Robustness Analysis of Structured Matrix Factorization via Self-Dictionary Mixed-Norm Optimization

Xiao Fu, *Member, IEEE*, and Wing-Kin Ma, *Senior Member, IEEE*

***Abstract*—We are interested in a low-rank matrix factorization problem where one of the matrix factors has a special structure; specifically, its columns live in the unit simplex. This problem finds applications in diverse areas such as hyperspectral unmixing, video summarization, spectrum sensing, and blind speech separation. Prior works showed that such a factorization problem can be formulated as a self-dictionary sparse optimization problem under some assumptions that are considered realistic in many applications, and convex mixed norms were employed as optimization surrogates to realize the factorization in practice. Numerical results have shown that the mixed-norm approach demonstrates promising performance. In this letter, we conduct performance analysis of the mixed-norm approach under noise perturbations. Our result shows that using a convex mixed norm can indeed yield provably good solutions. More importantly, we also show that using nonconvex mixed (quasi) norms is more advantageous in terms of robustness against noise.**

***Index Terms*—Matrix factorization, performance analysis, self-dictionary sparse optimization.**

I. Introduction

In SIGNAL processing and machine learning, there are scenarios in which the measured data points can be modeled as convex combinations of some vectors; i.e., for the ℓth measured data point $\mathbf{x}_\ell \in \mathbb{R}^M$, we have

$$\mathbf{x}_\ell \approx \mathbf{A}\mathbf{s}_\ell,\ \ell = 1, \ldots, L, \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N] \in \mathbb{R}^{M \times N}$ is a basis matrix with $N \leq M$, and $\mathbf{s}_\ell \in \mathbb{R}^N$ is a coefficient vector that satisfies

$$\mathbf{s}_\ell \geq \mathbf{0}, \quad \mathbf{1}^T\mathbf{s}_\ell = 1,\ \ell = 1, \ldots, L. \tag{2}$$

One particular example for which the above model applies is *hyperspectral unmixing* (HU) of remotely sensed hyperspectral images [1], [2]. There, $\mathbf{x}_\ell$ is a high dimensional pixel measured at multiple spectral bands, $\mathbf{a}_n$'s denote the spectral signatures of the materials that constitute the pixels, and $\mathbf{s}_\ell$ denotes the fractions of the materials contained in pixel ℓ, which is known to satisfy (2) by nature of the application. Another example is *nonnegative matrix factorization* (NMF), with applications such as text mining and document clustering. There, we have seen (1)–(2) being used to model the NMF problem [3]. Recently, this signal model also finds applications in blind source separation [4], power spectrum sensing [5], [6], and video summarization [7]. In the above applications, we are interested in recovering $\mathbf{A}$ and/or $\{\mathbf{s}_\ell\}_{\ell=1}^L$ from $\{\mathbf{x}[\ell]\}_{\ell=1}^L$; e.g., in HU we desire to extract the spectral signatures of the materials and their portions in each pixel. Estimating $\mathbf{A}$ and $\{\mathbf{s}_\ell\}$ can be viewed as a structured low-rank matrix factorization problem that aims at factoring the data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$ into $\mathbf{A}$ and $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_L]$, where the columns of $\mathbf{S}$ are constrained to lie in the unit simplex.

For the aforementioned factorization problem in the noiseless case (i.e., $\mathbf{X} = \mathbf{A}\mathbf{S}$), it was shown that the identifiability of $\mathbf{A}$ and $\mathbf{S}$ can be guaranteed up to an ordering permutation if $\mathbf{X}$ satisfies a special condition, namely, that the columns of $\mathbf{A}$ appear in $\mathbf{X}$. This condition is formally described as follows.

(A1) There exist indices $\ell_1, \ldots, \ell_N$, such that $\mathbf{s}_{\ell_n} = \mathbf{e}_n$ for $n = 1, \ldots, N$, where $\mathbf{e}_n$ denotes the unit vector with the nth element being one.

