On the Performance of Manhattan Non-negative Matrix Factorization

Tongliang Liu and Dacheng Tao, *Fellow, IEEE*

***Abstract*—Extracting low-rank and sparse structures from matrices has been extensively studied in machine learning, compressed sensing, and conventional signal processing, and has been widely applied to recommendation systems, image reconstruction, visual analytics, and brain signal processing. Manhattan non-negative matrix factorization (MahNMF) is an extension of the conventional non-negative matrix factorization (NMF), which models the heavy-tailed Laplacian noise by minimizing the Manhattan distance between a non-negative matrix X and the product of two non-negative low-rank factor matrices. Fast algorithms have been developed to restore the low-rank and sparse structures of X in MahNMF. In this paper, we study the statistical performance of MahNMF in the frame of the statistical learning theory. We decompose the expected reconstruction error of MahNMF into the estimation error and the approximation error. The estimation error is bounded by the generalization error bounds of MahNMF, while the approximation error is analyzed using the asymptotic results of the minimum distortion of vector quantization. The generalization error bound is valuable for determining the size of the training sample needed to guarantee a desirable upper bound for the defect between the expected and empirical reconstruction errors. Statistical performance analysis shows how the reduced dimensionality affects the estimation and approximation errors. Our framework can also be used for analyzing the performance of NMF.**

***Index Terms*—Manhattan distance, Non-negative matrix factorization, statistical analysis, estimation error, approximation error.**

I. INTRODUCTION

THE sheer volume of high-dimensional data generated from a wide variety of applications, such as data management [1], [2] and visual analytics [3], [4], present both a challenge and an opportunity for machine learning and algorithm development. Arguably, the high-dimensional data have low intrinsic dimensionality, particularly when sampled from a low-dimensional manifold, as shown by the success of manifold learning in many applications [5]–[7]. By stacking all data as column vectors of a matrix X, the assumed low intrinsic dimensionality means that X is constructed by perturbing a low-rank matrix L with small noise N, i.e., $X = L + N$. Low-rank structure has been extensively studied, for instance in principal component analysis (PCA) [8], sparse wavelets [9] and non-negative matrix factorization (NMF) [10], to either exactly or approximately represent an arbitrary example in a dataset as a weighted sum of a few bases W, i.e., $L = WH$.

Since examples in a dataset can be associated with specific structures, e.g., face images under different lighting conditions, it is insufficient to assume that L is perturbed by small noise N. This means that we need to model the random sparse structure S, i.e., $X = L + S + N$. If we simply ignore S, the estimated rank of L will be significantly increased; for this reason, low-rank and sparse matrix decomposition has been extensively studied. Although the decomposition model itself is simple, it is more difficult to exactly or approximately recover L and S, since the rank and cardinality constraints are not convex. Exact and unique decomposition does not always exist for an arbitrary matrix, and approximate decomposition results in a trade-off between low-rank and sparsity. Successful algorithms include rank-sparsity incoherence for matrix decomposition [11], robust PCA [12], GoDec [13], and Manhattan NMF (MahNMF) [14].

NMF, popularized by Lee and Seung [10], allows for the factorization of a non-negative data matrix into two low-rank non-negative matrices. Unlike other traditional dimensionality reduction methods, the found bases and the newly represented data matrices are required to be non-negative in NMF. The non-negativity allows only additive combinations and forces learning of parts-based representations [15]–[18]. NMF has been successfully applied as a dimensionality reduction approach in many fields, such as signal processing [19]–[21], data mining [22]–[24] and bioinformatics [25]–[27].

Numerous different algorithms have been developed for NMF applications, which can be classified into two groups depending on whether additional content information is used: the first group contains *content-free algorithms* [28], which attempt to find an optimization solution for NMF via numerical methods; the second contains *content-based algorithms* [29], in which the non-negative matrix is factorized with a constraint on W (or H, or both). Due to the successful use of NMF in real-world applications, several popular computing environments, such as Matlab, R and Oracle Data Mining, have developed NMF packages.

MahNMF is an important extension of NMF that learns the low-rank and sparse structures simultaneously. In particular, MahNMF models the heavy-tailed Laplacian noise by minimizing the Manhattan distance between $X \in \mathbb{R}_+^{m\times n}$ and its low-rank approximation $L = WH$, where $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{r\times n}$; it is therefore effective when data are contaminated by outliers. It has been comprehensively and

Manuscript received December 12, 2014; revised July 17, 2015; accepted July 18, 2015. This work was supported in part by Australian Research Council Projects DP-140102164 and FT-130101457.

T. Liu and D. Tao are with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia (Email: tliang.liu@gmail.com; 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.xxxx

empirically proven that MahNMF can robustly estimate the low-rank and sparse components of a non-negative matrix [14]. We also provide a thorough analysis to show how MahNMF is more robust to noise and outliers than NMF from an optimization viewpoint. Recently, learning with noisy examples has received a great deal of attention. The interested reader is referred to further examples [30]–[32].

Donoho and Stodden [33] discussed under what conditions NMF is unique (a sufficient condition for $X = WH$). The problem of when $X = WH$ has also been studied for a few special cases of NMF: Thomas [34] proved a sufficient and necessary condition for rank-r NMF, and Kaykobad [35] presented a sufficient condition for symmetric NMF (or a non-negative symmetric matirx to be completely positive). We therefore know that in most cases, the solutions of MahNMF (and indeed NMF) returned by content-free algorithms are not unique and $X \neq WH$, and that even though solutions returned by content-based algorithms may be unique, X is unlikely to be equal to WH. However, there are still two fundamental questions that have yet to be posed:

1) How much is $E\|X - W_nH_n\|$?
2) How much is $E\|X - W^*H^*\|$?

We use W_n and W^* to denote the learned and target bases of MahNMF, respectively, and H_n and H^* are the corresponding new representations. Maurer and Pontil [36] developed dimensionality-independent generalization error bounds for k-dimensional coding schemes, which can be used to analyze the dimensionality-independent generalization error bounds of NMF; this approach answers the first question for NMF. However, the MahNMF analysis is different, and therefore in this paper, we address these two questions specifically with respect to MahNMF.

Using statistic learning theory [37], [38], we first define the expected reconstruction error for MahNMF and decompose it into estimation and approximation errors. The estimation error, which is dependent on the used learning algorithm, is bounded by the generalization error bound of MahNMF. Using the Rademacher complexity [39] and covering number methods [40], we derive the generalization error bounds. The obtained bounds show, with high probability, that the expected reconstruction error of the learned bases W_n is not more than $\mathcal{O}(\sqrt{mr\ln n/n})$ worse than the empirical reconstruction error of the learned bases W_n; this answers the first question. The approximation error, which is dependent on the sample distribution, is analyzed by the asymptotic results of the minimum distortion of vector quantization [41]. In contrast to the estimation error, the approximation error is a non-increasing function of the reduced dimensionality r, and decreases in the order of $\mathcal{O}(r^{-1/m})$ as r approaches m. This answers the second question.

We note that our method based on directly bounding the covering number of the induced loss class can be used to derive tighter dimensionality-dependent generalization error bounds (than the dimensionality-independent bounds in [36]) for NMF, and that the method for deriving the approximation error bound can also be applied to the NMF using Euclidean distance loss.

The rest of this paper is organized as follows. In Section II, we present the setup of the problem, and provide the stochastic framework by decomposing the expected reconstruction error into estimation error and approximation error. Our main results about estimation error and approximation error are outlined in Section III and IV, respectively. The proofs of our main results are provided in Section V. Finally, Section VI concludes the paper.

II. Problem Setup

This section introduces notation that will be used throughout this paper, and then presents the normalized MahNMF. We also define reconstruction error for MahNMF and decompose the expected reconstruction error into estimation error and approximation error.

A. Notation

We denote $X = (x_1, \ldots, x_n) \in \mathbb{R}_+^{m\times n}$ as the data matrix consisting of n independent and identically distributed observations drawn from a space $\mathcal{X}$ with a Borel measure ρ. We use capital letters to denote matrices, A_i the i-th column of matrix A and A_{ij} the ij-th entry of A. The ℓ_2 (Euclidean) norm is denoted by $\|\cdot\|_2$, while the Manhattan distance between A and B is denoted by $\|A-B\|_M$ or $\|A-B\|_1$. The notation $\mathcal{O}(\cdot)$ represents the big Omicron notation in Bachmann-Landau notation.

