Minimum-Volume-Constrained Nonnegative Matrix Factorization: Enhanced Ability of Learning Parts

Guoxu Zhou, Shengli Xie, *Senior Member, IEEE*, Zuyuan Yang, Jun-Mei Yang, and Zhaoshui He

***Abstract*— Nonnegative matrix factorization (NMF) with minimum-volume-constraint (MVC) is exploited in this paper. Our results show that MVC can actually improve the sparseness of the results of NMF. This sparseness is L_0-norm oriented and can give desirable results even in very weak sparseness situations, thereby leading to the significantly enhanced ability of learning parts of NMF. The close relation between NMF, sparse NMF, and the MVC_NMF is discussed first. Then two algorithms are proposed to solve the MVC_NMF model. One is called quadratic programming_MVC_NMF (QP_MVC_NMF) which is based on quadratic programming and the other is called negative glow_MVC_NMF (NG_MVC_NMF) because it uses multiplicative updates incorporating natural gradient ingeniously. The QP_MVC_NMF algorithm is quite efficient for small-scale problems and the NG_MVC_NMF algorithm is more suitable for large-scale problems. Simulations show the efficiency and validity of the proposed methods in applications of blind source separation and human face images analysis.**

***Index Terms*— Blind source separation, nonnegative matrix factorization, sparse representation.**

NOTATION

$\mathcal{I}_N$	Set of positive integers less than $N+1$.
$\mathbb{R}_+^{M\times N}$	Set of nonnegative real $M \times N$ matrices.
$\mathbf{a}_i, a_{ij}, \underline{\mathbf{a}}_i$	ith column, ij element, ith row of matrix $\mathbf{A}$.
$\overline{\mathbf{A}_i}$	Submatrix of $\mathbf{A}$ by removing the ith column of $\mathbf{A}$.
$\mathbf{1}_N$	N-dimensional vector with all the elements equal to unity.
$\mathcal{P}_+ = \{\mathbf{P}_+\}$	Set of generalized permutation matrices $\mathbf{P}_+$, where $\mathbf{P}_+ = \mathbf{PD}$, $\mathbf{P}\in \mathbb{R}_+^{r\times r}$ is a permutation matrix, $\mathbf{D}\in \mathbb{R}_+^{r\times r}$ is an invertible diagonal matrix.
$\mathbf{I}$	Identity matrix with proper dimension.
$\|\cdot\|_F$	Frobenius norm of a matrix.
$\succeq, \succ$	Elementwise inequality.
$\odot, \frac{\mathbf{A}}{\mathbf{B}}$	Elementwise multiplication and division of matrices.

Manuscript received May 18, 2010; revised August 2, 2011; accepted August 4, 2011. Date of publication August 30, 2011; date of current version October 5, 2011. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2010CB731800 and by the National Natural Science Foundation of China under Grant U0835003, Grant 60874061, Grant 60974072, and Guangdong Natural Science Foundation under Grant S2011040005724.

G. Zhou, Z. Yang, and J. Yang are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: zhou.guoxu@mail.scut.edu.cn; yangzuyuan@yahoo.com.cn; yjunmei@scut.edu.cn).

S. Xie and Z. He are with the Faculty of Automation, Guangdong University of Technology, Guangzhou 510641, China (e-mail: eeoshlxie@scut.edu.cn; zhshhe@gdut.edu.cn).

Digital Object Identifier 10.1109/TNN.2011.2164621

I. INTRODUCTION

A FUNDAMENTAL problem in many data-analysis tasks is to find an informative decomposition of a given data matrix $\mathbf{V}$ such that

$$\mathbf{V} = \mathbf{WH}. \tag{1}$$

Nonnegative matrix factorization (NMF) is such a problem, where the involved matrices $\mathbf{V}$ (the observation matrix), $\mathbf{W}$ (the basis matrix), and $\mathbf{H}$ (the encoding matrix) are all nonnegative [1]. Because of the nonnegativity constraints, NMF is purely based on additions and thus ready to give parts-based representation of the objects [1], [2]. Consequently, NMF has attracted much attention in recent years [1]–[8].

To obtain desirable results, the uniqueness of NMF should be considered seriously. This issue has been presented by Donoho and Stodden as the following two questions [9].

1) Under what assumptions is the notion of nonnegative matrix factorization well defined, for example, is the factorization in some sense unique.
2) Under what assumptions is the factorization correct, recovering the "right answer."

In this paper, Donoho and Stodden developed a group of uniqueness conditions. These uniqueness conditions relate to both of the basis matrix and the encoding matrix, which limits their practical applications more or less.

It has been found that NMF with sparseness constraints may give unique factorization [2], [10]–[14]. Theis *et al.* proved that the sparse NMF method in [2] generates unique results when there are two observations and two sources with the same sparseness level [14]. Stadlthanner *et al.* obtained the uniqueness results while applying sparse NMF to microarray data [12]. Laurberg and Hansen proposed the affine sparse NMF to improve the uniqueness properties of the factorization [15]. Although the exploration on uniqueness of NMF is continuing, sparse NMF is the most likely way to generate unique solutions, and hence it has attracted increasing attention.

Another reason why sparse NMF should be especially concerned relates to the ability of the learning parts of NMF. If the sparseness levels of the factors are improved, the ability of the learning parts can be enhanced. In this sense, sparseness is rather an indispensable constraint to NMF instead of an

optional one. Because of these two major reasons, this paper is focused on sparse NMF.

A natural measure of sparseness is L_0-norm. However, the involved optimization has proved to be NP-hard (i.e., nondeterministic polynomial-time hard) and intractable. Hence, L_1-norm is frequently used instead to impose sparseness constraints [16], [17]. Although L_1-norm-based sparseness measure has achieved great success, it has two major disadvantages: first, it implicitly requires that the target signals are of high sparseness levels [18], as otherwise it cannot measure the sparseness properly, the other one is that the L_1-norm-based measure is simply ineffective in some situations. For example, in remote-sensing image analysis, the L_1-norms of the target signals (i.e., the abundances) are constant [19], [20]. Hence, it is necessary to design new sparseness measures.

In this paper, the minimum volume is suggested to impose the sparseness constraint on NMF. Minimum-volume-constrained NMF (MVC_NMF) was recently proposed and has shown its great competence in many applications such as dependent nonnegative source separation, remote sensing data interpretation, etc. [4], [19], [21], [22]. Although these methods have achieved great success, they have limitations. In theoretical respects, they fail to investigate the relation between the MVC and sparseness. The volume is simply interpreted as the *volume* of an r-parallelogram in [21], or as the correlation of sources in [4]. However, this is not enough. In algorithmic aspects, these methods are hard to guarantee the nonnegativity of results [4], [21]. Moreover, they may suffer from the disadvantage of high computational complexity [4], [19], [23], as shown in Section III-C. This paper aims to overcome these drawbacks.

In this paper, the significance of sparseness constraint to NMF is addressed first. Then we show that MVC_NMF can improve the sparseness of results even if the sparseness of the hiding encoding matrix is very weak, thereby leading to enhanced ability of the learning parts of NMF. This suggests that MVC_NMF can be regularly used to refine the results obtained by ordinary NMF methods. Then, two algorithms, named as quadratic programming_MVC_NMF (QP_MVC_NMF) and negative glow_MVC_NMF (NG_MVC_NMF) respectively, are proposed to optimize the MVC_NMF model. The former is based on a series of quadratic programming, which is efficient in solving small-scale problems, and the latter, using the multiplicative updates incorporating the natural gradient, is more suitable for large-scale problems. It can be seen that these two algorithms are quite competent in terms of efficiency and accuracy of factorization compared with the existing MVC_NMF methods.

Blind source separation (BSS) is a problem of recovering the sources from their mixtures when the mixing parameters are unknown [24]–[28]. If the encoding matrix $\mathbf{H}$ in NMF models the source signals and the basis matrix $\mathbf{W}$ models the mixing parameters, NMF with the uniqueness guarantee can be applied to BSS directly. A major advantage of NMF-based BSS methods is that the sources do not need to be mutually independent [4]. Thus, NMF with the uniqueness guarantee is a potential way to separate statistically dependent sources, which is often intractable using the traditional BSS methods. Since sparseness constraint often yields unique factorization, it is reasonable to apply sparse NMF to BSS. Compared with the traditional sparse BSS methods, the proposed methods can obtain desirable results even if the sources are of weak sparseness.

It is still fraught with difficulties by far to explore more relaxed conditions for uniqueness of NMF. As BSS requires the solutions to be unique implicitly, applying NMF to BSS can help us explore new uniqueness conditions. If NMF can always separate a kind of sources experimentally, by analyzing the features of the sources we may find some new uniqueness conditions and then turn to the theoretical proof.

The rest of this paper is organized as follows. In Section II, the general NMF model and the MVC model are introduced, and their relations to the sparseness are discussed as well. Section III is devoted to the development of the new algorithms. Numerical simulations are presented in Section IV, and finally, the conclusions are made in Section V.

II. Sparse NMF

Consider the following widely used cost function in NMF [1], [29]:

$$\begin{aligned} &\min: J(W,H) = \tfrac{1}{2}\|V - WH\|_{\mathrm{F}}^2 \\ &\text{s.t. } \mathbf{W} \in \mathbb{R}_+^{M\times r},\ \mathbf{H} \in \mathbb{R}_+^{r\times N} \end{aligned} \tag{2}$$

where $V \in \mathbb{R}_+^{M\times N}$, $\mathbf{W}$, and $\mathbf{H}$ denote the nonnegative observation matrix, the encoding matrix, and the basis matrix, respectively. For the convenience of presentation, the optimal solution to (2) such that $\mathbf{V} = \mathbf{WH}$ is denoted as

$$(\mathbf{W},\mathbf{H}) = \mathrm{NMF}(\mathbf{V}, r). \tag{3}$$

The following assumptions are made:

A1) $r \le M \ll N$ and $\mathbf{W}$ is of full column rank;
A2) $\mathbf{1}^{\mathrm{T}}\mathbf{w}_i = 1,\ i \in \mathcal{I}_r$.