Notice that, under (A1) and in the noiseless case, recovering $\mathbf{A}$ boils down to identifying $\Lambda = \{\ell_1, \ldots, \ell_N\}$, since $\mathbf{A} = \mathbf{X}_\Lambda$, where $\mathbf{X}_\Lambda$ denotes the submatrix of $\mathbf{X}$ consisting of the columns indexed by Λ. In the noisy case, if the noise is below a certain level, $\hat{\mathbf{A}} = \mathbf{X}_\Lambda$ can still serve as a good estimate of $\mathbf{A}$; after obtaining $\hat{\mathbf{A}}$, $\{\mathbf{s}_\ell\}$ can be easily estimated by solving a constrained least squares problem (cf. (1)–(2)). In fact, (A1) has been recognized as a reasonable and useful assumption in many areas, and it has different names in different contexts, such as *local dominance* in image and speech separation [4], [8], *pure pixel assumption* in HU [1], [9], and *separability condition* in NMF [3]. (A1) is considered particularly meaningful in applications where some data points living in a rank-one subspace can be found. Taking HU as an example, since there are pixels consisting of only one material in many cases, (A1) is considered a mild assumption.

A number of algorithms and approaches have been proposed to identify Λ; see [1], [2] for comprehensive surveys. Among the various approaches, we are interested in a recently proposed approach that uses sparse representation. To be specific, under (A1), identifying Λ amounts to finding a few columns of $\mathbf{X}$ that can represent all other columns of $\mathbf{X}$ by convex combinations [7], [10]. Hence, the formulated problem is similar to the

Manuscript received September 03, 2015; revised November 02, 2015; accepted November 02, 2015. Date of publication November 05, 2015; date of current version November 19, 2015. This work was supported in part by a General Research Fund of Hong Kong Research Grant Council under Project CUHK14205414. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Deanna Needell.

X. Fu is with Department of Electrical Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: xfu@umn.edu).

W. K. Ma is with Department of Electronic Engineering, the Chinese University of Hong Kong, NT, Hong Kong (e-mail: wkma@ieee.org).

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/LSP.2015.2498523

multiple measurement vectors (MMV) problem in compressive sensing [11], which selects a basis from an over-complete dictionary to represent a set of measurement vectors. The main difference is that the over-complete dictionary here is the data matrix itself, resulting in the so-called *self-dictionary MMV* (SD-MMV) formulation [9], [12].

Like MMV, SD-MMV is a cardinality optimization problem, which is computationally hard. Greedy pursuits were proposed in [9], [12] to tackle the SD-MMV problem, and robustness analysis in the presence of noise was presented. Another line of works [13]–[15] considered formulations that can be regarded as variants of SD-MMV, and relaxed them to linear programs. There, robustness in the noisy case was also shown. On the other hand, although mixed norm-based sparse optimization is considered 'natural' for basis selection-like problems [16], [17], and its application to SD-MMV indeed demonstrated empirically good performance in practice [10], [18], its robustness analysis against noise has not been investigated. In this work, we consider performance analysis of mixed norm-based self-dictionary sparse optimization. Specifically, we employ the convex mixed norm adopted in [10] and its nonconvex counterpart as optimization surrogates for SD-MMV, and analyze their performance in the presence of bounded noise. We show that, using such surrogates, two noise-robust variants of the SD-MMV formulation are theoretically guaranteed to identify Λ perfectly, if the noise is below a certain level. Notice that nonconvex mixed norms have not been considered for SD-MMV before, and our results are the first to show that employing nonconvex optimization surrogates can lead to provably better results. Numerical results are presented to support our analysis.

II. SD-MMV Sparse Optimization

We consider the following data model,

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{V}, \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\mathbf{S} \in \mathbb{R}^{N\times L}$ are defined as before, and $\mathbf{V}$ is noise. The columns of $\mathbf{V}$ are assumed to be bounded, i.e., $\|\mathbf{v}_\ell\|_2 \leq \epsilon$, where $\epsilon \geq 0$. As reviewed previously, our aim is to identify $\Lambda = \{\ell_1, \ldots, \ell_N\}$ under the assumption in (A1). To this end, consider the following SD-MMV formulation [9]:

$$\begin{aligned} \min_{\mathbf{C}\in\mathbb{R}^{L\times L}} \quad & \|\mathbf{C}\|_{\text{row}-0} \\ \text{s.t.} \quad & \|\mathbf{x}_\ell - \mathbf{X}\mathbf{c}_\ell\|_2 \leq \lambda,\ \ell = 1, \ldots, L, \\ & \mathbf{C} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{C} = \mathbf{1}^T. \end{aligned} \tag{4}$$

Here, $\|\mathbf{C}\|_{\text{row}-0}$ counts the number of the non-zero rows of $\mathbf{C}$, $\mathbf{c}_\ell$ denotes the ℓth column of $\mathbf{C}$, and $\lambda \geq 0$ is given. Problem (4) is called a self-dictionary sparse formulation because $\mathbf{X}$ is used as a dictionary to perform sparse optimization. To explain how SD-MMV leads to identification of Λ, let us assume that, without loss of generality, $\Lambda = \{1, \ldots, N\}$. It can be shown that, when noise is absent and $\lambda = 0$, Problem (4) has an optimal solution [9]:

$$\mathbf{C}^\star = \begin{bmatrix} \mathbf{S} \\ \mathbf{0} \end{bmatrix}; \tag{5}$$