B. MahNMF

NMF minimizes the Euclidean distance loss and KLNMF[1] minimizes the Kullback-Leibler (KL) divergence loss as follows:

$$\begin{aligned} \min\nolimits_{W,H} \quad & F_{\text{NMF}}(WH) = \|X - WH\|_2^2 \\ & = \sum_{i=1}^{n} \|x_i - WH_i\|_2^2, \\ s.t. \quad & W \in \mathbb{R}_+^{m\times r}, H \in \mathbb{R}_+^{r\times n} \end{aligned} \tag{1}$$

and

$$\begin{aligned} \min\nolimits_{W,H} \quad & F_{\text{KLNMF}}(WH) = D(X\|WH), \\ s.t. \quad & W \in \mathbb{R}_+^{m\times r}, H \in \mathbb{R}_+^{r\times n}, \end{aligned} \tag{2}$$

where

$$D(A\|B) = \sum_{ij}\left(A_{ij}\ln\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}\right)$$

and the reduced dimensionality r satisfies that $r < \min(m, n)$.

The Euclidean distance and KL divergence loss model the Gaussian noise and Poisson distribution, respectively. Assume that the residuals (or noise) $(x_i - WH_i), i = 1, \ldots, n$ are independently distributed with a Gaussian distribution. Problem (1) can be easily derived using the maximum likelihood approach.

[1]For clarity, we denote NMF and KLNMF as the matrix factorization procedures minimizing Euclidean distance loss and KL divergence loss, repectively.

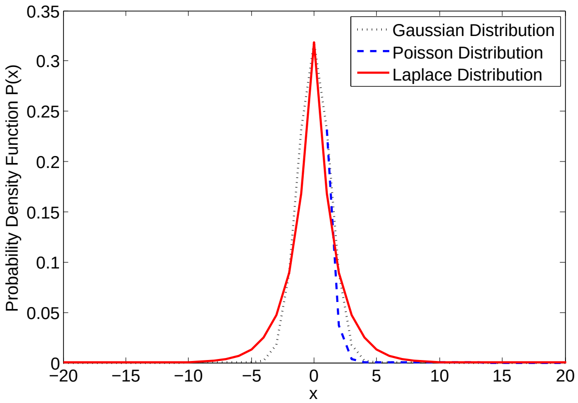

Fig. 1. Gaussian, Poisson and Laplace distributions. For a fair comparison, we set the parameters of the three distributions to be $\sigma = \sqrt{\pi/2}$, $\lambda = 1/\pi$ and $b = \pi/2$, respectively, such that $P_{\text{Laplace}}(0) = P_{\text{Gaussian}}(0)$ and $P_{\text{Gaussian}}(1) = P_{\text{Poisson}}(1)$. The figure shows that the Laplace distribution has a heavier tail than those of the other two.

To derive problem (2), we assume that $x_i \sim \text{Poisson}(WH_i)$, that is

$$P(x_i|WH_i) = \frac{e^{-WH_i}(WH_i)^{x_i}}{x_i!}.$$

Since we have assume that $x_i, i = 1, \ldots, n$ are independently distributed, the Poisson likelihood of observing $x_i, i = 1, \ldots, n$ given underlying parameters WH is given by

$$P(X|WH) = \prod_{i=1}^{n} \frac{e^{-WH_i}(WH_i)^{x_i}}{x_i!}. \quad (3)$$

To maximize the likelihood with respect to WH, take the log on both sides of Eq. (3), we have the following problem:

$$\begin{aligned} \min_{W,H} \quad & \sum_{i=1}^{n} WH_i - x_i \log(WH_i), \\ s.t. \quad & W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}, \end{aligned}$$

which is equal to the problem (2).

However, Gaussian and Poisson distributions are not heavy-tailed, as shown in Fig. 1. Considering that many practical contaminations, e.g., occlusions, are heavy-tailed, NMF and KLNMF therefore do not perform well in such practical scenarios.

Assume that the residuals $(x_i - WH_i), i = 1, \ldots, n$ are independently sampled from a Laplace distribution, which is heavy-tailed as shown in Fig. 1. The objective function of MahNMF can be easily derived using the maximum likelihood approach. Thus, MahNMF employs the Manhattan distance between X and WH to model the heavy-tailed Laplacian noise:

$$\begin{aligned} \min_{W,H} \quad & F_{\text{MahNMF}}(WH) = \|X - WH\|_M \\ & = \sum_{i=1}^{n} \|x_i - WH_i\|_1, \\ s.t. \quad & W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}, \end{aligned}$$

where the reduced dimensionality $r < \min(m, n)$.

Moreover, thanks to the Manhattan distance loss, MahNMF also effectively extracts the sparse structures, which makes MahNMF much more robust to outliers and capable to occlusions [14].

Besides the robustness advantage of MahNMF discussed from the viewpoint of noise distribution, we show the advantage of MahNMF from an optimization procedure viewpoint.

Let $F(WH)$ be the objective function of a non-negative matrix factorization problem. Let $f(t) = F(tWH), t \in \mathbb{R}$. We can verify that optimizing $F(WH)$ is equal to finding a WH such that $\delta f(1) = 0$, where $\delta f(t)$ denotes the subgradient of $f(t)$.

We have

$$\delta f_{\text{NMF}}(1) = \sum_{ij} 2(X - WH)_{ij}(-WH)_{ij} \quad (4)$$

and

$$\delta f_{\text{KLNMF}}(1) = \sum_{ij} (WH - X)_{ij} \quad (5)$$

and

$$\begin{aligned} & \delta f_{\text{MahNMF}}(1) \\ & = \sum_{ij} \frac{2}{|X - WH|_{ij}} (X - WH)_{ij}(-WH)_{ij}. \end{aligned} \quad (6)$$

We define $\frac{0}{0} \in [-1, 1]$ for MahNMF to satisfy the definition of subgradient.

We compare the robustness of MahNMF and NMF from an optimization viewpoint. Let

$$c(X_{ij}, WH) = 2(X - WH)_{ij}(-WH)_{ij}$$

be the contribution the jth entry of the ith training example made to the optimization procedure,

$$w_{\text{MahNMF}}(X_{ij}, WH) = 1/|X - WH|_{ij}$$

the weight to the contribution for the jth entry of the ith training example of MahNMF and

$$w_{\text{NMF}}(X_{ij}, WH) = 1$$

the weight to the contribution for the jth entry of the ith training example of NMF. Comparing equations (4) and (6), we notice that every entry has the same contribution but different weights. We further find that $|X - WH|_{ij}$ represents the noise error added to the ijth entry of X or the very large error introduced by an outlier X_{ij} with regard to W. We therefore interpret the optimization procedures as contribution-weighted optimization procedures regarding to the noise error of entries.

During the optimization procedure, a robust algorithm should assign a small weight to the contribution made by a large noise error entry, and a large weight to the contribution made by a small noise error entry. According to the contribution-weighted optimization procedure interpretation, MahNMF assigns a smaller weight to a contribution if the corresponding entry has a larger error while NMF provides the same weight to all entries. Thus, MahNMF is much more robust than NMF.

To compare the robustness of MahNMF and KLNMF, we set the contribution as

$$c(X_{ij}, WH) = (WH - X)_{ij},$$

the weight of MahNMF as

$$w_{\text{MahNMF}}(X_{ij}, WH) = 2/\left|1 - \frac{(X)_{ij}}{(WH)_{ij}}\right|$$

and the weight of KLNMF as

$$w_{\text{KLNMF}}(X_{ij}, WH) = 1.$$

Since the KL divergence is asymmetric, we compare the robustness in two directions: $X_{ij} > (WH)_{ij}$ and $X_{ij} < (WH)_{ij}$. For each direction, it can be easily verified that when the noise error $|X - WH|_{ij}$ is large, the weight function $w_{\text{MahNMF}}(X_{ij}, WH)$ will be small while the weight function of KLNMF assigns the same weight to all entries. MahNMF is therefore much more robust than NMF. We empirically show the robustness and good performance of MahNMF in Appendix B.

For fixed bases W, the representations H are determined by a convex problem. Therefore, it is sufficient to analyze the performance of MahNMF by studying the choice of bases W.

Note that $WH = WQ^{-1}QH$ for any scaling matrix Q, and we can normalize $\|W_i\|_1, i = 1, \ldots, r$ by choosing

$$Q = \begin{pmatrix} \|W_1\|_1 & & & \\ & \|W_2\|_1 & & \\ & & \ddots & \\ & & & \|W_r\|_1 \end{pmatrix}$$

to limit the choice of bases W without changing the ability to represent the solutions to a MahNMF problem. Here, we call MahNMF with ℓ_1 normalized bases W the ℓ_1-normalized MahNMF, which has the following property:

Lemma 1: For ℓ_1-normalized MahNMF problems, if $\|x\|_1 \leq R$, then $\|h\|_1 \leq 2R$.

We defer the proof until Section V.

C. Extensions of MahNMF

In the previous subsection, we showed that MahNMF is more robust (to noise) than NMF and KLNMF. Then, it may interest many readers to exploit the extensibility of MahNMF for practical applications.

Guan et al. [14] provided a flexible framework for developing various algorithms of MahNMF for practical applications. They presented detailed optimization algorithms for MahNMF that are restricted by box-constraint [42], manifold regularization [43], group saprsity [44], elastic net [45] and symmetric bases [46]. MahNMF of cause can be extended to many other different scenarios; see, for example, online NMF learning [47], NMF that handles time varying data [48], and NMF that exploits the pairwise constraints that indicate the similarity/dissimilarity between two observations [49].

