mr}/\epsilon)^{mr}$. Thus

$$
\begin{aligned}
&P\left\{\sup_{f_W \in F_{\mathcal{W},\epsilon}} |R(W) - R_n(W)| \ge \xi\right\} \\
&\qquad \le 2\left(\frac{4(R+r)\sqrt{mr}}{\epsilon}\right)^{mr} \exp\left(-\frac{2n\xi^2}{b^2}\right).
\end{aligned}
$$

Let

$$
2\left(\frac{4(R+r)\sqrt{mr}}{\epsilon}\right)^{mr} \exp\left(-\frac{2n\xi^2}{b^2}\right) = \delta.
$$

We get

$$
\xi = b\sqrt{\frac{mr \ln\left(\frac{4(R+r)\sqrt{mr}}{\epsilon}\right) - \ln\left(\frac{\delta}{2}\right)}{2n}}.
$$

Thus, with probability at least $1-\delta$, we have

$$
\begin{aligned}
&\sup_{f_W \in F_{\mathcal{W},\epsilon}} |R(W) - R_n(W)| \\
&\qquad \le \xi = b\sqrt{\frac{mk \ln\left(\frac{4(R+ck)\sqrt{mck}}{\epsilon}\right) - \ln\left(\frac{\delta}{2}\right)}{2n}}.
\end{aligned}
$$

Using inequality (7), we therefore have

$$
\begin{aligned}
&\sup_{f_W \in F_{\mathcal{W}}} |R(W) - R_n(W)| \\
&\quad \le 2\epsilon + \sup_{f_W \in F_{\mathcal{W},\epsilon}} |R(W) - R_n(W)| \\
&\quad \le 2\epsilon + b\sqrt{\frac{mr \ln\left(\frac{4(R+r)\sqrt{mr}}{\epsilon}\right) - \ln\left(\frac{\delta}{2}\right)}{2n}}.
\end{aligned}
$$

This has concluded the second part of the proof.

To prove the third part, let $\epsilon = \frac{1}{n}$. We have, for any $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$
\begin{aligned}
&\sup_{f_W \in F_{\mathcal{W}}} |R(W) - R_n(W)| \\
&\qquad \le \frac{2}{n} + b\sqrt{\frac{mr \ln(4(R+r)\sqrt{mr}n) - \ln\left(\frac{\delta}{2}\right)}{2n}}.
\end{aligned}
$$

This completes the proof of Lemma 3. ■

Because the bases W are normalized, we have that $b \le R$. Theorem 5 therefore follows form Lemma 3 by setting $R = 1$.

B. Proof of Theorem 5

The proof method is the same as that in [36]. We provide the detailed proof here for completeness.

Proof of Theorem 5: For LCNMF problems, the representative of an observation x is learned by optimizing the following function:

$$g(h) = \left\| x - \sum_{i=1}^{k} W_i h_i \right\|^2.$$

Assume that h is a minimizer of g and $\|h\| > R$.

Because W is normalized, $\|W_i\| = 1, i, \ldots, k$. Then

$$\begin{aligned} \left\| \sum_{i=1}^{k} W_i h_i \right\|^2 &= \|h\|^2 + \sum_{k \neq l} h_k h_l \langle W_k, W_l \rangle \\ &\geq \|h\|^2 > R^2. \end{aligned}$$

Let the real-valued function f be defined as $f(t) = g(th)$. Then

$$\begin{aligned} f'(1) &= 2\left(\left\| \sum_{i=1}^{k} W_i h_i \right\|^2 - \left\langle x, \sum_{i=1}^{k} W_i h_i \right\rangle \right) \\ &\quad \text{(Using Cauchy-Schwarz inequality)} \\ &\geq 2\left(\left\| \sum_{i=1}^{k} W_i h_i \right\|^2 - R \left\| \sum_{i=1}^{k} W_i h_i \right\| \right) \\ &= 2\left(\left\| \sum_{i=1}^{k} W_i h_i \right\| - R \right) \left\| \sum_{i=1}^{k} W_i h_i \right\| \\ &> 0. \end{aligned}$$

Therefore, f cannot have a minimum at 1, whence h cannot be a minimizer of g. Thus, the minimizer h^* must be contained in the ball with radius R in m-dimensional space. ■

C. Proof of Theorem 7

In [37, Proposition 3.2], the authors proved that the included Rademacher complexity for NMF is upper bounded by

$$\frac{\sqrt{2\pi}}{n}\left(\sqrt{8}E \sup_{W \in \mathbb{R}_+^{m \times r}} \sum_{i=1}^{n} \sum_{k=1}^{r} \gamma_{ik} \langle x_i, We_k \rangle + \sqrt{2}E \sup_{W \in \mathbb{R}_+^{m \times r}} \sum_{i=1}^{n} \sum_{l,k=1} \gamma_{ilk} \langle We_l, We_k \rangle \right)$$

where γ_{ik} and γ_{ilk} are orthogonal Gaussian sequences that are of independent Gaussian random variables with zero mean and unit standard deviation.