(where $\mathbf{S}$ is the true coefficient matrix in (3)). This means that Λ can be identified by inspecting the non-zero rows of $\mathbf{C}^\star$. In addition, it was shown in [9] that $\mathbf{C}^\star$ in (5) remains being the optimal solution to (4) in the presence of noise under some conditions. Nevertheless, dealing with the optimization objective $\|\mathbf{C}\|_{\text{row}-0}$ is hard. Here, we are interested in the following approximation:

$$\begin{aligned} \min_{\mathbf{C}\in\mathbb{R}^{L\times L}} \quad & \|\mathbf{C}\|_{q,p}^p \\ \text{s.t.} \quad & \|\mathbf{x}_\ell - \mathbf{X}\mathbf{c}_\ell\|_2 \leq \lambda, \quad \ell = 1, \ldots, L, \\ & \mathbf{C} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{C} = \mathbf{1}^T, \end{aligned} \tag{6}$$

where $\|\mathbf{C}\|_{q,p}^p = \sum_{\ell=1}^L \|\mathbf{c}^\ell\|_q^p$, $p, q > 0$, with $\mathbf{c}^\ell$ being the ℓth row of $\mathbf{C}$. Note that $\|\mathbf{C}\|_{q,p}$ is called the l_q/l_p mixed norm, and is convex for $p \geq 1$ and $q \geq 1$; otherwise it is a quasi-norm and is nonconvex. Particularly, the l_q/l_1 mixed norm, $q > 1$, is widely used to approximate regular MMV [16] and SD-MMV [7], [10], [18].

In this letter, we focus on the following analysis problem: analyze conditions under which the solution of Problem (6) guarantees perfect recovery of Λ in the presence of noise. Our analysis will concentrate on $0 < p \leq 1$, $q = \infty$, which covers both convex and nonconvex mixed (quasi-) norm optimization.

We noticed that although Problem (6) has not been explicitly considered in the literature, the following related formulation was popularly considered in practice [7], [10], [18], [19]:

$$\begin{aligned} \min_{\mathbf{C}\in\mathbb{R}^{L\times L}} \quad & \|\mathbf{C}\|_{q,p}^p + \lambda\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 \\ \text{s.t.} \quad & \mathbf{C} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{C} = \mathbf{1}^T, \end{aligned} \tag{7}$$

with $q > 1$ and $p = 1$. In this work, we will also analyze the performance of (7) with $0 < p \leq 1$ and $q = \infty$.

III. Main Results

We first consider characterizing the solution to Problem (6) with $q = \infty$ and $p \in (0, 1]$. Intuitively, for a solution $\mathbf{C}$, if $\|\mathbf{c}^\ell\|_\infty$ for $\ell \in \Lambda$ are significant while $\mathbf{c}^\ell \approx \mathbf{0}$ for $\ell \notin \Lambda$, Λ can be identified by taking the indices of the rows with large norms. With this in mind, let us first define

$$\alpha(\mathbf{A}) = \min_{n\in\{1,\ldots,N\}} \min_{\mathbf{1}^T\boldsymbol{\theta}=1, \boldsymbol{\theta}\geq\mathbf{0}} \|\mathbf{a}_n - \mathbf{A}_{-n}\boldsymbol{\theta}\|_2,$$

where $\mathbf{A}_{-n}$ is the submatrix of $\mathbf{A}$ with $\mathbf{a}_n$ being taken away, and

$$d(\mathbf{S}) = \max_{n\in\{1,\ldots,N\},\ \ell\notin\Lambda} s_{n,\ell},$$

where $s_{n,\ell}$ denotes the (n, ℓ)th entry of $\mathbf{S}$. We show the following result:

Theorem 1: *Let $\mathbf{C}_{\text{opt}}$ be an optimal solution to Problem (6) with $q = \infty$ and $p \in (0, 1]$, and let $\mathbf{c}_{\text{opt}}^\ell$ denote its ℓth row. Assume that $\lambda \geq 2\epsilon$ and $d(\mathbf{S}) < 1$. Then, under (A1), $\mathbf{C}_{\text{opt}}$ satisfies*

$$\|\mathbf{c}_{\text{opt}}^\ell\|_\infty > 1 - c, \quad \ell \in \Lambda$$

and

$$\|\mathbf{c}_{\text{opt}}^\ell\|_\infty \leq (N(1 - (1 - c)^p))^{1/p}, \quad \ell \notin \Lambda$$

where $c = \frac{2\epsilon+\lambda}{(1-d(\mathbf{S}))\alpha(\mathbf{A})}$.