Assumption A1) is common in NMF and A2) is used to eliminate the norm indeterminacy. Generally, if $(\mathbf{W},\mathbf{H})$ is a solution to (2), $(\mathbf{WP}_+^{-1}, \mathbf{P}_+\mathbf{H})$ is also a solution to (2), i.e., the solution is always non-unique. Nevertheless, this kind of non-uniqueness does not affect the results essentially and thus it is not in our consideration. In the following, we first introduce the concept of factorability, which is moderately weaker than uniqueness.

Definition 1: $\mathbf{H} \in \mathbb{R}_+^{r\times N}$ is factorable in the sense of NMF if and only if there exists a matrix $\mathbf{W} \notin \mathcal{P}_+$ such that $(\mathbf{W},\mathbf{H}') = \mathrm{NMF}(\mathbf{H}, r)$. Otherwise, $\mathbf{H}$ is called unfactorable.

Obviously, unfactorability is a necessary condition for unique factorization.

A. NMF and Sparseness

NMF has received increasing attention because of its ability to learn the parts of objects. The term "parts" means that only partial entries in the output matrix $\mathbf{W}$ and/or $\mathbf{H}$ are active (i.e., having nonzero values) while the others are

inactive (having zero values). This property is referred to as sparseness. For simplicity, the following sparseness measure is defined:

$$\mathcal{S}(\mathbf{H}) = 1 - \frac{\|\mathbf{H}\|_0}{rN} \quad \text{for} \quad \mathbf{H} \in \mathbb{R}_+^{r\times N} \tag{4}$$

where $\|\mathbf{H}\|_0$ is the number of nonzero entries in $\mathbf{H}$. Here, $\mathcal{S}(\mathbf{H})$ measures the sparseness level of $\mathbf{H}$ and satisfies:

1) $0 \le \mathcal{S}(\mathbf{H}) \le 1$;
2) $\mathcal{S}(\mathbf{H}) = 0$ if and only if $h_{ij} > 0$ for all $i \in \mathcal{I}_r$, $j \in \mathcal{I}_N$;
3) $\mathcal{S}(\mathbf{H}) = \mathcal{S}(\mathbf{P}_1\mathbf{H}) = \mathcal{S}(\mathbf{H}\mathbf{P}_2)$ for any $\mathbf{P}_1,\ \mathbf{P}_2 \in \mathcal{P}_+$.

With these preliminaries, we have the following proposition.

Proposition 1: Let $(\mathbf{W}, \mathbf{H}) = \mathrm{NMF}(\mathbf{V}, r)$, where $\mathbf{V} \in \mathbb{R}_+^{r\times N}$, then $\mathcal{S}(\mathbf{H}) \ge \mathcal{S}(\mathbf{V})$.

Proof: If $\mathbf{W} \in \mathcal{P}_+$, $\mathcal{S}(\mathbf{H}) = \mathcal{S}(\mathbf{V})$. Otherwise, assume that $w_{ii} > 0$ (possibly by properly re-permuting the columns of $\mathbf{W}$). From the fact that $\underline{\mathbf{v}}_i = \sum_{l=1}^{r} w_{il}\underline{\mathbf{h}}_l$ and $w_{il} \ge 0$, $\underline{\mathbf{h}}_l \ge 0$, we have $\|\underline{\mathbf{v}}_i\|_0 \ge \|\underline{\mathbf{h}}_i\|_0$ and hence $\|\mathbf{V}\|_0 \ge \|\mathbf{H}\|_0$, i.e., $\mathcal{S}(\mathbf{H}) \ge \mathcal{S}(\mathbf{V})$.

Proposition 1 shows that NMF leads to sparse factors naturally. To further increase the sparseness of the results, many researchers have considered NMF models with explicit sparseness constraints [10], [12], [13]

$$\begin{aligned} &\min: J = \tfrac{1}{2}\|\mathbf{V}-\mathbf{W}\mathbf{H}\|_{\mathrm{F}}^2 + \alpha_{\mathbf{W}} J_{\mathbf{W}}(\mathbf{W}) + \alpha_{\mathbf{H}} J_{\mathbf{H}}(\mathbf{H}) \\ &\text{s.t. } \mathbf{W} \in \mathbb{R}_+^{M\times r}, \mathbf{H} \in \mathbb{R}_+^{r\times N} \end{aligned} \tag{5}$$

where $\alpha_{\mathbf{W}}, \alpha_{\mathbf{H}} \ge 0$ balance the fitting error and the sparseness level, and $J_{\mathbf{W}}$, $J_{\mathbf{H}}$ are sparseness measure functions. Generally, it is unrealistic to require both the basis matrix $\mathbf{W}$ and the encoding matrix $\mathbf{H}$ to be sparse because of the following reasons. A general NMF can be further factorized as

$$\mathbf{V} = \mathbf{W}\mathbf{H} = \mathbf{W}_1\mathbf{G}\mathbf{H}_1 \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}_+^{M\times r}$, $\mathbf{H}_1 \in \mathbb{R}_+^{r\times N}$ are unfactorable and as sparse as possible, $\mathbf{G} \in \mathbb{R}_+^{r\times r}$. If $\mathbf{G} \in \mathcal{P}_+$, let $\mathbf{W} = \mathbf{W}_1$ and $\mathbf{H} = \mathbf{G}\mathbf{H}_1$, then both the basis matrix and the encoding matrix are the sparsest. Otherwise, if $\mathbf{G} \notin \mathcal{P}_+$, from Proposition 1, $\mathcal{S}(\mathbf{W}_1\mathbf{G}) \le \mathcal{S}(\mathbf{W}_1)$ and $\mathcal{S}(\mathbf{G}\mathbf{H}_1) \le \mathcal{S}(\mathbf{H}_1)$ hold. Consequently, we either accept $\mathbf{V} = \mathbf{W}_1(\mathbf{G}\mathbf{H}_1)$, where the basis matrix $\mathbf{W} = \mathbf{W}_1$ is the sparsest while the encoding matrix $\mathbf{H} = \mathbf{G}\mathbf{H}_1$ is less sparse, or else, accept $\mathbf{V} = (\mathbf{W}_1\mathbf{G})\mathbf{H}_1$, which is just the contrary. In this paper, we concentrate on the sparseness of the encoding matrix $\mathbf{H}$. Nevertheless, the proposed method can also be used to seek the sparsest basis matrix $\mathbf{W}$ as follows. Let $(\mathbf{W}, \mathbf{H}) = s_b\mathrm{NMF}(\mathbf{V}, r)$ denote a sparse NMF algorithm where the sparsest encoding matrix $\mathbf{H}$ is guaranteed. For the given observation matrix $\mathbf{V}$, the sparsest basis matrix$\mathbf{W}_1$ can be obtained by implementing $(\mathbf{W}, \mathbf{H}_1) = s_b\mathrm{NMF}(\mathbf{V}, r)$ and $(\mathbf{G}^{\mathrm{T}}, \mathbf{W}_1^T) = s_b\mathrm{NMF}(\mathbf{W}^{\mathrm{T}}, r)$ orderly.

In the following, the close relation between the sparseness and the uniqueness of NMF is investigated.

Proposition 2: Given $(\mathbf{W}, \mathbf{H}) = \mathrm{NMF}(\mathbf{V}, r)$, if there exists $i \in \mathcal{I}_r$ such that $\underline{\mathbf{h}}_i \succ 0$, there must exist another factorization $(\mathbf{W}_1, \mathbf{H}_1) = \mathrm{NMF}(\mathbf{V}, r)$ such that $\mathcal{S}(\mathbf{H}) < \mathcal{S}(\mathbf{H}_1)$.

The proof is somewhat straightforward and is omitted here. Proposition 2 shows that sparseness is necessary to unfactorability and uniqueness of NMF.

Proposition 3: For any $\mathbf{H} \in \mathbb{R}_+^{r\times N}$, let $\mathbf{H}_i^0 \in \mathbb{R}_+^{r\times L}$ be the matrix whose columns consist of specified columns of $\mathbf{H}$, e.g., $\mathbf{h}_t$, where the index t is chosen such that $h_{it} = 0$. If the rank of $\mathbf{H}_i^0$ is $r-1$, for any $i \in \mathcal{I}_r$, $\mathbf{H}$ is unfactorable.

The proof can be found in Appendix A. Proposition 3 gives a sufficient condition for unfactorability: if $\mathbf{H}$ is *sufficiently* sparse, it can not be further factorized.

From Proposition 3, we have the following corollary.

Corollary 1: If $\mathbf{H} \in \mathbb{R}_+^{r\times N}$ satisfies the pure sources assumption, $\mathbf{H}$ is unfactorable.

Here, the pure sources assumption means that, for any $i \in \mathcal{I}_r$, there exists at least one index t such that h_{it} is the only nonzero entry in this column [4]. This assumption is very popular in the uniqueness analysis of NMF.

Proposition 4: Let $(\mathbf{W}, \mathbf{H}) = \mathrm{NMF}(\mathbf{V}, r)$, where $\mathbf{V} \in \mathbb{R}_+^{r\times N}$. If $\mathbf{H}$ satisfies the pure sources assumption, then $\mathcal{S}(\widehat{\mathbf{H}}) \le \mathcal{S}(\mathbf{H})$ for any $(\widehat{\mathbf{W}}, \widehat{\mathbf{H}}) = \mathrm{NMF}(\mathbf{V}, r)$. Moreover, the equality holds if and only if $\mathbf{W} = \widehat{\mathbf{W}}\mathbf{P}$, where $\mathbf{P} \in \mathcal{P}_+$.

The proof is presented in Appendix B. It can be seen from Proposition 4 that the sparsest solution is essentially unique if the encoding matrix satisfies the pure sources assumption. [Note that even if $\mathbf{H}$ satisfies the pure sources assumption, the solution to (2) can be non-unique because $\mathbf{W}$ can be factorable, see (6). But the sparsest solution is unique.]