For orthogonal NMF, we have

$$\begin{aligned} \Re(F) &\leq \frac{\sqrt{2\pi}}{n}\left(\sqrt{8}E \sup_{W \in \mathbb{R}_+^{m \times r}} \sum_{i=1}^{n} \sum_{k=1}^{r} \gamma_{ik} \langle x_i, We_k \rangle + \sqrt{2}E \sup_{W \in \mathbb{R}_+^{m \times r}} \sum_{i=1}^{n} \sum_{l,k=1} \gamma_{ilk} \langle We_l, We_k \rangle \right) \\ &\leq \frac{\sqrt{2\pi}}{n}\left(\sqrt{8}E \sup_{W \in \mathbb{R}_+^{m \times r}} \sum_{i=1}^{n} \sum_{k=1}^{r} \gamma_{ik} \langle x_i, We_k \rangle + \sqrt{2}E \sup_{W \in \mathbb{R}_+^{m \times r}} \sum_{i=1}^{n} \sum_{l,k=1} \gamma_{ilk} \|We_k\|^2 \right) \\ &\leq \frac{\sqrt{2\pi}}{n}(r\sqrt{8n} + r\sqrt{2n}) = 6r\sqrt{\frac{\pi}{n}} \end{aligned}$$

where the second inequality holds, because the columns of W are orthogonal, and the third inequality holds because of [37, Lemma 3.3]. Theorem 7 can be proved according to Theorem 1. ■

REFERENCES

[1] S. Arora, R. Ge, R. Kannan, and A. Moitra, "Computing a nonnegative matrix factorization—Provably," in *Proc. 44th Annu. ACM Symp. Theory Comput.*, 2012, pp. 145–162.

[2] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Mar. 2003.

[3] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced Lectures on Machine Learning*. Heidelberg, Germany: Springer, 2004, pp. 169–207.

[4] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Trans. Neural Netw.*, vol. 19, no. 6, pp. 1090–1100, Jun. 2008.

[5] D. Cai, X. He, X. Wang, H. Bao, and J. Han, "Locality preserving nonnegative matrix factorization," in *Proc. IJCAI*, 2009, pp. 1010–1015.

[6] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proc. SIAM SDM*, 2005, pp. 606–610.

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

[8] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. SIGKDD*, 2006, pp. 126–135.

[9] P. Domingos, "A unified bias-variance decomposition," in *Proc. ICML*, 2000, pp. 231–238.

[10] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*, Cambridge, MA, USA: MIT Press, 2004.

[11] S. S. Dragomir and B. Mond, "On a property of Gram's determinant," *Extracta Math.*, vol. 11, no. 2, pp. 282–287, 1996.

[12] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," in *Proc. Amer. Control Conf.*, 2003, pp. 2156–2162.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[14] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[15] N. Gillis, "Sparse and unique nonnegative matrix factorization through data preprocessing," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 3349–3386, Nov. 2012.

[16] K. Gold, C. Havasi, M. Anderson, and K. Arnold, "Comparing matrix decomposition methods for meta-analysis and reconstruction of cognitive neuroscience results," in *Proc. AAAI*, 2011, pp. 21–26.

[17] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3469–3486, Jun. 2015.

[18] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor. (2012). "MahNMF: Manhattan non-negative matrix factorization." [Online]. Available: https://arxiv.org/abs/1207.3438

[19] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: An optimal gradient method for nonnegative matrix factorization," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2882–2898, Jun. 2012.

[20] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2551724.

[21] H. Xiong, T. Liu, D. Tao, and H. Shen, "Dual diversified dynamical Gaussian process latent variable model for video repair," *IEEE Trans. Image Process.*, to be published.

[22] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004.

[23] M. Kaykobad, "On nonnegative factorization of matrices," *Linear Algebra Appl.*, vol. 96, pp. 27–33, Nov. 1987.

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

[25] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using L21-norm," in *Proc. CIKM*, 2011, pp. 673–682.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[27] 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.

[28] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001, pp. 1–7.

[29] L.-X. Li, L. Wu, H.-S. Zhang, and F.-X. Wu, "A fast algorithm for nonnegative matrix factorization and its convergence," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1855–1863, Oct. 2014.

[30] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning spatially localized, parts-based representation," in *Proc. CVPR*, 2001, pp. I-207–I-212.

[31] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[32] C. J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, Nov. 2007.

[33] J. Liu, Z. Li, Z.-H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1233–1246, Jun. 2015.

[34] T. Liu and D. Tao, "On the performance of manhattan nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2015.2458986.

[35] T. Liu, D. Tao, M. Song, and S. J. Maybank, "Algorithm-dependent generalization bounds for multi-task learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2016.2544314.