By Theorem 1, when c is small and less than one, Λ can be easily identified by taking the indices corresponding to $\mathbf{c}_{\text{opt}}^\ell$'s such that $\|\mathbf{c}_{\text{opt}}^\ell\|_\infty \approx 1$, since the other rows are close to zero. It also says that using a small p can greatly suppress $\|\mathbf{c}_{\text{opt}}^\ell\|_\infty$ for $\ell \notin \Lambda$, when c and N are reasonably small–and this allows us to identify rows with 'significant' norms much easier. We also see that the result in Theorem 1 is consistent with the physical interpretations of $\alpha(\mathbf{A})$ and $d(\mathbf{S})$. Simply speaking, if $\alpha(\mathbf{A})$ is large, the columns of $\mathbf{A}$ are far away from each other on the affine set spanned by $\mathbf{a}_1, \ldots, \mathbf{a}_N$, resulting in a 'well-conditioned' $\mathbf{A}$. In addition, if $d(\mathbf{S})$ is small, it means that $\mathbf{s}_\ell$'s for all $\ell \notin \Lambda$ are far away from the unit vectors so that it is easy to distinguish $\mathbf{a}_n$ (or, $\mathbf{x}_\ell$ for $\ell \in \Lambda$) from $\mathbf{x}_\ell$ for $\ell \notin \Lambda$. Therefore, a large $\alpha(\mathbf{A})$ and a small $d(\mathbf{S})$ present an 'easy case' for identifying Λ.

For Problem (7), we also show that

Theorem 2: *Assume that $\mathbf{C}_{\rm opt}$ is an optimal solution to Problem (7) with $q = \infty$ and $0 < p \leq 1$, that $\alpha(\mathbf{A})$ and $d(\mathbf{S})$ are defined as before, and that $\lambda > 0$. Then, under* (`A1`), *we have*

$$\|\mathbf{c}^\ell_{\rm opt}\|_\infty > 1 - c', \quad \ell \in \Lambda$$

and

$$\left\|\mathbf{c}^\ell_{\rm opt}\right\|_\infty \leq \left(N(1-(1-c')^p) + 4\lambda(L-N)\epsilon^2\right)^{1/p}, \quad \ell \notin \Lambda,$$

where the constant $c' = \frac{\sqrt{\frac{N}{\lambda} + 4(L-N)\epsilon^2} + 2\epsilon}{(1-d(\mathbf{S}))\alpha(\mathbf{A})}$.

The result in Theorem 2 reflects the intuition of choosing λ: A small λ encourages the row-sparsity of the solution, and a large λ makes the solution concentrate more on data fitting accuracy. Another observation is that the upper bound of $\left\|\mathbf{c}^\ell_{\rm opt}\right\|_\infty$ for $\ell \notin \Lambda$ is scaled by L, which is less favorable than that in Theorem 1. Particularly, if $\left(N(1-(1-c')^p) + 4\lambda(L-N)\epsilon^2\right)^{1/p} > 1$ (which may happen when L or ϵ is large), the upper bound is very loose when p is small – and this implies that using $p < 1$ may not be helpful when noise is severe.

IV. Proof of The Theorems

A. Proof of Theorem 1

We show Theorem 1 step by step. To simplify the notations, we assume that $\Lambda = \{1, \ldots, N\}$ without loss of generality.

Step 1): For any feasible solution $\mathbf{C}$ of Problem (6), we consider $n \in \Lambda = \{1, \ldots, N\}$. We see that, by the triangle inequality,

$$\begin{aligned} \|\mathbf{x}_n - \mathbf{X}\mathbf{c}_n\|_2 &= \|\mathbf{a}_n + \mathbf{v}_n - (\mathbf{A}\mathbf{S} + \mathbf{V})\mathbf{c}_n\|_2 \\ &\geq \|\mathbf{a}_n - \mathbf{A}\mathbf{S}\mathbf{c}_n\|_2 - \|\mathbf{v}_n - \mathbf{V}\mathbf{c}_n\|_2. \end{aligned} \quad (8)$$