B. MVC NMF

Without loss of generality, we assume that there are no zero columns in the observation matrix $\mathbf{V}$. Then we perform the following transformation. Let $\tilde{\mathbf{V}} = \mathbf{V}\mathbf{D}$ and $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{D}$, where $D^{N\times N}$ is the diagonal matrix whose diagonal entries are $d_{tt} = 1/\sum_{i=1}^{M} v_{it}$, $t \in \mathcal{I}_N$. Thus we have

$$\widetilde{\mathbf{V}} = \mathbf{W}\widetilde{\mathbf{H}}. \tag{7}$$

It can be verified that $1 = \mathbf{1}_M^{\mathrm{T}}\widetilde{\mathbf{v}}_t = \mathbf{1}_M^{\mathrm{T}}\mathbf{W}\widetilde{\mathbf{h}}_t = \mathbf{1}_r^{\mathrm{T}}\widetilde{\mathbf{h}}_t$. Since this transformation does not change the original problem essentially, we still use $\mathbf{V}$ and $\mathbf{H}$ instead of $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{H}}$ hereafter, but with an additional assumption:

A3) $\mathbf{1}^{\mathrm{T}}\mathbf{h}_t = 1,\ t \in \mathcal{I}_N$.

Assumption A3) makes the proposed methods suitable for the remote-sensing image interpretation [20], [30]. Particularly, it provides a new interpretation of NMF. From the fact that $\mathbf{v}_t = \sum_{i=1}^{r} \mathbf{w}_i h_{it}$, $\sum_i h_{it} = 1$, and $h_{it} \ge 0$, each observation vector $\mathbf{v}_t$ is a convex combination of $\mathbf{w}_i$, and h_{it} are the associated coefficients. Particularly, when $M = r$, the observation points are totally enclosed by a simplex $\Omega(\mathbf{W})$ whose vertices are defined by the columns of $\mathbf{W}$ in $(M-1)$-dimensional space (i.e., on the hyperplane $\sum_{i=1}^{M} x_i = 1$ in M-dimensional space) [19], [20]. Consequently, the following MVC_NMF model is proposed:

$$\begin{aligned} &\min: \mathrm{Vol}(\Omega) \\ &\text{s.t. } \mathbf{V} = \mathbf{W}\mathbf{H},\ \mathbf{W} \in \mathbb{R}_+^{M\times r}, \mathbf{H} \in \mathbb{R}_+^{r\times N} \end{aligned} \tag{8}$$

where $\mathrm{Vol}(\Omega)$ is the volume of the simplex $\Omega(\mathbf{W})$. Note that the volume of $\Omega(\mathbf{W})$ is $1/(r-1)!\,|\det \mathbf{W}|$, which suggests

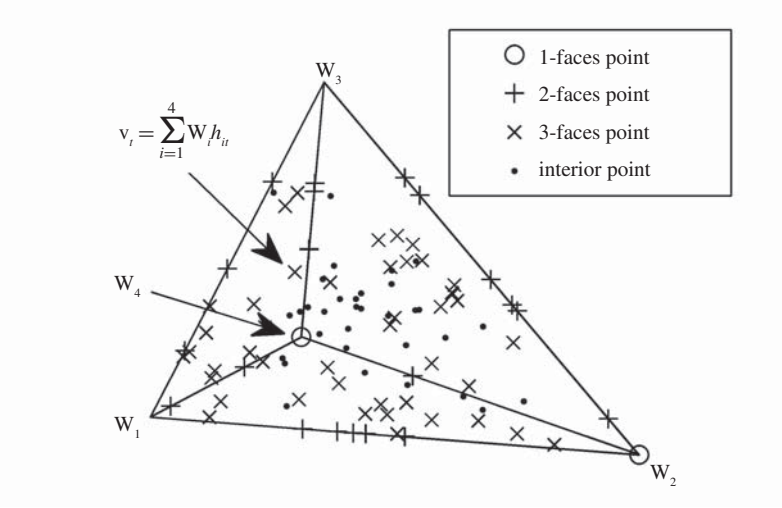

Fig. 1. Illustration of sparseness of H with $M = r = 4$ in MVC NMF.

that Vol (Ω) can be replaced by $|\det \mathbf{W}|$. However, $|\det \mathbf{W}|$ is valid only for $M = r$. To accommodate the case of $r < M$, some methods use the dimensionality reduction techniques in nonnegative space at first [4], [22], which is very inconvenient in practice. In this paper, the Gram determinant $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$ is used to replace the volume in (8). Geometrically, $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$ denotes the volume of an r-parallelogram defined by the columns of $\mathbf{W}$ [31].

Fig. 1 illustrates how model (8) actually imposes sparseness constraint on $\mathbf{H}$ for $M = r = 4$. In fact, when the volume of the simplex enclosing all the observation vectors is minimized, some observation vectors must be on the boundary (i.e., faces, edges, vertices) of the simplex, and thus the corresponding coefficient vector $\mathbf{h}_t$ is sparse, which is consistent with Proposition 2. Compared with the L_1-norm-based sparseness constraint, the minimum-volume-based sparseness constraint has a quite important advantage, i.e., it is able to measure different levels of sparseness uniformly. Generally, for a column vector of $\mathbf{H} \in \mathbb{R}_+^{r\times N}$, say $\mathbf{h}_t$, if there are total $(r–k)$ zero entries in $\mathbf{h}_t$, the corresponding observation vector $\mathbf{v}_t$ just lies in a k-face of the simplex $\Omega(\mathbf{W})$ (thus $\mathbf{v}_t$ is called a k-faces point hereafter). Particularly, the 1-faces points coincide with the vertices of $\Omega(\mathbf{W})$, and the 2-faces points are situated on the edges, and the $(r-1)$-faces points lie in the facets. If there are no zero entries in $\mathbf{h}_t$, $\mathbf{v}_t$ is called a interior point since it is in the interior of the simplex $\Omega(\mathbf{W})$ completely. This phenomenon is illustrated clearly by Fig. 1, where the coefficient vectors $\mathbf{h}_t$ are as sparse as possible, and at the same time they are of different levels of sparseness. This sparseness relates to L_0-norm directly. It is worth noticing that there can be many interior points. These points affect the L_1-norm radically but do not affect the volume of the simplex. Hoyer also proposed an NMF method with sparseness constraints (NMFSC) [2]. In NMFSC, all the columns of $\mathbf{H}$ (and/or $\mathbf{W}$) are of the same sparseness specified by the user. Consequently, all the coefficient vectors $\mathbf{h}_t$ are required to be $k-$faces points. This restricts the applications of NMFSC. In summary, the MVC is more suitable to impose sparseness on $\mathbf{H}$.

The following augmented Lagrangian function is introduced to solve model (8) (ignoring nonnegativity constraints):

$$L = \det\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right) + trace\left(\Lambda^{\mathrm{T}}(\mathbf{V}-\mathbf{WH})\right) + \frac{c_k}{2}\|\mathbf{V}–\mathbf{WH}\|_{\mathrm{F}}^2. \tag{9}$$

There are two mechanisms to guarantee the convergence: 1) by taking Λ close to the Lagrangian multipliers Λ^*, 2) by taking c_k very large and $0 < c_k < c_{k+1}$ for all k, and $c_k \to +\infty$ [32]. To be consistent with (5), we simply set $\Lambda = \mathbf{0}$ and consider the following MVC NMF model:

$$\begin{aligned} &\min: J = \tfrac{1}{2}\|\mathbf{V}–\mathbf{WH}\|_{\mathrm{F}}^2 + \tfrac{\alpha}{2}\det\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right) \\ &\text{s.t. } \mathbf{W} \in \mathbb{R}_+^{M\times r}, \mathbf{H} \in \mathbb{R}_+^{r\times N} \end{aligned} \tag{10}$$

where the penalty parameter α satisfies $0 \le \alpha_{k+1} < \alpha_k$, and k denotes the iteration number in optimization.

About the uniqueness of (8), we have the following.

Lemma 1: If $\mathbf{H}$ satisfies the pure sources assumption, the optimal solution of model (8) is unique and the sparsest.

The proof of uniqueness is similar to [4, Lemma 1], except that the authors therein use the inverse matrix of $\mathbf{W}$. From Proposition 4, we know that the unique solution is also the sparsest one. Hence, the proof on sparseness is also complete. This Lemma shows that (8) can give the sparsest and unique solution to (2) under pure source assumption. However, to obtain this sparsest and unique solution, the corresponding optimization algorithm is required to be of global convergence, which is often a big challenge in practice. More generally, we have the following proposition.

Proposition 5: Let $(\mathbf{W}, \mathbf{H})$ be a local minimum of (8) or (10). Then, $\mathcal{S}\left(\underline{\mathbf{h}}_i\right) > 0$, $\forall i \in \mathcal{I}_r$ must hold.

The proof of Proposition 5 can be found in Appendix C. Proposition 5 shows an important feature of MVC_NMF that its suboptimal solutions still have sparseness. (From Fig. 1, when the enclosing simplex is minimized, intuitively, there are at least $r-1$ points on each facet. Thus we conjecture that Proposition 5 can be further improved to be somewhat parallel with Proposition 3. It is currently an open problem.)

There is another interesting interpretation of MVC_NMF from (8). Note that $0 < \det(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \le 1$ holds under the assumptions $\mathbf{1}^{\mathrm{T}}\mathbf{w}_i = 1$, $\mathbf{w}_i 0$, $i \in \mathcal{I}_r$ (Readers can refer to Appendix B in [4] for a similar proof). Let $\mathbf{W} = \mathbf{I}$ and $\mathbf{H} = \mathbf{V}$, then we have $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W}) = 1$ and the volume of the simplex $\Omega(\mathbf{W})$ is maximized. In this situation, $\mathbf{W}$ is the sparsest while $\mathbf{H}$ is the most unsparse. On the contrary, when $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$, i.e., the volume of $\Omega(\mathbf{W})$, is minimized, $\mathbf{H}$ is the sparsest and $\mathbf{W}$ is on the contrary. Therefore, from $\mathbf{V} = \mathbf{IV} = \mathbf{WH}$ we know, NMF starts with the sparsest basis matrix and goes toward the sparsest encoding matrix.