Note that the generalization error bounds of MahNMF derived by employing the complexity measures that measure the whole predefined loss class (such as the generalization error bounds in Theorems 1 and 2) can also be the upper bounds for the batch learning [50] extensions of MahNMF. Constraints will help produce a small active loss class, which is a subset of the predefined loss class. Then the complexity of the active loss class will be small. Taking the Rademacher complexity for an example, Bartlett and Mendelson have proven this property in Theorem 12 of [39]. So, if a constraint produces a small active loss class, the corresponding learning algorithm will share the generalization bound obtained by analyzing the complexity of the whole predefined loss class.

However, for different extensions of MahNMF, the estimation and approximation errors may vary. It is hard to analyze them in a uniform approach. We therefore, in the rest of the paper, analyze the performance of the empirical risk minimization (ERM) algorithm of MahNMF.

D. Estimation Error and Approximation Error

This subsection sets a framework for our results based on statistic learning theory. For any bases $W \in \mathbb{R}_+^{m \times r}$, we define the *reconstruction error* of an example x as follows:

$$f_W(x) = \min_{h \in \mathbb{R}_+^r} \|x - Wh\|_1.$$

The *expected reconstruction error* provides a criterion for choosing the bases W, such that the expected reconstruction error is minimized.

Definition 1 (Expected reconstruction error): If the measurable space $\mathcal{X}$ has a Borel measure ρ with distribution density function $p(x)$, we define the expected reconstruction error of the bases W as

$$R(W) = \int_{\mathcal{X}} f_W(x) d\rho(x) = \int_{\mathcal{X}} f_W(x) p(x) dx.$$

MahNMF attempts to search the bases W to make a small expected reconstruction error. However, the distribution density function $p(x)$ is usually unknown, and $R(W)$ cannot be directly minimized. The ERM algorithm provides a way to learn an approximation by minimizing the *empirical reconstruction error*.

Definition 2 (Empirical reconstruction error): If data $X = (x_1, \cdots, x_n)$ are drawn independently and identically from $\mathcal{X}$, we define the empirical reconstruction error of the bases W as

$$R_n(W) = \frac{1}{n} \sum_{i=1}^{n} f_W(x_i).$$

Following statistic learning theory, we define the set of functions $F_{\mathcal{W}_r}$ indexed by $\mathcal{W}_r = \mathbb{R}_+^{m \times r}$ as the *loss class* of MahNMF. For every $f_W \in F_{\mathcal{W}_r}$, $f_W(x)$ measures the reconstruction error of x using the bases $W \in \mathcal{W}_r$. Thus, choosing the bases W with a small empirical reconstruction error is equal to choosing a $f_W \in F_{\mathcal{W}_r}$ such that the empirical reconstruction error is minimized.

Our goal is to analyze the expected reconstruction error of the learned bases $W_{n,r}$ of MahNMF, where

$$W_{n,r} = \arg\min_{W \in \mathcal{W}_r} R_n(W).$$

We can decompose it into the *estimation error* and *approximation error* as follows

$$R(W_{n,r}) = \underbrace{R(W_{n,r}) - R(W_r)}_{\text{Estimation error}} + \underbrace{R(W_r),}_{\text{Approximation error}}$$

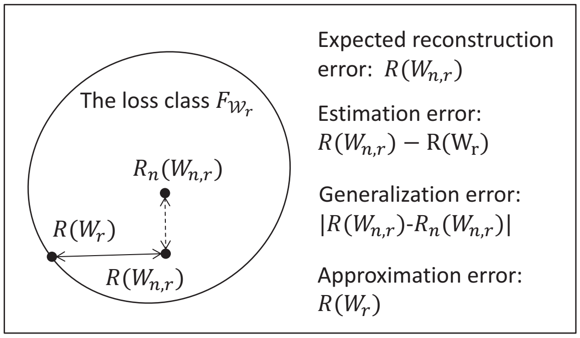

Fig. 2. A schematic diagram of the loss class and errors. The circle represents the loss class of MahNMF. W_r are the target bases in $\mathcal{W}_r$, and $R(W_r)$ is the corresponding expected reconstruction error. $W_{n,r}$ are the bases in $\mathcal{W}_r$ hypothesized by the learner on the basis of the training data, and $R(W_{n,r})$ and $R_n(W_{n,r})$ are the expected and empirical reconstruction errors, respectively. The relationship between the expected reconstruction error, the estimation error, the generalization error, and the approximation error are shown by points and arrows.

where

$$W_r = \arg \min_{W \in \mathcal{W}_r} R(W).$$

In this paper, the estimation error and approximation error are bounded, respectively. The following inequalities provide a method to upper bound the estimation error.

$$\begin{aligned} & R(W_{n,r}) - R(W_r) \\ & \leq |R(W_{n,r}) - R_n(W_{n,r})| \\ & \quad + R_n(W_{n,r}) - R_n(W_r) + |R_n(W_r) - R(W_r)| \\ & \leq |R(W_{n,r}) - R_n(W_{n,r})| + |R_n(W_r) - R(W_r)| \\ & \leq 2 \sup_{W \in \mathbb{R}_+^{m \times r}} |R(W) - R_n(W)|. \end{aligned} \tag{7}$$

The second inequality holds because $W_{n,r}$ is defined to be the global solution as $W_{n,r} = \arg\min_W R_n(W)$. The defect in the last line is referred to as the *generalization error*.

The expected reconstruction error $R(W_{n,r})$ can be either bounded by bounding the estimation error $R(W_{n,r}) - R(W_r)$ and approximation error $R(W_r)$, or bounding the generalization error $R(W_{n,r}) - R_n(W_{n,r})$.

Figure 2 is a schematic diagram illustrating the loss class and errors. It is well known that in the classical statistical learning theory, the estimation error mostly depends on learning algorithms and the approximation error mostly depends on the choice of the loss class. However, for a MahNMF (and NMF) problem, the loss class is fixed because $\mathcal{W}_r = \mathbb{R}_+^{m\times r}$ and the loss function $f_{W_r} = \arg\min_{f_W \in F_{\mathcal{W}_r}} R(W)$ is always in the loss class[2]. Therefore, the approximation error for MahNMF (and NMF) depends only on the distribution of $\mathcal{X}$ (if r is fixed). We will study the estimation error, generalization error and approximation error of MahNMF in the rest of the paper.

[2]If the minimum cannot be attained, we set $f_{W_r} = \arg\inf_{f_W \in F_{\mathcal{W}_r}} R(W)$, which means there is a sequence $\{f_{W_i}, i = 1, 2, \ldots\} \subseteq F_{\mathcal{W}_r}$ such that when $i \to \infty$, $f_{W_r} = f_{W_i} \to \arg\inf_{W \in \mathcal{W}_r} R(W)$. Throughout this paper, we always assume the minimum exists.

III. Generalization and Estimation Error Bounds

We bound the estimation error using the bound of the generalization error and inequalities (7). The generalization error is usually bounded by measuring the complexity of the loss class. The method based on Rademacher complexity is one of the most frequently used methods for deriving generalization error bounds, based on which the following theorem is obtained.

Theorem 1: For ℓ_1-normalized MahNMF problems, assume that $\|x\|_1 \leq R$. For any $W \in \mathcal{W}_r$ and any $\delta > 0$, with probability at least $1 - \delta$, we have

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

The Rademacher complexity usually leads to dimensionality-independent generalization bounds. However, our obtained bound (in Theorem 1) is dimensionality-dependent. One reason for this is that in the proof we used $\|x\|_1 \leq \sqrt{m}\|x\|_2$ to find a proper Gaussian process to upper bound the Rademacher complexity. We conjecture that the Rademacher complexity can be used to derive a dimensionality-independent generalization error bound for MahNMF by exploiting a more subtle proving approach.

The method of Rademacher complexity will introduce the worst-case dependency with regard to r, which means the generalization bound in Theorem 1 is loose. Combining the Rademacher complexity and covering number to measure the complexity of the loss class may avoid the worse-case dependency (see, e.g., the proof method of Theorem 2 in [36]). The price to pay for this is that the obtained convergence rate is slower than that of Theorem 1.

To combine the measurements of Rademacher complexity and covering number, Maurer and Pontil [36] factored the bases W as $W = US$ for NMF, where S is an $r \times r$ matrix and U is an $m \times r$ isometry[3], and measured the complexities induced by the $r \times r$ matrices and isometries by exploiting the covering number and Rademacher complexity, respectively. However, for MahNMF problems, we cannot derive any tighter generalization bounds by using this method.

We introduce a method to directly bound the covering number of the loss function class induced by the reconstruction error. The obtained generalization bound will be much tighter.