Notice that, also by the triangle inequality and the nonnegativity of $\mathbf{C}$, we have

$$\begin{aligned} \|\mathbf{v}_n - \mathbf{V}\mathbf{c}_n\|_2 &\leq \|\mathbf{v}_n\|_2 + \sum_{\ell=1}^{L} c_{\ell,n}\|\mathbf{v}_\ell\|_2 \\ &\leq \|\mathbf{v}_n\|_2 + \max_{\ell \in \{1,\ldots,L\}} \|\mathbf{v}_\ell\|_2 \leq 2\epsilon, \end{aligned} \quad (9)$$

where the second inequality results from the sum-to-one property of $\mathbf{c}_n$. Combining (8)–(9) and the sphere constraint in (6), we see that

$$\|\mathbf{a}_n - \mathbf{A}\mathbf{S}\mathbf{c}_n\|_2 \leq \lambda + 2\epsilon. \quad (10)$$

Step 2): Now we show that $\|\mathbf{a}_n - \mathbf{A}\mathbf{S}\mathbf{c}_n\|_2$ is also lower bounded following the insight of Lemma 17 in [14], with proper modifications. Since $s_{n,n} = 1$ for $n = 1, \ldots, N$ under the assumption of $\Lambda = \{1, \ldots, N\}$, we have

$$\begin{aligned} \mathbf{A}\mathbf{S}\mathbf{c}_n &= \mathbf{a}_n \mathbf{s}^n \mathbf{c}_n + \mathbf{A}_{-n}\mathbf{S}^{-n}\mathbf{c}_n, \\ &= \mathbf{a}_n \underbrace{\left(c_{n,n} + \sum_{\ell \neq n} s_{n,\ell}c_{\ell,n}\right)}_{\triangleq \eta} + \mathbf{A}_{-n}\underbrace{\mathbf{S}^{-n}\mathbf{c}_n}_{\triangleq \boldsymbol{\zeta}}, \end{aligned}$$

where $\mathbf{S}^{-n}$ denotes a submatrix of $\mathbf{S}$ with the nth row being taken away. It can be verified that $0 \leq \eta \leq 1$. Now, suppose that $\eta < 1$. Consider

$$\begin{aligned} \|\mathbf{a}_n - \mathbf{A}\mathbf{S}\mathbf{c}_n\|_2 &= \|\mathbf{a}_n - \mathbf{a}_n\eta - \mathbf{A}_{-n}\boldsymbol{\zeta}\|_2 \\ &= (1-\eta)\left\|\mathbf{a}_n - \mathbf{A}_{-n}\frac{\boldsymbol{\zeta}}{1-\eta}\right\|_2. \end{aligned} \quad (11)$$

Let $\theta = \frac{\zeta}{1-\eta}$, which can be shown to satisfy $\theta \geq \mathbf{0}$ and $\mathbf{1}^T\theta = 1$. Hence, we have

$$\|\mathbf{a}_n - \mathbf{A}\mathbf{S}\mathbf{c}_n\|_2 \geq (1-\eta)\alpha(\mathbf{A}) \quad (12)$$

by the definition of $\alpha(\mathbf{A})$. Now, by Eq. (10), we have

$$(1-\eta)\alpha(\mathbf{A}) \leq \lambda + 2\epsilon. \quad (13)$$

Also notice that

$$\eta = \left(c_{n,n} + \sum_{\ell \neq n} s_{n,\ell}c_{\ell,n}\right) \leq c_{n,n} + d(\mathbf{S})(1 - c_{n,n}), \quad (14)$$

where the inequality is obtained by the assumption that $s_{n,\ell} \leq d(\mathbf{S})$ for $\ell \notin \Lambda$ and the fact that $s_{n,k} = 0$ for $k \neq n$ and $k \in \Lambda$. Thus, combining (13)–(14), we see that $\|\mathbf{c}^n\|_\infty$ for $n \in \Lambda$ is lower bounded:

$$1 - \frac{2\epsilon + \lambda}{(1-d(\mathbf{S}))\alpha(\mathbf{A})} \leq c_{n,n} \leq \|\mathbf{c}^n\|_\infty, \ \forall n \in \Lambda. \quad (15)$$

The above bound is derived for the case of $\eta < 1$. For $\eta = 1$, (15) is still valid; this can be seen from (14).