Remark 1: The matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ captures all the geometric information about the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$, i.e., the length of the vectors and the angles between them. And, this is the only information it contains [33]. Thus it is reasonable that the authors in [4] use $\det \mathbf{W}$ to measure the correlations of the sources. Our analysis shows that the MVC can actually impose sparseness constraint on $\mathbf{H}$.

III. Algorithms

In this section we focus on the optimization of (10).

A. QP_MVC_NMF Algorithm

Motivated by $\mathbf{WH} = \sum_{i=1}^{r} \mathbf{w}_i \underline{\mathbf{h}}_i$, we will optimize (10) with respect to $\mathbf{w}_i$ and $\underline{\mathbf{h}}_i$ alternately. For the first term, it holds that

$$\begin{aligned} \|\mathbf{V} - \mathbf{WH}\|_{\mathrm{F}}^2 &= \left\| \mathbf{V} - \sum_{k \neq i} \mathbf{w}_k \underline{\mathbf{h}}_k - \mathbf{w}_i \underline{\mathbf{h}}_i \right\|_{\mathrm{F}}^2 \\ &= \left\| \mathbf{V}_i - \mathbf{w}_i \underline{\mathbf{h}}_i \right\|_{\mathrm{F}}^2 \\ &= \left(\mathbf{w}_i^{\mathrm{T}} \mathbf{w}_i \right) \left(\underline{\mathbf{h}}_i \underline{\mathbf{h}}_i^{\mathrm{T}} \right) - 2 \underline{\mathbf{h}}_i \mathbf{V}_i^{\mathrm{T}} \mathbf{w}_i + c \end{aligned} \quad (11)$$

where $\mathbf{V}_i = \mathbf{V} - \sum_{k \neq i} \mathbf{w}_k \underline{\mathbf{h}}_k$, and $c = \mathrm{trace}(\mathbf{V}_i^{\mathrm{T}} \mathbf{V}_i)$ is a constant irrelevant to $\mathbf{w}_i$ and $\underline{\mathbf{h}}_i$.

Regarding $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$ in the second term, we have:

Proposition 6:

1) $\det\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right) = \gamma\, \mathbf{w}_i^{\mathrm{T}} \left[\mathbf{I} - \overline{\mathbf{W}_i} \left(\overline{\mathbf{W}_i}^{\mathrm{T}} \overline{\mathbf{W}_i} \right)^{-1} \overline{\mathbf{W}_i}^{\mathrm{T}} \right] \mathbf{w}_i$

where $\gamma = \det\left(\overline{\mathbf{W}_i}^{\mathrm{T}} \overline{\mathbf{W}_i} \right)$ is irrelevant to $\mathbf{w}_i$;

2) $\mathbf{w}_i^{\mathrm{T}} \left[\mathbf{I} - \overline{\mathbf{W}_i} \left(\overline{\mathbf{W}_i}^{\mathrm{T}} \overline{\mathbf{W}_i} \right)^{-1} \overline{\mathbf{W}_i}^{\mathrm{T}} \right] \mathbf{w}_i = \mathbf{w}_i^{\mathrm{T}} \mathbf{C}_i \mathbf{C}_i^{\mathrm{T}} \mathbf{w}_i$

where $\mathbf{C}_i = \mathrm{Null}(\overline{\mathbf{W}_i}^{\mathrm{T}})$ is an orthonormal basis for the null space of $\overline{\mathbf{W}_i}$.

The proof is presented in Appendix D. Note that $\mathbf{C}_i$ can be computed via singular value decomposition. Proposition 6 suggests that $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$ can be calculated from $\det(\overline{\mathbf{W}_i}^{\mathrm{T}} \overline{\mathbf{W}_i}) \mathbf{w}_i^{\mathrm{T}} \mathbf{C}_i \mathbf{C}_i^{\mathrm{T}} \mathbf{w}_i$ efficiently, thus the computation of the inverse matrix of $\overline{\mathbf{W}_i}^{\mathrm{T}} \overline{\mathbf{W}_i}$ is avoided.

From Proposition 6 and (11), optimizing (10) with respect to $\mathbf{w}_i$ is equivalent to solving the following problem:

$$\begin{aligned} &\text{(P1):} \quad \min\ J(\mathbf{w}_i) = \mathbf{w}_i^{\mathrm{T}} \mathbf{Q}_i \mathbf{w}_i + \mathbf{f}_i^{\mathrm{T}} \mathbf{w}_i \\ &\text{s.t. } \mathbf{w}_i 0, \mathbf{1}^{\mathrm{T}} \mathbf{w}_i = 1 \end{aligned}$$

where $\mathbf{Q}_i = 1/2(\underline{\mathbf{h}}_i \underline{\mathbf{h}}_i^{\mathrm{T}} \mathbf{I} + \alpha \gamma\, \mathbf{C}_i \mathbf{C}_i^{\mathrm{T}})$ is positive definite, and $\mathbf{f}_i = -\mathbf{V}_i \underline{\mathbf{h}}_i^{\mathrm{T}}$. Thus (P1) can be solved efficiently by quadratic programming algorithms, and the solution is unique. For simplicity, the solution of (P1) is denoted by $\mathbf{w}_i = \mathrm{QP}(\mathbf{Q}_i, \mathbf{f}_i, *)$.

As for the optimization with respect to $\underline{\mathbf{h}}_i$, only the fitting error needs to be considered. From (11) again, $\underline{\mathbf{h}}_i$ is obtained by solving the following model:

$$\begin{aligned} &\text{(P2):} \quad \min\ J\left(\underline{\mathbf{h}}_i\right) = \mathbf{w}_i^{\mathrm{T}} \mathbf{w}_i \underline{\mathbf{h}}_i \underline{\mathbf{h}}_i^{\mathrm{T}} - 2 \mathbf{w}_i^{\mathrm{T}} \mathbf{V}_i \underline{\mathbf{h}}_i^{\mathrm{T}} \\ &\text{s.t.} \qquad \geq \underline{\mathbf{h}}_i 0. \end{aligned} \quad (12)$$

The Lagrangian function of (12) is

$$L = J\left(\underline{\mathbf{h}}_i\right) - \sum_{t=1}^{N} \lambda_t h_{it}. \quad (13)$$

By setting $(\partial L / h_{it}) = 0$, we have

$$\underline{\mathbf{h}}_i = \frac{1}{\mathbf{w}_i^{\mathrm{T}} \mathbf{w}_i} \left(\mathbf{w}_i^{\mathrm{T}} \mathbf{V}_i + \underline{\lambda} \right) \quad (14)$$

where $\underline{\lambda}$ is a row vector whose entries consist of the Lagrange multipliers λ_t. Let $\mathbf{x} = \mathbf{w}_i^{\mathrm{T}} \mathbf{V}_i$. From the Karush–Kuhn–Tucker necessary conditions [32], we have $\underline{\mathbf{h}}_i \geq 0$, $\underline{\lambda} \geq 0$, and $h_{it} \lambda_t = 0$. Thus if $x_i \geq 0$, let $\lambda_i = 0$, then $h_{it} > 0$, otherwise, if $x_i < 0$, let $\lambda_i = -x_i > 0$, then $h_{it} = 0$. In other words

$$\underline{\mathbf{h}}_i = \max\left\{ \frac{\mathbf{w}_i^{\mathrm{T}} \mathbf{V}_i}{\mathbf{w}_i^{\mathrm{T}} \mathbf{w}_i}, 0 \right\}. \quad (15)$$

It can be seen that (15) gives the unique and optimal solution to (12). Since $\mathbf{w}_i$ and $\underline{\mathbf{h}}_i$ are updated by a series of quadratic programming, this algorithm is called QP_MVC_NMF and summarized as follows.

Algorithm 1 QP_MVC_NMF Algorithm

Initialization: $\mathbf{V} = \mathbf{VD}$, where $D^{N \times N}$ is a diagonal matrix whose diagonal elements are $d_{tt} = 1/\sum_{i=1}^{M} v_{it}$, $t \in \mathcal{I}_N$;

While it does not converge

  For i = 1 to r

  $\mathbf{C}_i = \mathrm{Null}\left(\overline{\mathbf{W}_i}^{\mathrm{T}}\right)$, $\mathbf{Q}_i = \frac{1}{2}\left(\underline{\mathbf{h}}_i \underline{\mathbf{h}}_i^{\mathrm{T}} \mathbf{I} + \alpha \gamma\, \mathbf{C}_i \mathbf{C}_i^{\mathrm{T}}\right)$,

  $\mathbf{f} = -\mathbf{V}_i \underline{\mathbf{h}}_i^{\mathrm{T}}$,

  where $\gamma = \det\left(\overline{\mathbf{W}_i}^{\mathrm{T}} \overline{\mathbf{W}_i}\right)$.

  $\mathbf{w}_i = \mathrm{QP}(\mathbf{Q}_i, \mathbf{f}, *)$;

  Update $\underline{\mathbf{h}}_i$ according to (15);

  end

end.

The QP_MVC_NMF algorithm uses the block coordinate descent methods. Since the solutions of both subproblems, i.e., (P1) and (P2), are unique, the QP_MVC_NMF is at least of local convergence [32].

B. NG_MVC_NMF Algorithm

Generally, the QP_MVC_NMF algorithm is efficient when the number of the sources, i.e., r, is relatively small. Otherwise, the following multiplicative updating rule based algorithm is recommended.