Theorem 2: For ℓ_1-normalized MahNMF problems, assume that $\|x\|_1 \leq R$. For any $W \in \mathcal{W}_r$, any $\xi > 0$ and any $\delta > 0$, and $n \geq \frac{8R^2}{\xi^2}$, we have

$$\begin{aligned} & P\left\{|R(W) - R_n(W)| \geq \xi\right\} \\ & \leq 8\left(\frac{16mR}{\xi}\right)^{mr} \exp\left(-\frac{n\xi^2}{32R^2}\right), \end{aligned}$$

[3]Let U be a bounded linear transformation from $\mathbb{R}^r$ to $\mathbb{R}^m$. If U is an isometry, it holds that $\|Ux\| = \|x\|$ for all $x \in \mathbb{R}^r$, where $\|\cdot\|$ denotes a norm on the Euclidian space. It can be easily verified that the ℓ_1-normalized bases W compose an isometry with regard to the ℓ_1 norm.

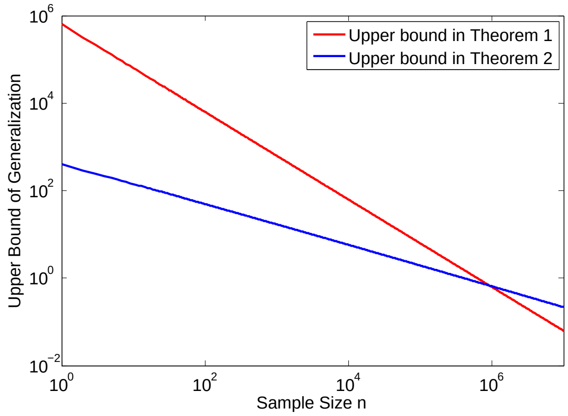

Fig. 3. Comparison of the upper generalization bounds for MahNMF. We set $\delta = 0.01$, $R = 1$, $m = 1000$, $r = 40$ and n within the range $[1, 10^7]$.

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

$$\begin{aligned}|R(W) - R_n(W)| \quad &\le R\sqrt{\frac{mr\ln(2mnR)}{2n}} \\ &+\frac{2}{n} + R\sqrt{\frac{\ln(2/\delta)}{2n}}.\end{aligned}$$

See the proof in Section V.

In Theorem 2, the dependence of the upper bound on m and r, whose order is $\mathcal{O}\left(\sqrt{\frac{mr\ln n}{n}}\right)$, has been greatly reduced compared to that of Theorem 1, whose order is $\mathcal{O}\left(\sqrt{\frac{m^2r^3}{n}}\right)$.

Theorem 2 also implies that the expected reconstruction error of the learned bases W_n for MahNMF is, with high probability, not more than

$$\frac{2}{n} + R\sqrt{\frac{mr\ln(2mnR)}{2n}} + R\sqrt{\frac{\ln(2/\delta)}{2n}}$$

worse than the empirical reconstruction error of the learned bases W_n. This can be used to design the size of the training sample.

The proof procedures of Theorems 1 and 2 show that the generalization error bounds are increasing functions of $\|H\|_2$. Then, a small value of $\|H\|_2$ will lead to small generalization error bounds. This provides a theoretical justification for many successful NMF applications [22], [51], [52] that employ the ℓ_2-regularization on the representations H.

We compare the upper bounds of the generalization error in Theorems 1 and 2 in Fig. 3. It shows that when the sample size n is small, the upper bound in Theorem 2 will be smaller; and that when the sample size n is large, the upper bound in Theorem 1 will be smaller.

In practical applications, the sample sizes are usually small. We therefore combine Theorem 2 and inequalities (7) to obtain the following theorem:

Theorem 3: For ℓ_1-normalized MahNMF problems, assume that $\|x\|_1 \le R$. For any $\delta > 0$, with probability at least $1-\delta$, we have

$$\begin{aligned}R(W_{n,r}) - R(W_r) \quad &\le 2R\sqrt{\frac{mr\ln(2mnR)}{2n}} \\ &+\frac{4}{n} + 2R\sqrt{\frac{\ln(2/\delta)}{2n}}.\end{aligned}$$

Theorem 4 shows that given m and r, the estimation error, which is derived from the generalization bound, depends only on the learning algorithm of MahNMF, because some learning algorithms can further narrow the complexity of the loss class by using regularization terms.

IV. Asymptotic Approximation Error Bound

The approximation errors $R(W_r)$ are different for different distributions of $\mathcal{X}$. It is hard to derive a non-asymptotic uniform upper bound for the approximation error of different distributions. However, we can instead provide a tight asymptotic uniform upper bound for the approximation error, as follows:

Theorem 4: For MahNMF problems, when the reduced dimensionality r approaches m, we have

$$R(W_r) \le \mathcal{O}(r^{-\frac{1}{m}}).$$

The order of r is optimal.

A detailed proof is provided in the next section.

The asymptotic approximation error bound depends only on the reduced dimensionality r. Our approximation error bound is somewhat weak (because $m^{-1/m} \to 1, m \to \infty$); however, it is the first to be derived for MahNMF (and NMF) and it provides insight into the problem, namely that when r and m are large, the approximation error will decrease with respect to the increase of r of order $O(r^{-1/m})$.

V. Proofs

In this section, we present the proofs of the results in Sections II, III and IV. We begin by introducing the concentration inequalities, which play an important role in proving generalization error bounds.

A. Auxiliary Results

The following concentration inequality, well known as Hoeffding's inequality [53], is widely used for deriving generalization error bounds.

Theorem 5: Let $X_1, \ldots, X_n$ be independent random variables with the range $[a_i, b_i]$ for $i = 1, \ldots, n$. Let $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$, the following inequalities hold:

$$\Pr\{S_n - ES_n \ge \epsilon\} \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

$$\Pr\{ES_n - S_n \ge \epsilon\} \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Another well-known concentration inequality is MicDiarmid's inequality [54] (also known as the bounded-difference inequality).

Theorem 6: Let $X = (x_1, \ldots, x_n)$ be an independent and identically distributed sample and X^i a new sample with the

i-th example in X being replaced by an independent example x'_i. If there exists $c_1, \ldots, c_n > 0$ such that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following conditions:

$$|f(X) - f(X^i)| \le c_i, \forall i \in \{1, \ldots, n\}.$$

Then for any $X \in \mathcal{X}^n$ and $\epsilon > 0$, the following inequalities hold:

$$\Pr\{f(X) - Ef(X) \ge \epsilon\} \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (c_i)^2}\right),$$

$$\Pr\{Ef(X) - f(X) \ge \epsilon\} \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (c_i)^2}\right).$$

Glivenko-Cantelli Theorem [40] is often used to analyze the uniform convergence of the empirical reconstruction error to the expected reconstruction error. A relatively small complexity of the loss class is essential to prove a Glivenko-Cantelli class, which provides consistent generalization error bounds. The Rademacher complexity and covering number are the most used complexity measures.

The *Rademacher complexity* and *Gaussian complexity* are data-dependent complexity measures. They are often used to derive dimensionality-independent generalization error bounds and defined as follows:

Definition 3: Let $\sigma_1, \ldots, \sigma_n$ and $\gamma_1, \ldots, \gamma_n$ be independent Rademacher variables and independent standard normal variables, respectively. Let $x_1, \ldots, x_n$ be an independent and identically distributed sample and F a function class. The empirical Rademacher complexity and empirical Gaussian complexity are defined as:

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

and

$$\mathcal{G}_n(F) = E_\gamma \sup_{f \in F} \frac{2}{n} \sum_{i=1}^n \gamma_i f(x_i),$$

respectively. And the Rademacher complexity and Gaussian complexity are defined as:

$$\mathfrak{R}(F) = E_x \mathfrak{R}_n(F)$$

and

$$\mathcal{G}(F) = E_x \mathcal{G}_n(F),$$

respectively.

Using the symmetric distribution property of random variables, the following theorem holds:

Theorem 7: Let F be a real-valued function class on $\mathcal{X}$ and $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$. Let

$$\Phi(X) = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n (E_x f(x) - f(x_i)).$$

Then,

$$E_x \Phi(X) \le \mathfrak{R}(F).$$

proof. Let $X' = (x'_1, \ldots, x'_n) \in \mathcal{X}^n$ be an another sample distributed independently of X. We have

$$\begin{aligned}
&E_x \Phi(X) \\
&= E_x \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n (E_x f(x'_i) - f(x_i)) \\
&\le E_x \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n (f(x'_i) - f(x_i)) \\
&\text{(Since } f(x'_i) - f(x_i) \text{ has a symmetric distribution)} \\
&= E_{x,\sigma} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i (f(x'_i) - f(x_i)) \\
&= E_{x,\sigma} \sup_{f \in F} \frac{2}{n} \sum_{i=1}^n \sigma_i f(x_i) \\
&= \mathfrak{R}(F).
\end{aligned}$$

This concludes the proof. ■

The following theorem [55] proven utilizing Theorem 7 and Hoeffding's inequality plays an important role in proving generalization error bounds.