Step 3): The proof of this step is based on the observation that, under $\lambda \geq 2\epsilon$, $\mathbf{C}^\star = [\mathbf{S}^T, \mathbf{0}]^T$ is a feasible solution to Problem (6) [9]. Notice that it is obvious $\|\mathbf{C}^\star\|^p_{\infty,p} = N$ for $0 < p \leq 1$. Hence, $\sum_{\ell=1}^L \|\mathbf{c}^\ell_{\rm opt}\|^p_\infty \leq N$ has to be satisfied. Consequently, for $\ell \notin \Lambda$, we have

$$\begin{aligned} &\|\mathbf{c}^\ell_{\rm opt}\|^p_\infty \leq \sum_{\ell \notin \Lambda} \|\mathbf{c}^\ell_{\rm opt}\|^p_\infty \leq N - \sum_{\ell \in \Lambda} \|\mathbf{c}^\ell_{\rm opt}\|^p_\infty \\ &\Rightarrow \|\mathbf{c}^\ell_{\rm opt}\|_\infty \leq (N(1-(1-c)^p))^{1/p}, \quad \ell \notin \Lambda. \end{aligned}$$

B. Proof of Theorem 2

Let us define $v(\mathbf{C})$ the objective value of Problem (7) for a given feasible $\mathbf{C}$. Apparently, we have $v(\mathbf{C}_{\rm opt}) \leq v(\mathbf{C}^\star)$, where $\mathbf{C}^\star$ is defined as before. Since we know $\|\mathbf{x}_\ell - \mathbf{X}\mathbf{c}^\star_\ell\|_2 \leq 2\epsilon$, it is easy to see that $v(\mathbf{C}^\star) \leq N + 4\lambda(L-N)\epsilon^2$. Hence, for any $\mathbf{C}$ satisfying $v(\mathbf{C}) \leq v(\mathbf{C}^\star)$, we get

$$\begin{aligned} N + 4\lambda(L-N)\epsilon^2 &\geq \|\mathbf{C}\|^p_{\infty,p} + \lambda\|\mathbf{X} - \mathbf{X}\mathbf{C}\|^2_F \\ &\geq \lambda\|\mathbf{X} - \mathbf{X}\mathbf{C}\|^2_F \\ &\geq \lambda\|\mathbf{x}_n - \mathbf{X}\mathbf{c}_n\|^2_2, n \in \Lambda \\ &\geq \lambda(\|\mathbf{a}_n - \mathbf{A}\mathbf{S}\mathbf{c}_n\|_2 - 2\epsilon)^2. \end{aligned}$$

Thus, for $n \in \Lambda$, we have

$$\begin{aligned} \sqrt{\frac{N}{\lambda} + 4(L-N)\epsilon^2} + 2\epsilon &\geq \|\mathbf{a}_n - \mathbf{A}\mathbf{S}\mathbf{c}_n\|_2 \\ &\geq (1-\eta)\alpha(\mathbf{A}), \end{aligned} \quad (16)$$

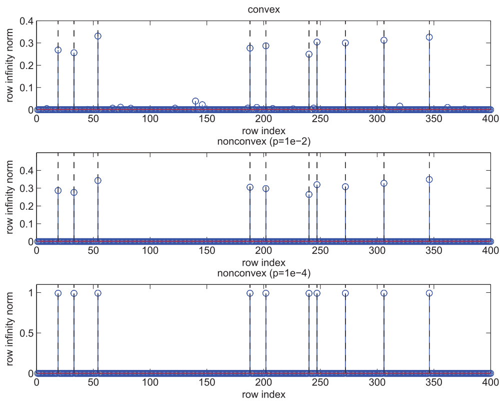

Fig. 1. $\{\|\mathbf{c}^\ell\|_\infty\}_{\ell=1}^L$ yielded by the proposed criterion with different p's. The dot lines correspond to the indices belonging to Λ.

where (16) follows from (12). Consequently, we see that

$$\|\mathbf{c}^n\|_\infty \geq c_{n,n} > 1 - \frac{\sqrt{\frac{N}{\lambda} + 4(L-N)\epsilon^2} + 2\epsilon}{(1-d(\mathbf{S}))\alpha(\mathbf{A})}, \; n \in \Lambda,$$

following the same derivation of obtaining (15) from (12). On the other hand, since $v(\mathbf{C}) \leq v(\mathbf{C}^\star)$, $\sum_{\ell\in\Lambda}\|\mathbf{c}^\ell\|_\infty^p + \sum_{\ell\notin\Lambda}\|\mathbf{c}^\ell\|_\infty^p \leq N + 4\lambda(L-N)\epsilon^2$ has to be satisfied. Rearranging the above terms leads to

$$\|\mathbf{c}_{\rm opt}^\ell\|_\infty \leq \left(N(1-(1-c')^p) + 4\lambda(L-N)\epsilon^2\right)^{1/p}, \; \ell \notin \Lambda.$$