Consider the partial derivatives of J with respect to $\mathbf{W}$ and $\mathbf{H}$

$$\frac{\partial J}{\partial \mathbf{H}} = \mathbf{W}^{\mathrm{T}} \mathbf{WH} - \mathbf{W}^{\mathrm{T}} \mathbf{V} \quad (16)$$

$$\frac{\partial J}{\partial \mathbf{W}} = \mathbf{WHH}^{\mathrm{T}} - \mathbf{VH}^{\mathrm{T}} + \alpha \det\left(\mathbf{W}^{\mathrm{T}} \mathbf{W}\right) \mathbf{W} \left(\mathbf{W}^{\mathrm{T}} \mathbf{W}\right)^{-1}. \quad (17)$$

The updating rules for $\mathbf{H}$ is

$$\mathbf{H} = \mathbf{H} - \eta_{\mathbf{H}} \odot \frac{\partial J}{\partial \mathbf{H}}. \quad (18)$$

Let $\eta_{\mathbf{H}} = (\mathbf{H} / \mathbf{W}^{\mathrm{T}} \mathbf{WH})$, and it follows that:

$$\mathbf{H} = \mathbf{H} \odot \frac{\mathbf{W}^{\mathrm{T}} \mathbf{V} + \delta}{\mathbf{W}^{\mathrm{T}} \mathbf{WH} + \delta} \quad (19)$$

where δ is a small positive constant to avoid numerical instabilities [34]. Note that we cannot derive the updating rule for $\mathbf{W}$ in the same way, because $(\partial J / \partial W)$ includes the computation of the inverse matrix of $\mathbf{W}^{\mathrm{T}}\mathbf{W}$, which probably contains many negative entries. To overcome this difficulty, the so-called natural gradient is introduced [24], [35]. In fact, since $\mathbf{W}$ is of full column rank, there exists a matrix

$\mathbf{X}$ such that $\mathbf{XW} = \mathbf{I}$. Consequently, the parameter matrix $\mathbf{W}$ possesses a special algebraic structure, namely Li group structure, which is not possessed by vectors. This property makes the variables behave like a curved manifold (Riemann manifold). It is known that the natural gradient, instead of the ordinary gradient, is the steepest descent direction in the Riemann manifold [24], [35]. By using natural gradient, the updating rule for $\mathbf{W}$ should be

$$\mathbf{W} = \mathbf{W} - \eta_{\mathbf{W}} \odot \frac{\partial J}{\partial \mathbf{W}} \mathbf{W}^{\mathrm{T}}\mathbf{W}. \tag{20}$$

Let $\eta_{\mathbf{W}} = (\mathbf{W}/\mathbf{WHH}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}W + \alpha \det(\mathbf{W}^{\mathrm{T}}\mathbf{W})W)$, we have

$$\mathbf{W} = \mathbf{W} \odot \frac{\mathbf{VH}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \delta}{\mathbf{WHH}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \alpha\gamma\,\mathbf{W} + \delta} \tag{21}$$

where $\gamma = \det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$.

It can be seen that the use of natural gradient in (21) is crucial because it can not only possibly speed up the convergence but also avoid the computation of the inverse matrix, and the nonnegativity of parameters is therefore maintained. Due to the use of natural gradient, the algorithm based on (19) and (21) is called NG_MVC_NMF.

Remark 2: A general convergence for NG_MVC_NMF is difficult. However, in perfect factorization cases, the parameter α approaches zero gradually, and thus NG_MVC_NMF reduces to be the unconstrained NMF algorithm ultimately which has the convergence guarantee, see [29], [36], [37]. For the gradient-based algorithms, the learning step should be sufficiently small to ensure the convergence. For this reason, one may consider replacing $\eta_{\mathbf{W}}$ by $\rho\eta_{\mathbf{W}}$, where $0 < \rho \leq 1$. Thus the following update formula yields:

$$\mathbf{W} = (1-\rho)\mathbf{W} + \rho\mathbf{W} \odot \frac{\mathbf{VH}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \delta}{\mathbf{WHH}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \alpha\gamma\,\mathbf{W} + \delta}. \tag{22}$$

When $\rho = 1$, (22) reduces to (21). NG_MVC_NMF with a smaller value of ρ is generally more stable but suffers from a slower convergence speed. Besides, the improved version of multiplicative updates in [36] can be incorporated into the proposed algorithm. Nevertheless, we find that (21) actually works quite well in practice.

C. Parameters, Initialization, and Complexity

The parameter α balances the fitting error and the volume of the simplex formed by $\mathbf{W}$, and plays a critical role in the proposed method. In this paper, α is selected using the exponential rule $\alpha_k = \sigma \exp(-\tau\kappa)$ in [38], where κ is a nondecreasing function of the iteration number, and $\sigma \geq 0$ and $\tau \geq 0$ are constants such that $0 \leq \alpha_{k+1} \leq \alpha_k$. Obviously, the proposed method reduces to be unconstrained NMF if $\sigma = 0$. By letting $\tau = 0$ and $\sigma \neq 0$, a tradeoff between the fitting error and the sparseness of $\mathbf{H}$ can be achieved. If $\sigma > 10$, $\tau > 0$, the sparsest solution with the smallest fitting error is desired. A major problem is that the value of $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$ is generally quite small when M is large. To make the volume constraint effective, two terms in (10), i.e., the fitting error $\|\mathbf{V} - \mathbf{WH}\|_{\mathrm{F}}^2$ and the volume constraint $\det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$, should be properly scaled in practice. Note that $\|\mathbf{V}\|_{\mathrm{F}}^2 \leq N$ as each column of $\mathbf{V}$ sums to unity. Let $\lambda = \max((M - r/M)N, 1)$. It can be easily verified that $\min \|\mathbf{V} - \mathbf{WH}\|_{\mathrm{F}}^2 \leq (M - r/M)\|\mathbf{V}\|_{\mathrm{F}}^2 \leq (M - r/M)N \leq \lambda$. On the other hand, we have $0 \leq \lambda \det(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \leq \lambda$ from $0 \leq \det(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \leq 1$. Let $\alpha = \lambda\gamma_*\sigma \exp(-\tau\kappa)$, where $\gamma_* = 1/\det(\mathbf{W}_*^{\mathrm{T}}\mathbf{W}_*)$ and $\mathbf{W}_*$ is the true basis matrix. Then there holds that $0 \leq \alpha \det(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \to \lambda\sigma \exp(-\tau\kappa)$ since $\gamma_* \det(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \to 1$ on condition that $\mathbf{W}_k \to \mathbf{W}_*$ even if $\det(\mathbf{W}_*^{\mathrm{T}}\mathbf{W}_*)$ is very small. By this trick, the selection of σ and τ can be more intuitive. The problem is that the true basis mixing matrix (and thus γ_*) is generally unknown. In practice, we can run unconstrained NMF algorithms by setting $\sigma = 0$ at first to approximate the scale of γ_* and then turn to the constrained NMF.[1] In the constrained NMF period, especially when the algorithm is close to convergence, we can set $\gamma_* \approx 1/\det(\mathbf{W}_k^{\mathrm{T}}\mathbf{W}_k)$ in the $(k+1)$th iteration and make sure that $0 \leq \alpha_{k+1} \leq \alpha_k$ at the same time. Considering that $\det(\mathbf{W}_{k+1}^{\mathrm{T}}\mathbf{W}_{k+1}) \approx \det(\mathbf{W}_k^{\mathrm{T}}\mathbf{W}_k)$ when the algorithm almost converges, we may simply let $\gamma_* = 1/\det(\mathbf{W}_{k+1}^{\mathrm{T}}\mathbf{W}_{k+1})$. This suggests that we can let $\gamma_*\gamma = 1$ and do not need to compute γ (see Proposition 6 and (21) for their definitions) in each iteration. Hereafter, the parameter σ denotes $\sigma\gamma_*\gamma$ and $\alpha = \lambda\sigma \exp(-\tau\kappa)$.

Regarding the initialization, first of all, the columns of $\mathbf{V}$ should be normalized to sum to unity, see (7) and the corresponding text for details. Finally, the output matrix $\mathbf{H}$ should be adjusted accordingly.

An optional initialization step is to provide initial values for $\mathbf{W}$ and $\mathbf{H}$ as good as possible, which can speed up the convergence of NMF algorithms significantly and escape from some local minima. Moreover, as discussed above, it makes the selection of α much easier. For this issue, Boutsidisa and Gallopoulosb developed SVD-based initialization techniques [39], and Xue *et al.* proposed the clustering-based initialization method [40]. In fact, any NMF algorithms can be used as the initialization procedure. However, the efficiency should be considered in practice, because it is often not worth spending much time in the initialization. In our experiments, the VCA algorithm proposed in [20] is shown to be a suitable choice because of its striking efficiency.

Another optional initialization technique is about the data reduction, which has been mentioned in [4]. Suppose $\overline{\mathbf{V}}$ is the matrix whose columns are chosen from the columns of $\mathbf{V}$ and form a convex hull of $\mathbf{v}_t$. Hence, for any $t \in \mathcal{I}_N$, there exists a vector $\mathbf{x}_t \geq 0$ such that $\mathbf{v}_t = \overline{\mathbf{V}}\mathbf{x}_t$, $\mathbf{1}^{\mathrm{T}}\mathbf{x}_t = 1$. Suppose that $(\mathbf{W}, \overline{\mathbf{H}}) = \mathrm{NMF}(\overline{\mathbf{V}}, r)$, and we have $\mathbf{v}_t = \overline{\mathbf{V}}\mathbf{x}_t = \mathbf{W}\overline{\mathbf{H}}\mathbf{x}_t$. Let $\mathbf{h}_t = \overline{\mathbf{H}}\mathbf{x}_t$, then $\mathbf{1}^{\mathrm{T}}\mathbf{h}_t = \mathbf{1}^{\mathrm{T}}\overline{\mathbf{H}}\mathbf{x}_t = \mathbf{1}^{\mathrm{T}}\mathbf{x}_t = 1$ yields, which means that $(\mathbf{W}, \mathbf{H}) = \mathrm{NMF}(\mathbf{V}, r)$. In other words, we only need to run NMF algorithms on the significantly reduced observation data $\overline{\mathbf{V}}$. However, searching the convex hull in high-dimensional spaces is in itself computationally expensive For example, the computational complexity of the gift wrapping algorithm is $\mathrm{O}\left(N^{\lfloor M/2 \rfloor + 1}\right)$, where $\lfloor x \rfloor$ is the highest integer lower than or equal to x, N is the number of samples, and M is the dimensionality of points [20]. Therefore,

[1] However, in our simulations we find that it is often unnecessary to run unconstrained NMF in advance.