Theorem 8: Let F be a $[a, b]$-valued function class on $\mathcal{X}$ and $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$. For any $\delta > 0$, with probability at least $1 - \delta$, we have

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

Proof. Let

$$\Phi(X) = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n (E_x f(x) - f(x_i)).$$

We have

$$|\Phi(X) - \Phi(X^i)| \le \frac{1}{n} |f(x_i) - f(x'_i)| \le \frac{b - a}{n}.$$

Combining Theorem 7 and MicDiarmid's inequality, we have

$$\begin{aligned}
&\Pr\{\Phi(X) - \mathfrak{R}(F) \ge \epsilon\} \\
&\le \Pr\{\Phi(X) - E\Phi(X) \ge \epsilon\} \\
&\le \exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b - a)^2}\right).
\end{aligned}$$

Solving $\delta = \exp \frac{-2n\epsilon^2}{(b-a)^2}$ for ϵ gives the result. ■

The *covering number* [56] is also a data-dependent complexity measure of the complexity of a loss class.

Definition 4: 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 $\forall f \in F, \exists v_j$:

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

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

The following well-known result is due to Pollard [57], whose proof combines the Hoeffding's inequality and the covering number.

Theorem 9: Let $X_1^{2n} = \{x_1, \ldots, x_{2n}\}$ be $2n$ independent and identically distributed observations. For a function class F with the range $[a,b]$. Let $Ef = \int f(x)d\rho(x)$ and $E_n f = \frac{1}{n}\sum_{i=1}^n f(x_i)$. For any $\xi > 0$ and any $n \geq \frac{8(b-a)^2}{\xi^2}$, we have

$$\begin{aligned} &\Pr\left\{\sup_{f\in F}|Ef - E_n f| \geq \xi\right\} \\ &\leq 8E\mathcal{N}_1(F, \xi/8, X_1^{2n})\exp\left(\frac{-n\xi^2}{32(b-a)^2}\right). \end{aligned}$$

We provide an outline in Appendix A for the sake of self-containment of this paper.

B. Proof of Lemma 1

Let

$$h^* = \min_h \|x - Wh\|_1.$$

We have

$$\begin{aligned} &\|Wh^*\|_1 - \|x\|_1 \\ &\leq \|x - Wh^*\|_1 \\ &\leq \|x - W\mathbf{0}\|_1 \\ &= \|x\|_1. \end{aligned}$$

Then

$$\|Wh^*\|_1 \leq 2\|x\|_1 \leq 2R.$$

We also have that

$$\begin{aligned} &\|Wh^*\|_1 \\ &= \sum_{j=1}^m |(Wh^*)_j| = \sum_{j=1}^m \left|\sum_{k=1}^r W_{jk}h_k^*\right| \\ &= \sum_{j=1}^m \sum_{k=1}^r W_{jk}h_k^* = \sum_{k=1}^r \left(\sum_{j=1}^m W_{jk}\right) h_k^* \\ &= \sum_{k=1}^r h_k^* \\ &= \|h^*\|_1. \end{aligned}$$

Thus,

$$\|h^*\|_1 \leq 2R.$$

This concludes the proof of Lemma 1. ■

C. Proof of Theorem 1

Because the reconstruction error

$$f_W(x) = \min_{h\in\mathbb{R}_+^r} \|x - Wh\|_1$$

has a minimum in its function, it is hard to bound the Rademacher complexity $\mathfrak{R}(F_\mathcal{W})$ directly. We use the following two lemmas (see proofs in [58]) to bound $\mathfrak{R}(F_\mathcal{W})$ by finding a proper Gaussian process which can be easily bounded to upper bound the Rademacher complexity $\mathfrak{R}(F_\mathcal{W})$.

Lemma 2 (Slepian's Lemma): A Gaussian process of X is defined as $\Omega_X = \sum_{i=1}^n \gamma_i x_i$. Let Ω and Ξ be mean zero, separable Gaussian processes indexed by a common set $\mathcal{S}$, such that

$$E(\Omega_{s_1} - \Omega_{s_2})^2 \leq E(\Xi_{s_1} - \Xi_{s_2})^2, \forall s_1, s_2 \in \mathcal{S}.$$

Then

$$E\sup_{s\in\mathcal{S}}\Omega_s \leq E\sup_{s\in\mathcal{S}}\Xi_s.$$

The Gaussian complexity is related to the Rademacher complexity by the following lemma.

Lemma 3:

$$\mathfrak{R}(F) \leq \sqrt{\pi/2}\mathcal{G}(F).$$

We are now going to find a proper Gaussian process to upper bound the Rademacher complexity $\mathfrak{R}(F_\mathcal{W})$.

Lemma 4: For ℓ_1-normalized MahNMF problems, assume that $\|x\|_1 \leq R$. Then,

$$\mathfrak{R}(F_\mathcal{W}) \leq \frac{\sqrt{2\pi r}mrR}{\sqrt{n}}.$$

Proof. Let

$$\Omega_W = \sum_{k=1}^n \gamma_k \min_h \|x_k - Wh\|_1$$

and

$$\Xi_W = 2\sqrt{m}R\sum_{k=1}^n\sum_{i=1}^r\sum_{j=1}^m \gamma_{kij}\left\langle W^T e_j, e_i\right\rangle,$$

where e_i, e_j are natural basis and γ_k and γ_{kij} are independent standard norm variables indexed by k, i, j.

We first prove that

$$E(\Omega_{W_1} - \Omega_{W_2})^2 \leq E(\Xi_{W_1} - \Xi_{W_2})^2.$$

We have

$$\begin{aligned} &E(\Omega_{W_1} - \Omega_{W_2})^2 \\ &= E\left(\sum_{k=1}^n\left(\gamma_k \min_{h\in\mathbb{R}_+^r}\|x_k - W_1 h\|_1\right.\right. \\ &\quad \left.\left. -\gamma_k \min_{h\in\mathbb{R}_+^r}\|x_k - W_2 h\|_1\right)\right)^2 \\ &= \sum_{k=1}^n\left(\min_{h\in\mathbb{R}_+^r}\|x_k - W_1 h\|_1 - \min_{h\in\mathbb{R}_+^r}\|x_k - W_2 h\|_1\right)^2 \\ &= \sum_{k=1}^n\left(\min_{h\in\mathbb{R}_+^r}\max_{h'\in\mathbb{R}_+^r}\left(\|x_k - W_1 h\|_1 - \|x_k - W_2 h'\|_1\right)\right)^2 \\ &\leq \sum_{k=1}^n\left(\max_h \|x_k - W_1 h\|_1 - \|x_k - W_2 h\|_1\right)^2 \\ &\leq \sum_{k=1}^n\left(\max_h \|W_1 h - W_2 h\|_1\right)^2 \\ &= \sum_{k=1}^n\left(\max_h \sum_{i=1}^r h_i\|(W_1 - W_2)e_i\|_1\right)^2 \\ &\quad \text{(Using Cauchy-Schwarz inequality)} \\ &\leq \sum_{k=1}^n \max_h \sum_{i=1}^r h_i^2 \sum_{i=1}^r \|(W_1 - W_2)e_i\|_1^2 \end{aligned}$$

$$
\begin{aligned}
&\text{(Using } \|x\|_2 \le \|x\|_1) \\
&\le 4R^2 \sum_{k=1}^{n}\sum_{i=1}^{r} \|(W_1 - W_2)e_i\|_1^2 \\
&\text{(Using Cauchy-Schwarz inequality again)} \\
&\le 4mR^2 \sum_{k=1}^{n}\sum_{i=1}^{r} \|(W_1 - W_2)e_i\|_2^2 \\
&= 4mR^2 \sum_{k=1}^{n}\sum_{i=1}^{r}\sum_{j=1}^{m} \left(\langle W_1^T e_j, e_i\rangle - \langle W_2^T e_j, e_i\rangle\right)^2 \\
&= E(\Xi_{W_1} - \Xi_{W_2})^2.
\end{aligned}
$$

Now, we are ready to upper bound the Redamacher complexity $\mathfrak{R}(F_{\mathcal{W}})$.