V. Numerical Results

In this section, we provide numerical results to support our analysis. To deal with the optimization problems for $0 < p < 1$, we employ a successive convex approximation approach following the insight in [20]. Due to space limitation, we only describe the key steps concisely. For Problem (6), at each iteration, we solve a weighted l_∞/l_1 problem with the cost function being $\|\mathbf{WC}\|_{\infty,1}$ under the same constraints in (6), where $\mathbf{W} = {\rm Diag}(w_1,\ldots,w_L)$ and $w_\ell = p(\hat{\mathbf{c}}^\ell)^{p-1}$, and $\hat{\mathbf{C}}$ denotes the current solution of $\mathbf{C}$. By solving this subproblem, we calculate a new $\mathbf{W}$, and solve another weighted subproblem until some stopping criterion is satisfied. For each subproblem, we solve it by the *alternating direction method of multipliers* [21]. For Problem (7), the same iterative reweighting technique can be applied.

Fig. 1 shows an illustrative example, where the elements of $\mathbf{A} \in \mathbb{R}^{224\times 10}$ follow the uniform distribution between zero and one, and $\mathbf{s}_\ell$ for $\ell = 1,\ldots,L$ is generated following the uniform Dirichlet distribution; $\mathbf{s}_{\ell_1},\ldots,\mathbf{s}_{\ell_N}$ are then manually set to unit vectors so that (A1) is satisfied. The noise vector $\mathbf{v}_\ell$ follows the zero-mean i.i.d. Gaussian distribution with variance σ^2. The signal to noise ratio (SNR, which is defined by SNR$= \|\mathbf{A}\mathbf{s}_\ell\|_2^2/(ML\sigma^2)$) is set to 15 dB. We apply the formulation in (6) with $p = 1$, 0.01 and 0.0001 and set $\lambda = 2\epsilon$ following Theorem 1. We see that when $p = 1$, $\|\mathbf{c}^\ell\|_\infty$'s for $\ell \in \Lambda$ admit most significant values; however, some $\|\mathbf{c}^\ell\|_\infty$ for $\ell \notin \Lambda$ are visible. Using $p = 0.01$, the visible residues are successfully suppressed, which is consistent to the result in Theorem 1. By using $p = 0.0001$, the obtained result is very close to the groundtruth.

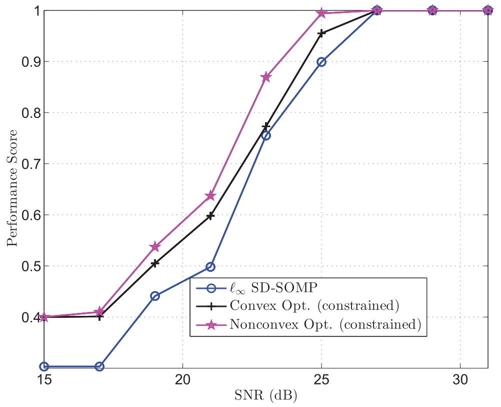

Fig. 2. The performance scores of the algorithms.

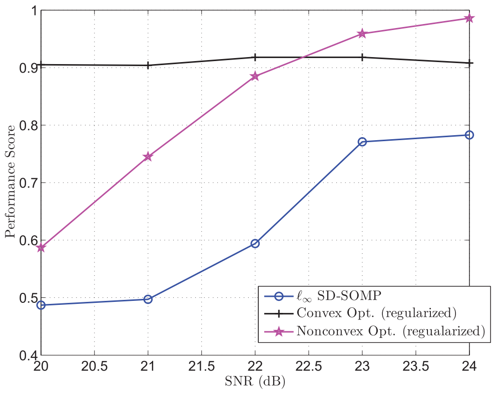

Fig. 3. The performance scores of the algorithms; $\lambda = 0.1$.