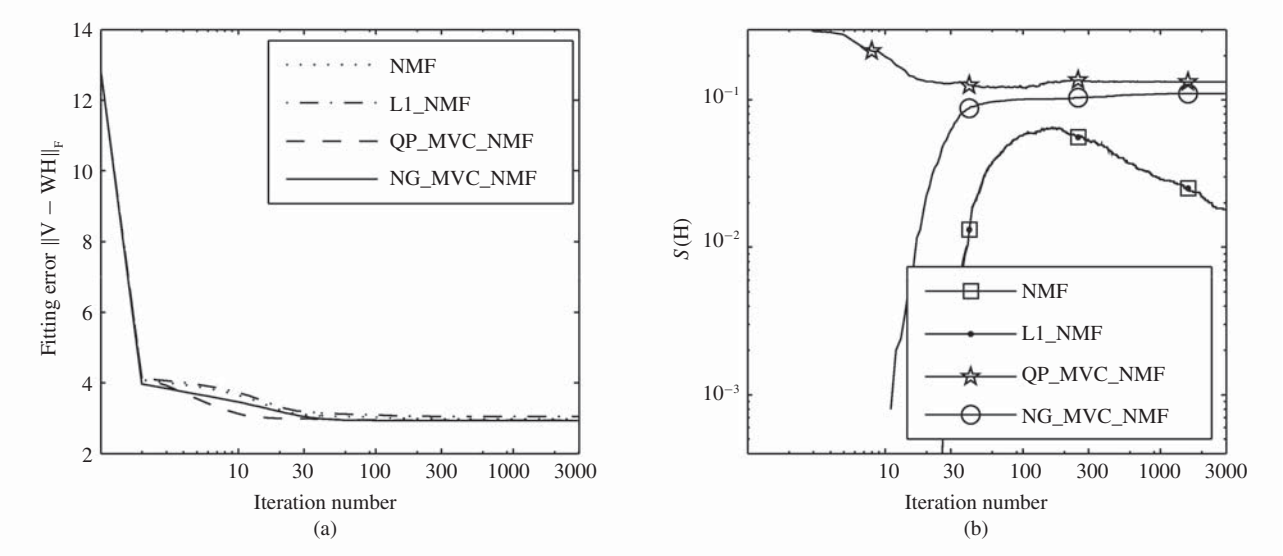

Fig. 2. Evolution of (a) fitting error and (b) sparseness of the encoding matrix versus the iteration number of the four NMF algorithms.

searching for a convex hull in high-dimensional spaces should be avoided.

It can be easily verified that the computational complexity of the NG_MVC_NMF algorithm is $O(MrN)$ per iteration. Note that (P1) is a quadratic programming problem with M variables and $M+1$ constraints. If M is small and $M \ll N$, the time consumed by solving (P1) is nearly negligible and thus the total computational complexity of QP_MVC_NMF is about $O(MrN)$ per iteration. For small-scale problems, QP_MVC_NMF only needs much fewer iterations to converge than NG_MVC_NMF and thus is much more efficient. In summary, the proposed methods have competent complexity order compared to other NMF methods [4], [41]. It is worth noticing that, although nLCA_IVM also has a complexity of $O(MrN)$ on average, the complexity would be $O(MrN^{1.5})$ in the worst case [4], [41]. Moreover, nLCA_IVM is not very competent when $r < M$, because a dimensionality reduction procedure should be performed in nonnegative space [4], [42]. As a result, nLCA_IVM may suffer from a slow convergence speed when N is very large and $r < M$.

IV. Simulations

Four simulations are performed to test the proposed algorithms. In the first one, the convergence of the proposed algorithms is investigated. Then the proposed algorithms are applied to BSS and human face images analysis. In BSS applications, the following signal-to-noise ratio (SNR) is utilized to evaluate the recovery accuracy of the sources [24]:

$$SNR = 10\log\frac{E[s]^2}{E[s-\widehat{s}]^2} \tag{23}$$

where $\widehat{s}$ is an estimate of s, and $\widehat{s}$, s are normalized to be of zero mean and unit variance. The operator $E[\cdot]$ denotes the mathematical expectation. The maximum iteration number is 3000, if not explicitly specified. For the proposed algorithms, κ is simply equal to the iteration number.

Note that, although (8) is able to give the unique sparsest solutions to (2) theoretically, both the proposed algorithms are of local convergence and thus cannot always converge to the sparsest solution. For this reason, the VCA algorithm in [20] is used as the initialization procedure in BSS tasks to increase the probability of global convergence. To investigate the robustness of the proposed methods to the suboptimal solutions, the initial factors in simulations 1 and 4 are randomly generated with elements drawn from independent uniform distributions between 0 and 1.

Simulation 1: Investigation of convergence. The entries of the matrix $\mathbf{V}$ are drawn from a uniform distribution between 0 and 1, $M = 10$, $r = 5$, and $N = 1000$. Besides, its columns are normalized to sum to unity. The proposed methods are compared with the NMF algorithm [29] and the L_1-NMF algorithm [16]. Because both the L_1-NMF algorithm and the NMF algorithm in [17] use L_1-norm-based sparseness constraint, only L_1-NMF is compared (The MATLAB code for L_1-NMF can be found in [43]). NMFSC in [2] is not compared because it is based on the assumption that all the columns of $\mathbf{H}$ are of the same sparseness level, which does not hold here. All the algorithms start from the same random initial matrices of $\mathbf{W}$ and $\mathbf{H}$. For QP_MVC_NMF, $\sigma = 10^{-2}$, $\tau = 0$ and for NG_MVC_NMF, $\sigma = 10^{-3}$, $\tau = 0$ is set. For the L_1-NMF algorithm, $\alpha_{\mathbf{H}} = 1$ is set to control the sparseness in all the simulations. Fig. 2 shows the evolution of the fitting error ($\|\mathbf{V} - \mathbf{WH}\|_F$) and the sparseness of the encoding matrix versus the iteration number. From the figure, the proposed methods are seen to outperform the others. Particularly, QP_MVC_NMF achieves the least fitting error and the highest sparseness, and has a fast convergence speed. Also, it can be seen that L_1-NMF cannot improve the sparseness substantially. This is mainly because L_1-norm describes the sparseness poorly in weak sparseness situations.

Simulation 2: Applications in BSS when $M = r$. The sources are six human face images chosen from the benchmark

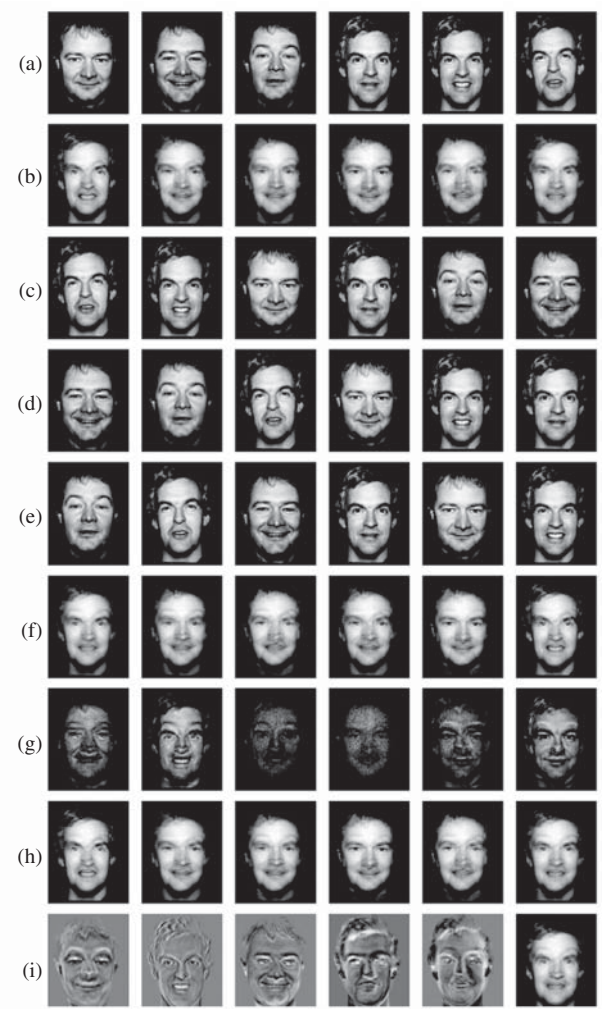

Fig. 3. Separation of human face images in the absence of noise. (a) sources, (b) observations and the sources estimated by, (c) QP_MVC_NMF, (d) NG_MVC_NMF, (e) nLCA_IVM, (f) NMF, (g) L1-NMF, (h) VCA, and (i) FastICA.