$$
\begin{aligned}
&\mathfrak{R}(F_{\mathcal{W}}) \\
&\quad\text{(Using Lemma 3)} \\
&\le \frac{\sqrt{\pi/2}}{n} E \sup_W \Omega_W \\
&\quad\text{(Using Lemma 2)} \\
&\le \frac{\sqrt{\pi/2}}{n} E \sup_W \Xi_W \\
&= \frac{\sqrt{2\pi m}R}{n} E \sup_W \sum_{k=1}^{n}\sum_{i=1}^{r}\sum_{j=1}^{m} \gamma_{kij} \left\langle W^T e_j, e_i\right\rangle \\
&\quad\text{(Using Cauchy-Schwarz inquality)} \\
&\le \frac{\sqrt{2\pi m}R}{n} E \sup_W \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\sum_{k=1}^{n}\gamma_{kij}\right)^2} \\
&\quad \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{m} (\langle W^T e_j, e_i\rangle)^2} \\
&\le \frac{\sqrt{2\pi m}R}{n} E \sup_W \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\sum_{k=1}^{n}\gamma_{kij}\right)^2} \\
&\quad \sum_{i=1}^{r}\sum_{j=1}^{m} \left|\left\langle W^T e_j, e_i\right\rangle\right| \\
&\quad\text{(Using } \|W_i\|_1 = 1, i = 1, \ldots, r) \\
&\le \frac{\sqrt{2\pi m}rR}{n} E \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\sum_{k=1}^{n}\gamma_{kij}\right)^2} \\
&\quad\text{(Using Jensen's inquality)} \\
&\le \frac{\sqrt{2\pi m}rR}{n} \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{m} E\left(\sum_{k=1}^{n}\gamma_{kij}\right)^2} \\
&\quad\text{(Using orthogaussian properties of the } \gamma_{kij}) \\
&\le \frac{\sqrt{2\pi m}rR}{n} \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{m} n} \\
&\le \frac{\sqrt{2\pi r}mrR}{\sqrt{n}}.
\end{aligned}
$$

■

Note that

$$f_W(x) = \min_{h\in\mathbb{R}_+^r} \|x - Wh\|_1 \le \|x - W\mathbf{0}\|_1 \le R.$$

Theorem 1 can be proven combing Theorem 8 and Lemma 4.

D. Proof of Theorem 2

The following result, which bounds the covering number of the loss class of MahNMF, plays a central role in proving Theorem 2.

Lemma 5: Let $F_{\mathcal{W}_r}$ be the loss function class of ℓ_1-normalized MahNMF. Then

$$\ln \mathcal{N}_1(F_{\mathcal{W}_r}, \xi, 2n) \le mr \ln\left(\frac{2mR}{\xi}\right).$$

Proof. We will bound the covering number of $F_{\mathcal{W}}$ by bounding the covering number of the class $\mathcal{W}$. Cutting the subspace $[0,1]^m \subset \mathbb{R}^m$ into small m-dimensional regular solids with width ξ, there are a total of

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

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

$$\left\lceil\frac{1}{\xi}\right\rceil^{mr} \le \left(\frac{2}{\xi}\right)^{mr}$$

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

We will prove that for every W, there exists a $W' \in \mathcal{S}$ such that $|f_W - f_{W'}| \le \xi'$, where $\xi' = m\xi R$.

$$
\begin{aligned}
&|f_W - f_{W'}| \\
&= \left|\min_h \|x - Wh\|_{\ell_1} - \min_h \|x - W'h\|_{\ell_1}\right| \\
&\le \left|\max_h (\|x - Wh\|_{\ell_1} - \|x - W'h\|_{\ell_1})\right| \\
&\le \max_h |(\|x - Wh\|_{\ell_1} - \|x - W'h\|_{\ell_1})| \\
&\le \max_h \|(x - Wh) - (x - W'h)\|_{\ell_1} \\
&\le \max_h \sum_{i=1}^{r} h_i \|(W - W')e_i\|_{\ell_1} \\
&\le \max_h \sum_{i=1}^{r} h_i m\xi/2 \\
&\quad\text{(Using Lemma 1)} \\
&\le m\xi R \\
&= \xi'.
\end{aligned}
$$

The third inequality holds because $||a| - |b|| \le |a - b|$.

Let the metric d be the absolute difference metric. According to Definition 4, for $\forall f_W \in F_{\mathcal{W}}$, there is a $W' \in \mathcal{S}$ such that

$$
\begin{aligned}
&\|d(f_W(X), f_{W'}(X))\|_{\ell_1} \\
&= \left[\sum_{i=1}^{2n} d(f_W(x_i), f_{W'}(x_i))\right] \\
&\le 2n\xi'.
\end{aligned}
$$

Thus,

$$\mathcal{N}_1(F_{\mathcal{W}}, \xi', 2n) \le |\mathcal{S}| \le \left(\frac{2}{\xi}\right)^{mr} = \left(\frac{2mR}{\xi'}\right)^{mr}.$$

Taking log on both sides, we have

$$\ln \mathcal{N}_1(F_{\mathcal{W}}, \xi', 2n) \leq mr \ln\left(\frac{2mR}{\xi'}\right).$$

■

The first part of Theorem 2 can be proven combing Theorem 9 and Lemma 5.

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

$$\sup_{f_W \in F_{\mathcal{W}}} |Ef_W - E_n f_W| \leq 2\varepsilon + \sup_{f_{W,\varepsilon} \in F_{\mathcal{W},\varepsilon}} |Ef_{W,\varepsilon} - E_n f_{W,\varepsilon}|.$$

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

$$\begin{aligned} &P\left\{\sup_{f_{W,\varepsilon} \in F_{\mathcal{W},\varepsilon}} |Ef_W - E_n f_W| \geq \xi\right\} \\ &\leq 2|F_{\mathcal{W},\varepsilon}| \exp(-2nR^{-2}\xi^2). \end{aligned}$$

Let

$$\begin{aligned} &2|F_{\mathcal{W},\varepsilon}| \exp(-2nR^{-2}\xi^2) \\ &= 2\left(\frac{2mR}{\varepsilon}\right)^{mr} \exp(-2nR^{-2}\xi^2) \\ &= \delta. \end{aligned}$$

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

$$\begin{aligned} &\sup_{f_{W,\varepsilon} \in F_{\mathcal{W},\varepsilon}} |Ef_W - E_n f_W| \\ &\leq R\sqrt{\frac{mr\ln(2mR/\varepsilon) + \ln(2/\delta)}{2n}}. \end{aligned}$$

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

$$\begin{aligned} &\sup_{f_W \in F_{\mathcal{W}}} |Ef_W - E_n f_W| \leq 2\varepsilon \\ &+R\sqrt{\frac{mr\ln(2mR/\varepsilon) + \ln(2/\delta)}{2n}}. \end{aligned}$$

Let $\varepsilon = 1/n$, with probability at least $1-\delta$, we have

$$\begin{aligned} &\sup_{f_W \in F_{\mathcal{W}}} |Ef_W - E_n f_W| \\ &\leq \frac{2}{n} + R\sqrt{\frac{mr\ln(2mnR) + \ln(2/\delta)}{2n}} \\ &\leq \frac{2}{n} + R\sqrt{\frac{mr\ln(2mnR)}{2n}} + R\sqrt{\frac{\ln(2/\delta)}{2n}}. \end{aligned}$$

This concludes the proof of the second part of Theorem 2. ■

E. *Proof of Theorem 4*

According to Theorem 3 of Gruber [41], asymptotical results have been proven for the distortion of high resolution vector quantization.

Let $C_1, \ldots, C_k$ be k measurable sets which tile a measurable set $C \in \mathbb{R}^m$ and $\{c_1, \ldots, c_k \in \mathbb{R}^m\}$ be a *codebook*. To each signal $x \in C$ assign the *codevector* c_i if $x \in C_i$. Let p be the density of the source which generates the signal x. Then, the distortion is defined as

$$\int_C \min_{i=1,\ldots,k} \{f(\|c_i - x\|)\} p(x)dx,$$

where f is a measure (loss function) for the quality of the quantizer and $\|\cdot\|$ is a norm on $\mathbb{R}^m$.

The minimum distortion is given by

$$\inf_{C_1,\ldots,C_r} \left\{\int_C \min_{i=1,\ldots,r} \{f(\|c_i - x\|)\} p(x)dx\right\}. \tag{8}$$

The following theorem describes the asymptotic result of the minimum distortion of vector quantization [41].

Theorem 10: Let $f : [0, +\infty) \to [0, +\infty)$ satisfying $f(0) = 0$, f is continuous and strictly increasing and, for any given $s > 1$, the quotient $f(st)/f(t)$ is decreasing and bounded above for $t > 0$. Then there are constants $A, B > 0$, depending only on f and $\|\cdot\|$, such that the following statement holds: let $C \in \mathbb{R}^m$ be compact and measurable with $|C| > 0$ and p be the distribution density function of the source which generates x. Then (8) is asymptotically equal to

$$B\left(\int_C p(x)^{\frac{m}{A+m}} dx\right)^{\frac{A+m}{m}} f(r^{-\frac{1}{m}}),$$

as $r \to \infty$.

Now, we are ready to prove Theorem 4.