In Fig. 2, we consider the HU application, and use real hyperspectral signatures as the columns of $\mathbf{A}$ for a Monte Carlo Simulation. The signatures are selected from the U.S. Geology Survey (U.S.G.S.) library [22]. After solving Problem (6), we select a set of indices, $\hat{\Lambda}$, by taking the indices of the rows with $\|\mathbf{c}^\ell\|_\infty > 0.05$. We define Performance Score $= \frac{\max\{|\mathcal{S}_1|-|\mathcal{S}_2|,0\}}{N}$ to measure the performance, where $\mathcal{S}_1 = \Lambda\bigcap_{\hat{\Lambda}}$ and $\mathcal{S}_2 = \hat{\Lambda}\backslash\{\mathcal{S}_1\}$. The first term in the denominator counts the correctly identified indices and the second discounts the over-estimated ones. Notice that the performance score is between zero and one, and one is the best that an algorithm can obtain, indicating $\hat{\Lambda} = \Lambda$. We benchmark our algorithm against a greedy SD-MMV algorithm called ℓ_∞ SD-SOMP [9]. We set $L = 200$, $p = 1$ and p= 10^{-5} respectively, and average the result from 100 independent trials. We see that the mixed-norm approach outperforms the greedy pursuit under all the SNRs. Particularly, the one using a nonconvex surrogate exhibits the best performance. We also apply the formulation in (7) under the same problem settings, and the result is shown in Fig. 3. We see that using $p = 1$ yields better performance scores than that of using $p = 0.01$ when SNR < 23 dB. This also verifies our analysis: Using $p < 1$ may not be helpful when ϵ or L is large under this formulation.

VI. Conclusion

In this letter, we analyzed the performance of mixed-norm SD-MMV optimization for structured matrix factorization, and showed that they are provably robust to bounded noise. Our research also showed that using the nonconvex l_∞/l_p quasi-norm can lead to better results under some conditions.

References

[1] W.-K. Ma, J. Bioucas-Dias, P. Gader, T.-H. Chan, N. Gillis, A. Plaza, A. Ambikapathi, and C.-Y. Chi, "An signal processing perspective on hyperspectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, 2014.

[2] J. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 5, no. 2, pp. 354–379, 2012.

[3] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, p. 257, 2014.

[4] X. Fu, W.-K. Ma, K. Huang, and N. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May 2015.

[5] X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Tensor-based power spectra separation and emitter localization for cognitive radio," in *Proc. IEEE SAM 2014*, 2014, pp. 421–424.

[6] X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Power spectra separation via structured matrix factorization," *IEEE J. Sel. Topics Signal Process*, 2015, submitted to.

[7] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. IEEE CVPR 2012*, 2012, pp. 1600–1607.

[8] T.-H. Chan, W.-K. Ma, C.-Y. Chi, and Y. Wang, "A convex analysis framework for blind separation of non-negative sources," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5120–5134, Oct. 2008.

[9] X. Fu, W.-K. Ma, T.-H. Chan, and J. Bioucas-Dias, "Self-dictionary sparse regression for hyperspectral unmixing: Greedy pursuit and pure pixel search are related," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1128–1141, Oct. 2015.

[10] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3239–3252, Jul. 2012.

[11] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[12] X. Fu, W.-K. Ma, T.-H. Chan, J. M. Bioucas-Dias, and M.-D. Iordache, "Greedy algorithms for pure pixels identification in hyperspectral unmixing: A multiple-measurement vector viewpoint," in *Proc. EUSIPCO 2013*, 2013, pp. 1–5.

[13] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Advances in Neural Information Processing Systems*, 2012, pp. 1214–1222.

[14] N. Gillis and R. Luce, "Robust near-separable nonnegative matrix factorization using linear optimization," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1249–1280, 2014.

[15] N. Gillis, "Robustness analysis of hottopixx, a linear programming model for factoring nonnegative matrices," *SIAM J. Matrix Anal. Applicat.*, vol. 34, no. 3, pp. 1189–1212, 2013.

[16] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.

[17] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.

[18] R. Ammanouil, A. Ferrari, C. Richard, and D. Mary, "Blind and fully constrained unmixing of hyperspectral images," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5510–5518, Dec. 2014.

[19] Q. Qu, N. Nasrabadi, and T. Tran, "Subspace vertex pursuit: A fast and robust near-separable nonnegative matrix factorization method for hyperspectral unmixing," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1142–1155, 2015.

[20] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Applicat., Special Issue on Sparsity*, no. 5, pp. 877–905, Dec. 2009.

[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, pp. 1–122, 2011.

[22] R. Clark, G. Swayze, R. Wise, E. Livo, T. Hoefen, R. Kokaly, and S. Sutley, "USGS digital spectral library splib06a: U.S. geological survey, digital data series 231," in , 2007 [Online]. Available: http://speclab.cr.usgs.gov/spectral.lib06