TABLE I
CORRELATION COEFFICIENTS OF THE SIX HUMAN FACE IMAGE SIGNALS

	h_1	h_2	h_3	h_4	h_5	h_6
h_1	1.0000	0.9106	0.9008	0.8421	0.8388	0.8265
h_2	0.9106	1.0000	0.8804	0.8379	0.8363	0.8237
h_3	0.9008	0.8804	1.0000	0.8085	0.8017	0.8361
h_4	0.8421	0.8379	0.8085	1.0000	0.9279	0.8451
h_5	0.8388	0.8363	0.8017	0.9279	1.0000	0.8592
h_6	0.8265	0.8237	0.8361	0.8451	0.8592	1.0000

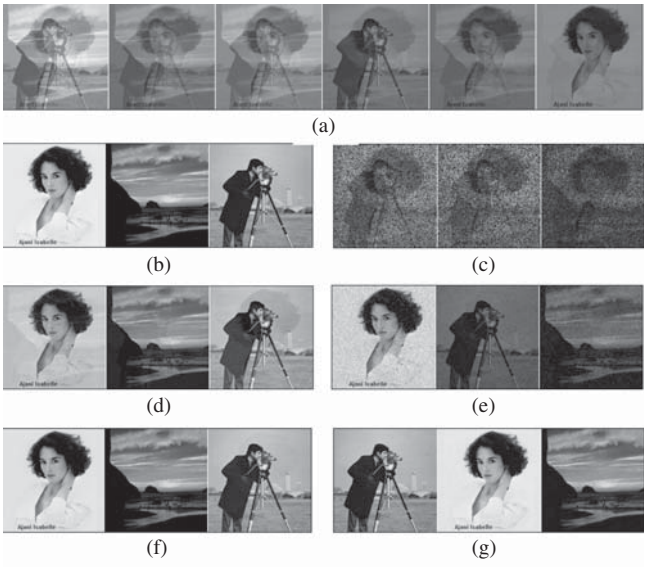

Fig. 4. Separation of natural images in the presence of additive noise with SNR = 30 dB (a) mixtures, (b) sources, and the estimated sources obtained by, (c) NMF, (d)L1_NMF, (e) nLCA_IVM, (f) NG_MVC_NMF, (g) QP_MVC_NMF

of NMFLABIP [44] with 96 250 samples. In this simulation, the proposed two algorithms are compared with nLCA_IVM [4], NMF, L_1-NMF, VCA [20], and FastICA [45], [46]. Both nLCA_IVM and QP_MVC_NMF use the reduced observations by running the quickhull algorithm [47]. Here, set $\sigma = \tau = 0.01$. The separation results in a typical run are shown in Fig. 3. It can be seen from the figure that the FastICA algorithm cannot give satisfactory results. In fact, the six face image signals are highly correlated. The correlation coefficients between any two of them are higher than 0.8, as shown in Table I. The images estimated by QP_MVC_NMF, NG_MVC_NMF, and nLCA_IVM match the original sources better than those estimated by the others. Table II shows the separation results averaged over 100 Monte Carlo runs. As is seen from the table, nLCA_IVM achieves the highest separation accuracy among all the methods and the proposed methods win the second. However, the time consumption of QP_MVC_NMF is only half of nLCA_IMV. Also note that the VCA method converges quite fast, which makes it very suitable for the initialization.

Simulation 3: Applications in BSS when $M > r$ in presence of noise. In this experiment, three natural images are mixed by a 6×3 matrix in each run. The observations are corrupted by an additive truncated Gaussian noise with SNR = 30 dB. Fig. 4 shows the results of a typical run obtained by NMF, L_1-NMF, nLCA_IVM, NG_MVC_NMF ($\sigma = 0.001$, $\tau = 0.1$) and QP_MVC_NMF ($\sigma = 0.01$, $\tau = 0.1$), respectively. By visual comparison, NMF cannot separate the sources. L_1-NMF performs better, but there exist some unwanted interferences from the other images. The nLCA_IVM algorithm separates the three images successfully, but evident noise exists in each picture. Compared with these methods, the proposed methods achieve higher separation accuracy. Table III shows the separation results averaged over 100 Monte Carlo runs. From the table, QP_MVC_NMF outperforms the other algorithms in terms of accuracy and efficiency, and NG_MVC_NMF also achieves desirable results, which shows the validity of the MVC. Particularly, QP_MVC_NMF is not only faster than nLCA_IVM this time but also achieves higher separation accuracy, which may suggest that, compared with nLCA_IVM, the proposed methods are more competent in overdetermined and noise cases.

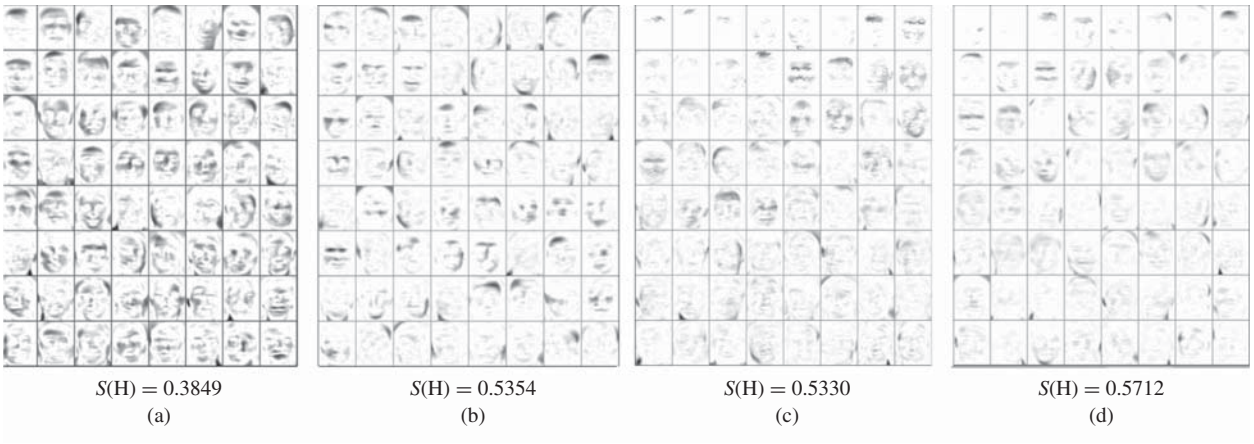

Fig. 5. Applications of NMF algorithms in ORL human face images analysis. (a) NMF (also L1-NMF) gives rather global representation. (b), (c), and (d) NMFSC, QP_MVC_NMF, and NG_MVC_NMF can give more parts-based representation.

TABLE II
AVERAGED SNR BETWEEN SOURCES AND THEIR ESTIMATES, AVERAGED COMPUTATION TIME OF VARIOUS METHODS OVER 100 MONTE CARLO RUNS

Method	SNR (dB)						Time(s)
QP_MVC_NMF	40.9136	43.3909	52.9528	56.4548	59.8593	64.3552	14.83
NG_MVC_NMF	33.2196	34.7493	44.6217	48.2681	54.6563	60.1248	25.58
nLCA_IVM	47.4586	50.2105	55.9140	62.4813	70.2974	100.2495	34.39
NMF	6.1033	6.8163	7.5157	8.4690	10.3801	11.2237	0.45
L_1-NMF	9.3169	9.9094	10.3645	10.7434	11.2448	12.2464	221.08
VCA	9.3162	9.7520	10.2189	10.5791	11.0558	11.9074	0.28
FastICA	5.4493	5.7903	6.1342	7.8638	9.4769	12.6132	9.49

TABLE III
AVERAGED SNR BETWEEN SOURCES AND THEIR ESTIMATES, AVERAGED COMPUTATION TIME OF VARIOUS METHODS OVER 100 MONTE CARLO RUNS WHEN $M = 6$ AND $R = 3$ IN PRESENCE OF ADDITIVE NOISE

Method	SNR (dB)			Time(s)
NMF	−1.5701	−1.2801	1.8992	0.10
L_1-NMF	8.3224	11.2138	15.2281	25.9
nLCA_IVM	12.0741	15.6102	19.1660	3.2
QP_MVC_NMF	19.7129	25.0928	27.5469	3.07
NG_MVC_NMF	19.4473	23.0795	27.4660	8.63

Simulation 4: Applications in human face images analysis. A major advantage of NMF is its ability of learning parts. However, in some cases the NMF representation is global rather than parts-based, for example, when the classic NMF algorithm is applied to the analysis of the Olivetti Research Laboratory (ORL) face images [2] (the ORL face database can be found in the nmfpack [43], or at http://www.uk.research.att.com/facedatabase.html). In this simulation, the proposed algorithms are applied to the analysis of ORL face images. There are total 400 human face images in the ORL database with the size of 46×56 pixel. In this experiment, NMFSC is compared, where the sparseness of $\mathbf{H}$ is 0.6 and there is no sparseness constraint on $\mathbf{W}$. For the proposed algorithms, $\sigma = 0$ is set at first to obtain the initial guess of the decomposition. Then $\sigma = 10^{-4}$, $\tau = 0$, $\sigma = 10^{-6}$, and $\tau = 0$ are set in QP_MVC_NMF and NG_MVC_NMF, respectively. All the algorithms start from the same initial random matrices of $\mathbf{W}$ and $\mathbf{H}$. $r = 64$ is set and the results are shown in Fig. 5 (The result obtained by L_1-NMF is omitted because it obtains almost the same results as NMF). From the figure, NMFSC, QP_MVC_NMF, and NG_MVC_NMF give much sparser solutions, i.e., more parts-based representations. A major limitation of NMFSC is that each column of the factors must be of the same sparseness. For example, in this experiment $\mathcal{S}(\mathbf{h}_t) = 0.5938$ for all t. This limitation sometimes increases the fitting error and causes explanation problems in practice. Note that this experiment deals with a large-scale problem with $M = 400$, hence NG_MVC_NMF is much faster than QP_MVC_NMF. In a typical run, QP_MVC_NMF consumes 2153 s, whereas NG_MVC_NMF consumes only 295 s. Since they give almost the same results, NG_MVC_NMF is preferred when r is large.

V. CONCLUSION

A study on MVC NMF model was presented in this paper, together with several theoretical results on the relation between NMF, sparse NMF, and MVC NMF. Also, a quadratic-programming-based algorithm QP_MVC_NMF and a natural-gradient-based algorithm NG–MVC–NMF were developed to solve the proposed model. The former is quite efficient when the problem is of relatively small size, whereas the latter is more appropriate for the large-scale problems. The main

advantage of the proposed methods is that they can obtain the desired results even if the sparseness is weak. The proposed methods are essentially L_0-norm-oriented and can enhance the ability of learning parts of NMF significantly. Simulations showed the superiority of the proposed algorithms over the traditional methods (NMF, L_1-NMF, etc).