Proof of Theorem 4. MahNMF can be viewed as a coding scheme. Therefore, we can analyze MahNMF using Theorem 10. We relate the reconstruction error and the distortion of quantization as follows:

$$\begin{aligned} &R(W_r) \\ &= \min_{W \in \mathbb{R}_+^{m\times r}} \int_{\mathcal{X}} f_W(x) d\rho \\ &= \min_{W \in \mathbb{R}_+^{m\times r}} \int_{\mathcal{X}} \min_{h \in \mathbb{R}_+^r} \|x - Wh\|_1 p(x)dx \\ &\leq \min_{W \in \mathbb{R}_+^{m\times r}} \int_{\mathcal{X}} \min_{x' \in \{W_1,\ldots,W_r\}} \|x - x'\|_1 p(x)dx \\ &= \int_{\mathcal{X}} \min_{x' \in \{W_1,\ldots,W_r \in \mathbb{R}_+^m\}} \|x - x'\|_1 p(x)dx. \end{aligned}$$

We can carefully choose $\{W_1, \ldots, W_r \in \mathbb{R}_+^m\}$ such that

$$\int_{\mathcal{X}} \min_{x' \in \{W_1,\ldots,W_r \in \mathbb{R}_+^m\}} \|x - x'\|_1 p(x)dx$$

is equal to the quantization distortion problem (8). Then, the minimum of $R(W_r)$ is upper bounded by the minimum distortion.

Let $f(x) = x$ and $\|\cdot\| = \|\cdot\|_1$. It can be easily verified that f satisfies $f : [0, +\infty) \to [0, +\infty)$, $f(0) = 0$, f is continuous and strictly increasing and, for any given $s > 1$, the quotient $f(st)/f(t)$ is non-increasing and bounded above for $t > 0$. According to Theorem 10, the approximation error has an asymptotical bound as:

$$R(W_r) \leq B\left(\int_{\mathcal{X}} p(x)^{\frac{m}{A+m}} dx\right)^{\frac{A+m}{m}} f(r^{-\frac{1}{m}}).$$

Since the reduced dimensionality $r < \min(m, n)$, to use Theorem 10, we have to assume that the feature dimensionality m is sufficiently large. Thus, when m is sufficiently large and r approaches m, it holds that $R(W_r) \leq Br^{-\frac{1}{m}}$. When m is finite and r approaches m, it trivially holds that $R(W_r) = 0 \leq$

$Br^{-\frac{1}{m}}$. We therefore have proven the inequality in Theorem 4.

We then prove that this bound is tight about the order of the reduced dimensionality r. We will show that if H is orthogonal and $\|H_i\|_1 = 1, i = 1, \ldots, k$, then

$$R(W_r) = \int_{\mathcal{X}} \min_{x' \in \{W_1, \ldots, W_r \in \mathbb{R}_+^r\}} \|x - x'\|_1 p(x) dx.$$

We have

$$\begin{aligned}
&R(W_r) \\
&= \min_{W \in \mathbb{R}_+^{m \times r}} \int_{\mathcal{X}} \min_{h \in \mathbb{R}_+^r} \|x - Wh\|_1 p(x) dx \\
&= \min_{W \in \mathbb{R}_+^{m \times r}} \int_{\mathcal{X}} \min_{h \in \mathbb{R}_+^r} \left\| x - \sum_{j=1}^{k} h_j W_j \right\|_1 p(x) dx \\
&\quad \text{(Because of } \|H_i\|_1 = 1.) \\
&= \min_{W \in \mathbb{R}_+^{m \times r}} \int_{\mathcal{X}} \min_{h \in \mathbb{R}_+^r} \left\| \sum_{j=1}^{k} h_j (x - W_j) \right\|_1 p(x) dx \\
&= \min_{W \in \mathbb{R}_+^{m \times r}} \int_{\mathcal{X}} \min_{W_j, j=1,\ldots,r} \|x - W_j\|_1 p(x) dx \\
&= \int_{\mathcal{X}} \min_{x' \in \{W_1, \ldots, W_r \in \mathbb{R}_+^m\}} \|x - x'\|_1 p(x) dx.
\end{aligned}$$

The fourth equality holds because H is orthogonal. The orthogonality condition implies that in each column of H only one element is nonzero and thus $h_j \in \{0, 1\}$.

We have proven that the approximation error of MahNMF is upper bounded by and sometimes equal to the minimum distortion. Therefore, no tighter bound with respect to r can be obtained. ■

We note that this asymptotic approximation error bound can be applied for NMF problems as well. Using the same proof method, the following holds:

Theorem 11: For NMF problems, when the reduced dimensionality r approaches m, we have

$$R(W_r) \leq \mathcal{O}(r^{-\frac{2}{m}}).$$

The order of r is optimal.

VI. CONCLUSIONS

In this paper, we have analyzed the performance of MahNMF, which effectively and efficiently restores the low-rank and sparse structures of a data matrix. In particular, we decomposed the expected reconstruction error into the estimation and approximation errors. The estimation error was bounded using the generalization error bound. We then obtained the generalization error bounds by implementing both the Rademacher complexity and the covering number methods. While the obtained bounds are dimensionality-dependent, we conjecture that a more subtle approach can be used for the Rademacher complexity method to derive dimensionality-independent generalization error bounds and estimation errors.

For given m and r, the approximation error depends only on the distribution of $\mathcal{X}$. We verified this by deriving an asymptotic uniform upper bound for the approximation error by using the asymptotic results of the minimum distortion of vector quantization. Moreover, we proved that the bound is tight regarding to r, and in doing so revealed a clear relationship between the approximation error and reduced dimensionality for MahNMF.

APPENDIX A
THE OUTLINE OF THE PROOF OF THEOREM 9

The outline of proof is of four steps.

(1) Symmetrization (see a detailed proof in [59]). Let X' and X be independent and identically distributed sample sets drawn from $\mathcal{X}$, respectively. For any $n \geq \frac{8(b-a)^2}{\xi^2}$, we have

$$\begin{aligned}
&\Pr\left\{ \sup_{f \in F} |Ef - E_n f| \geq \xi \right\} \\
&\leq 2\Pr\left\{ \sup_{f \in F} |E'_n f - E_n f| \geq \xi/2 \right\},
\end{aligned}$$

where $E'_n f = \frac{1}{n} \sum_{i=1}^n f(x'_i)$.

(2) Permutation. Using the symmetric distribution property of $f(x'_i) - f(x_i)$ and the union bound, it gives

$$\begin{aligned}
&\Pr\left\{ \sup_{f \in F} |E'_n f - E_n f| \geq \xi/2 \right\} \\
&\leq 2\Pr\left\{ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \geq \xi/4 \right\}.
\end{aligned}$$

(3) Combination. Setting G be an $\xi/8$-cover of F with respect to ℓ_1 norm and applying the union bound, we have

$$\begin{aligned}
&\Pr\left\{ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \geq \xi/4 \right\} \\
&\leq \Pr\left\{ \sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \geq \xi/8 \right\} \\
&\leq E\mathcal{N}_1(F, \xi/8, X_1^{2n}) \Pr\left\{ \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \geq \xi/8 \right\}.
\end{aligned}$$

(4) Concentration. By the union bound of the Hoeffding's inequality, we get

$$\Pr\left\{ \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \geq \xi/8 \right\} \leq 2 \exp\left(\frac{-n\xi^2}{32(b-a)^2} \right).$$

At last, combining the four steps, we have

$$\begin{aligned}
&\Pr\left\{ \sup_{f \in F} |Ef - E_n f| \geq \xi \right\} \\
&\leq 8E\mathcal{N}_1(F, \xi/8, X_1^{2n}) \exp\left(\frac{-n\xi^2}{32(b-a)^2} \right).
\end{aligned}$$

■

TABLE I
THE RELATIVE ERRORS OF THE RECONSTRUCTIONS BY NMF, KLNMF AND MAHNMF ON THE PIE DATASET

Algorithm	NMF	KLNMF	MahNMF	NMF	KLNMF	MahNMF
PIE Dataset	Training set			Test set		
No noise	.010±.000	**.009±.000**	.012±.001	**.012±.000**	**.012±.000**	.014±.000
Occlusion	.075±.002	.113±.003	**.063±.002**	.070±.001	.106±.001	**.064±.001**
Laplace noise	0.72±.002	.079±.003	**.046±.001**	.056±.001	.054±.001	**.044±.001**
Salt & Pepper	.097±.004	.115±.003	**.021±.003**	.052±.001	.049±.001	**.020±.001**
Gaussian noise	**.024±.001**	.028±.001	**.024±.001**	**.023±.000**	.024±.000	**.023±.000**
Poisson noise	.013±.000	**.011±.000**	.012±.000	.014±.000	**.013±.000**	.014±.000

APPENDIX B
THE ROBUSTNESS AND PERFORMANCE OF MAHNMF

In Section II-B we theoretically analyzed that MahNMF is more robust (to noise) than NMF and KLNMF. Guan et al. [14] had conducted extensive empirical experiments to validate this point. In this appendix, we borrow one reconstruction experiment on the PIE dataset [60] from [14] to show both the robustness and good performance of MahNMF.