[36] T. Liu, D. Tao, and D. Xu, "Dimensionality-dependent generalization bounds for k-dimensional coding schemes," *Neural Comput.*, to be published.

[37] A. Maurer and M. Pontil, "K-dimensional coding schemes in Hilbert spaces," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5839–5846, Nov. 2010.

[38] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 765–777, Mar. 2007.

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

[40] A. Papaioannou and S. Zafeiriou, "Principal component analysis with complex kernel: The widely linear model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1719–1726, Sep. 2014.

[41] D. Pollard, *Convergence of Stochastic Processes*. New York, NY, USA: Springer-Verlag, 1984.

[42] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur, "Two algorithms for orthogonal nonnegative matrix factorization with application to clustering," *Neurocomputing*, vol. 141, pp. 15–25, Oct. 2014.

[43] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Proc. NIPS*, 2012, pp. 1214–1222.

[44] R. Sandler and M. Lindenbaum, "Nonnegative matrix factorization with Earth mover's distance metric for image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1590–1602, Aug. 2011.

[45] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 46–51.

[46] K. Takeuchi, K. Ishiguro, A. Kimura, and H. Sawada, "Non-negative multiple matrix factorization," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1713–1720.

[47] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[48] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[49] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[50] F. J. Theis, K. Stadlthanner, and T. Tanaka, "First results on uniqueness of sparse non-negative matrix factorization," in *Proc. EUSIPCO*, Sep. 2005, pp. 1–4.

[51] L. B. Thomas, "Rank factorization of nonnegative matrices," *SIAM Rev.*, vol. 16, no. 3, pp. 393–394, 1974.

[52] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," *J. Mach. Learn. Res.*, vol. 12, pp. 3259–3281, Feb. 2011.

[53] V. Vapnik, *The Nature of Statistical Learning Theory*. Heidelberg, Germany: Springer, 2000.

[54] F.-Y. Wang, C.-Y. Chi, T.-H. Chan, and Y. Wang, "Nonnegative least-correlated component analysis for separation of dependent sources by volume maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 875–888, May 2010.

[55] R. Wang and D. Tao, "Non-local auto-encoder with collaborative stabilization for image restoration," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2117–2129, May 2016.

[56] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

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

[58] M. Ye, Y. Qian, and J. Zhou, "Multitask sparse nonnegative matrix factorization for joint spectral–spatial hyperspectral imagery denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2621–2639, May 2015.

[59] J. Yoo and S. Choi, "Nonnegative matrix factorization with orthogonality constraints," *J. Comput. Sci. Eng.*, vol. 4, no. 2, pp. 97–109, 2010.

[60] T. Zhang, "Covering number bounds of certain regularized linear function classes," *J. Mach. Learn. Res.*, vol. 2, pp. 527–550, Mar. 2002.

[61] W.-S. Zheng, S. Z. Li, J. H. Lai, and S. Liao, "On constrained sparse matrix factorization," in *Proc. IEEE ICCV*, Oct. 2007, pp. 1–8.

[62] D. Zhou *et al.*, "Learning multiple graphs for document recommendations," in *Proc. ACM WWW*, 2008, pp. 141–150.

[63] G. Zhou, S. Xie, Z. Yang, J.-M. Yang, and Z. He, "Minimum-volume-constrained nonnegative matrix factorization: Enhanced ability of learning parts," *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1626–1637, Oct. 2011.

[64] G. Zhou, Z. Yang, S. Xie, and J.-M. Yang, "Online blind source separation using incremental nonnegative matrix factorization with volume constraint," *IEEE Trans. Neural Netw.*, vol. 22, no. 4, pp. 550–560, Apr. 2011.

Tongliang Liu received the B.E. degree in electronics engineering and information science from the University of Science and Technology of China, Hefei, China, in 2012. He is currently pursuing the Ph.D. degree in computer science with the University of Technology Sydney, Ultimo, NSW, Australia.

His current research interests include statistical learning theory, causal reference, computer vision, and optimization.

Mr. Liu received the best paper award in the IEEE International Conference on Information Science and Technology 2014.

Mingming Gong received the B.S. degree in electrical engineering from Nanjing University, Nanjing, China, and the M.S. degree in communications and information systems from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the School of Software, University of Technology Sydney, Ultimo, NSW, Australia.

His current research interests include causal inference, domain adaptation, kernel methods, and deep learning.

Dacheng Tao (F'15) is Professor of Computer Science with the Centre for Quantum Computation & Intelligent Systems, and the Faculty of Engineering and Information Technology in the University of Technology Sydney. He mainly applies statistics and mathematics to data analytics problems and his research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 200+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM'07, the best student paper award in IEEE ICDM'13, and the 2014 ICDM 10-year highest-impact award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research. He is a Fellow of the IEEE, OSA, IAPR and SPIE.