The relative reconstruction error is defined as $\|X - X'\|_F^2/\|X\|_F^2$, wherein X and X' denote the original image and reconstructed image, respectively. The smallest reconstruction error in each subline of Table I is shown in bold. The experiment settings are referred to [14]. Table I shows that MahNMF reconstructs the face images better in both the training and test sets when the training set is contaminated by occlusion, Laplace noise, salt and pepper noise and Gaussian noise. MahNMF therefore successfully handles the heavy-tailed noise and performs robustly in the presence of outliers. We do not repeat the other experiment results and suggest readers refer to [14] for detailed information.

ACKNOWLEDGMENT

The authors would like to thank Dr. Naiyang Guan for providing useful suggestions and codes of MahNMF.

REFERENCES

[1] C. Lynch, “Big data: How do your data grow?” *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.
[2] M. Piernik, D. Brzezinski, and T. Morzy, “Clustering xml documents by patterns,” *Knowledge and Information Systems*, pp. 1–28, 2015.
[3] D. Tao, L. Jin, Y. Wang, and X. Li, “Person reidentification by minimum classification error-based kiss metric learning,” *IEEE Transactions on Cybernetics*, vol. 45, pp. 242–252, Feb. 2015.
[4] X. Wen, L. Shao, W. Fang, and Y. Xue, “Efficient feature selection and classification for vehicle detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, pp. 508–517, Mar. 2015.
[5] Y. Luo, D. Tao, B. Geng, C. Xu, and S. Maybank, “Manifold regularized multitask learning for semi-supervised multilabel image classification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 22, pp. 523–536, Feb. 2013.
[6] Z. Zhang and K. Zhao, “Low-rank matrix approximation with manifold regularization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1717–1729, Jul. 2013.
[7] D. Tao, L. Jin, W. Liu, and X. Li, “Hessian regularized support vector machines for mobile image annotation on the cloud,” *IEEE Transactions on Multimedia*, vol. 15, pp. 833–844, Jun. 2013.
[8] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.
[9] R. Yan, L. Shao, and Y. Liu, “Nonlocal hierarchical dictionary learning using wavelets for image denoising,” *IEEE Transactions on Image Processing*, vol. 22, pp. 4689–4698, Dec. 2013.
[10] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition,” *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.
[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of the ACM*, vol. 58, no. 3, pp. 11: 1–37, 2011.
[13] T. Zhou and D. Tao, “GoDec: Randomized low-rank & sparse matrix decomposition in noisy case,” in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, WA, Jul. 2011, pp. 33–40.
[14] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, “MahNMF: Manhattan non-negative matrix factorization,” *arXiv preprint arXiv:1207.3438*, 2012.
[15] L. Liu, M. Yu, and L. Shao, “Multiview alignment hashing for efficient image search,” *IEEE Transactions on Image Processing*, vol. 24, pp. 956–966, Mar. 2015.
[16] C. Xu, D. Tao, and C. Xu, “Multi-view intact space learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2015.2417578, 2015.
[17] L. Zhang, X. Zhen, and L. Shao, “Learning object-to-class kernels for scene classification,” *IEEE Transactions on Image Processing*, vol. 23, pp. 3241–3253, Aug. 2014.
[18] A. Khan, M. A. Jaffar, and L. Shao, “A modified adaptive differential evolution algorithm for color image segmentation,” *Knowledge and Information Systems*, vol. 43, no. 3, pp. 583–597.
[19] M. N. Schmidt, J. Larsen, and K. Lyngby, “Wind noise reduction using non-negative sparse coding,” in *IEEE International Workshop on Machine Learning for Signal Processing*, Thessaloniki, Greece, Aug. 2007, pp. 431–436.
[20] T. Zhang, B. Fang, Y. Y. Tang, G. He, and J. Wen, “Topology preserving non-negative matrix factorization for face recognition,” *IEEE Transactions on Image Processing*, vol. 17, pp. 574–584, Apr. 2008.
[21] A. Dessein, A. Cont, and G. Lemaitre, “Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence,” in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, Aug. 2010, pp. 489–494.
[22] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, “Text mining using non-negative matrix factorizations,” in *Proceedings of SIAM International Conference on Data Mining (SDM)*. SIAM, Lake Buena Vista, Florida, Apr. 2004, pp. 452–456.
[23] S. Arora, R. Ge, and A. Moitra, “Learning topic models–going beyond svd,” in *Proceedings of the 53rd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, Hyatt Regency, New Jersey, Oct. 2012, pp. 1–10.
[24] J. M. Tomczak and A. Gonczarek, “Decision rules extraction from data stream in the presence of changing context for diabetes treatment,” *Knowledge and Information Systems*, vol. 34, no. 3, pp. 521–546, 2013.
[25] 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.

[26] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS computational biology*, vol. 4, no. 7, p. e1000029, 2008.

[27] S. Zhang, Q. Li, J. Liu, and X. J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules," *Bioinformatics*, vol. 27, no. 13, pp. 401–409, 2011.

[28] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: An optimal gradient method for nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 60, pp. 2882–2898, Jun. 2012.

[29] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, Dec. 2012, pp. 1214–1222.

[30] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, pp. 845–869, May. 2014.

[31] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI 10.1109/TPAMI.2015.2456899.

[32] T. Yang, M. Mahdavi, R. Jin, L. Zhang, and Y. Zhou, "Multiple kernel learning from noisy labels by stochastic programming," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, Edinburgh, Scotland, Jun. 2012, pp. 233–240.

[33] D. Donoho and V. Stodden, "When does non-negative matrix factorization give correct decomposition into parts?" in *Proceedings of Advances in neural information processing systems (NIPS)*, Lake Tahoe, Nevada, Dec. 2003.

[34] L. B. Thomas, "Problem 73-14, rank factorization of nonnegative matrices," *SIAM Review*, vol. 16, no. 3, pp. 393–394, 1974.

[35] M. Kaykobad, "On nonnegative factorization of matrices," *Linear Algebra and Its Application*, vol. 96, pp. 27–33, 1987.

[36] A. Maurer and M. Pontil, "K-dimensional coding schemes in Hilbert spaces," *IEEE Transactions on Information Theory*, vol. 56, pp. 5839–5846, Nov. 2010.

[37] V. Vapnik, *The nature of statistical learning theory*. springer, 2000.

[38] T. Poggio and C. Shelton, "On the mathematical foundations of learning," *American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[39] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2003.

[40] S. Mendelson, "A few notes on statistical learning theory," in *Advanced Lectures on Machine Learning*, S. Mendelson and A. J. Smola, Eds. Springer, 2003, pp. 1–40.

[41] P. M. Gruber, "Optimum quantization and its applications," *Advances in Mathematics*, vol. 186, no. 2, pp. 456–497, 2004.

[42] X. Zhang, N. Guan, L. Lan, D. Tao, and Z. Luo, "Box-constrained projective nonnegative matrix factorization via augmented lagrangian method," in *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, Beijing, China, Jul. 2014.

[43] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[44] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[45] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[46] D. Kuang, H. Park, and C. H. Ding, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of SIAM International Conference on Data Mining (SDM)*. SIAM, Anaheim, California, Apr. 2012, pp. 106–117.

[47] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 1087–1099, Jul. 2012.

[48] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, Beijing, China, Aug. 2012, pp. 453–461.

[49] F. Wang, T. Li, and C. Zhang, "Semi-supervised clustering via matrix factorization," in *Proceedings of SIAM Conference on Data Mining*. SIAM, Atlanta, Georgia, Apr. 2008, pp. 1–12.

[50] O. Dekel, "From online to batch learning with cutoff-averaging," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Vancouver, B.C., Dec. 2009, pp. 377–384.

[51] V. P. Pauca, J. Piper, and R. J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear algebra and its applications*, vol. 416, no. 1, pp. 29–47, 2006.

[52] 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.

[53] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[54] C. McDiarmid, "Concentration," in *Probabilistic methods for algorithmic discrete mathematics*, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, Eds. Springer, 1998, pp. 195–248.

[55] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. The MIT Press, 2012.

[56] T. Zhang, "Covering number bounds of certain regularized linear function classes," *Journal of Machine Learning Research*, vol. 2, pp. 527–550, 2002.

[57] D. Pollard, *Convergence of stochastic procresses*. Springer-Verlag, 1984.

[58] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[59] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Lusburg, and G. Rätsch, Eds. Springer, 2004, pp. 169–207.

[60] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1615–1618, Dec. 2003.

Tongliang Liu received the B.E. degree in electronics engineering and information science from the University of Science and Technology of China, in 2012. He is currently pursuing the Ph.D. degree in computer science from the University of Technology, Sydney. He won the best paper award in the IEEE International Conference on Information Science and Technology 2014.

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

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 100+ 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 ICDM07, the best student paper award in IEEE ICDM13, and the 2014 ICDM 10 Year Highest-Impact Paper Award.