The proposed methods are very suitable for BSS (especially for the separation of correlated sources), remote-sensing image interpretation, images analysis, etc. Note that the uniqueness is often crucial for NMF problems. In this paper, we only consider the simplest case that can offer unique solutions. Exploring relaxed uniqueness conditions is still a challenging task and should be our future work.

APPENDIX A
PROOF OF PROPOSITION 3

Let $\mathbf{H} = \mathbf{WZ}$ where $\mathbf{W} \in \mathbb{R}_+^{r\times r}$ and $\mathbf{Z} \in \mathbb{R}_+^{r\times N}$. First we show that there is only one nonzero element in the ith row of $\mathbf{W}$ if the rank of $\mathbf{H}_i^0$ is $r-1$. Suppose that $w_{ij} > 0$ and there exists $k \neq j$ such that $w_{ik} > 0$. Let $\mathbf{H}_i^0 = \mathbf{WQ}$ where $\mathbf{Q}$ consists of the columns of $\mathbf{Z}$ associated with $\mathbf{H}_i^0$. Note that the ith row of $\mathbf{H}_i^0$ are all zeros and, consequently, the jth and kth rows of $\mathbf{Q}$ are all zeros, too. That is to say, the rank of $\mathbf{Q}$ is not larger than $r-2$, hence the rank of $\mathbf{H}_i^0$ is not larger than $r-2$, which contradicts the assumptions. Thus there is only one nonzero element in the ith row of $\mathbf{W}$.

As a result, if the rank of $\mathbf{H}_i^0$ is $r-1$ for any $i \in \mathcal{I}_r$, $\mathbf{W}$ must be a generalized permutation matrix, i.e., $\mathbf{H}$ is not factorable.

APPENDIX B
PROOF OF PROPOSITION 4

Without loss of generality, we let $\mathbf{H} = \begin{bmatrix}\mathbf{H}_1 & \mathbf{H}_2\end{bmatrix}$ and $\widehat{\mathbf{H}} = \begin{bmatrix}\widehat{\mathbf{H}}_1 & \widehat{\mathbf{H}}_2\end{bmatrix}$ where $\mathbf{H}_1 \in \mathcal{P}_+^{r\times r}$ and $\widehat{\mathbf{H}}_1 \in \mathbb{R}_+^{r\times r}$. Note that $\mathbf{W}\begin{bmatrix}\mathbf{H}_1 & \mathbf{H}_2\end{bmatrix} = \widehat{\mathbf{W}}\begin{bmatrix}\widehat{\mathbf{H}}_1 & \widehat{\mathbf{H}}_2\end{bmatrix}$. From $\mathbf{H}_1 \in \mathcal{P}_+$, we have $\mathbf{W} = \widehat{\mathbf{W}}\widehat{\mathbf{H}}_1\mathbf{H}_1^{-1} \geq 0$.

If $\widehat{\mathbf{H}}_1 \in \mathcal{P}_+$, then $\widehat{\mathbf{H}}_1\mathbf{H}_1^{-1} = \mathbf{P}_+ \in \mathcal{P}_+$, hence $\mathbf{W} = \widehat{\mathbf{W}}\mathbf{P}_+$ and $\mathbf{H} = \mathbf{P}_+^{-1}\widehat{\mathbf{H}}$, i.e., $\mathcal{S}(\mathbf{H}_2) = \mathcal{S}(\mathbf{H}_1)$.

If $\widehat{\mathbf{H}}_1 \notin \mathcal{P}_+$, then $\|\widehat{\mathbf{H}}_1\|_0 > \|\mathbf{H}_1\|_0$. Moreover, from $\widehat{\mathbf{W}}\widehat{\mathbf{H}}_2 = \mathbf{WH}_2 = \widehat{\mathbf{W}}\widehat{\mathbf{H}}_1\mathbf{H}_1^{-1}\mathbf{H}_2$, $\widehat{\mathbf{H}}_2 = \widehat{\mathbf{H}}_1\mathbf{H}_1^{-1}\cdot\mathbf{H}_2$ holds, which means that $\|\widehat{\mathbf{H}}_2\|_0 \geq \|\mathbf{H}_2\|_0$. That is, $\mathcal{S}(\widehat{\mathbf{H}}) < \mathcal{S}(\mathbf{H})$. The proof is complete.

APPENDIX C
PROOF OF PROPOSITION 5

Suppose that there exists i such that $\mathcal{S}(\underline{\mathbf{h}}_i) = 0$. It is to be shown that there always exists $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{H}}$ such that $\|\mathbf{V}-\mathbf{WH}\|_{\mathrm{F}} = \|\mathbf{V}-\widehat{\mathbf{W}}\widehat{\mathbf{H}}\|_{\mathrm{F}}$ and $\det(\widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}}) < \det(\mathbf{W}^{\mathrm{T}}\mathbf{W})$. Let $\varphi = \min_j(h_{ij}/h_{kj})$, for some $k \neq i$ and all $h_{kj} > 0$. Let $\mathbf{Z}$ equal to an $r\times r$ identity matrix except that the ikth element is $-\varphi$ and the kkth element is $1+\varphi$. Let $\widehat{\mathbf{W}} = \mathbf{WZ}^{-1}$, $\widehat{\mathbf{H}} = \mathbf{ZH}$, it can be easily verified that $\widehat{\mathbf{W}}$, $\widehat{\mathbf{H}}$ satisfy the assumption A2, A3, and $\mathbf{WH} = \widehat{\mathbf{W}}\widehat{\mathbf{H}}$, $\mathcal{S}(\widehat{\mathbf{H}}) > \mathcal{S}(\mathbf{H})$. Particularly, we have

$$\det\left(\widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}}\right) = \left(\frac{1}{1+\varphi}\right)^2 \det\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right) < \det\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right).$$

The proof is complete.

APPENDIX D
PROOF OF PROPOSITION 6

Let $\mathbf{W} = \begin{bmatrix}\mathbf{w}_i & \overline{\mathbf{W}_i}\end{bmatrix}\mathbf{P}$, where $\mathbf{P}$ is a permutation matrix. Thus we have

$$\det\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right) = \det\left(\begin{bmatrix}\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i & \mathbf{w}_i^{\mathrm{T}}\overline{\mathbf{W}_i} \\ \overline{\mathbf{W}_i}^{\mathrm{T}}\mathbf{w}_i & \overline{\mathbf{W}_i}^{\mathrm{T}}\overline{\mathbf{W}_i}\end{bmatrix}\right)$$
$$= \det\left(\overline{\mathbf{W}_i}^{\mathrm{T}}\overline{\mathbf{W}_i}\right)\cdot\mathbf{w}_i^{\mathrm{T}}\left(\mathbf{I}-\overline{\mathbf{W}_i}\left(\overline{\mathbf{W}_i}^{\mathrm{T}}\overline{\mathbf{W}_i}\right)^{-1}\overline{\mathbf{W}_i}^{\mathrm{T}}\right)\mathbf{w}_i.$$

The proof of 1) is completed.

From $\mathbf{C}_i = Null(\overline{\mathbf{W}_i}^{\mathrm{T}})$, we have $\mathbf{C}_i^{\mathrm{T}}\mathbf{C}_i = \mathbf{I}$ and $\mathbf{C}_i^{\mathrm{T}}\overline{\mathbf{W}_i} = \mathbf{0}$. Note that the columns of $\mathbf{C}_i$ and $\overline{\mathbf{W}_i}$ form a base of the M-dimensional vector space. Let $\mathbf{w}_i = \mathbf{C}_i\mathbf{x} + \overline{\mathbf{W}_i}\mathbf{y}$. Then we have

$$\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i = \mathbf{x}^{\mathrm{T}}\mathbf{x} + \mathbf{y}^{\mathrm{T}}\overline{\mathbf{W}_i}^{\mathrm{T}}\overline{\mathbf{W}_i}\mathbf{y} \tag{24}$$

and

$$\mathbf{w}_i^{\mathrm{T}}\overline{\mathbf{W}_i}\left(\overline{\mathbf{W}_i}^{\mathrm{T}}\overline{\mathbf{W}_i}\right)^{-1}\overline{\mathbf{W}_i}^{\mathrm{T}}\mathbf{w}_i = \mathbf{y}^{\mathrm{T}}\overline{\mathbf{W}_i}^{\mathrm{T}}\overline{\mathbf{W}_i}\mathbf{y}. \tag{25}$$

From (24) and (25), we have

$$\mathbf{w}_i^{\mathrm{T}}\mathbf{C}_i\mathbf{C}_i^{\mathrm{T}}\mathbf{w}_i = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \mathbf{w}_i^{\mathrm{T}}\left(\mathbf{I}-\overline{\mathbf{W}_i}\left(\overline{\mathbf{W}_i}^{\mathrm{T}}\overline{\mathbf{W}_i}\right)^{-1}\overline{\mathbf{W}_i}^{\mathrm{T}}\right)\mathbf{w}_i.$$

The proof is completed.

ACKNOWLEDGMENT

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions that led to the present improved version of the original manuscript.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[2] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Nov. 2004.
[3] D. Lee and S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. 13*, 2001, pp. 556–562.
[4] F. Y. Wang, C. Y. Chi, T. H. Chan, and Y. Wang, "Nonnegative least-correlated component analysis for separation of dependent sources by volume maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 875–888, May 2010.
[5] S. Zafeiriou, "Discriminant nonnegative tensor factorization algorithms," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 217–235, Feb. 2009.
[6] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Trans. Neural Netw.*, vol. 19, no. 6, pp. 1090–1100, Jun. 2008.
[7] Z. R. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 734–749, May 2010.
[8] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 683–695, May 2006.
[9] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Proc. Adv. Neural Inf. Process. Syst. 16*, 2003, pp. 1141–1148.
[10] A. Cichocki, A. Phan, R. Zdunek, and L.-Q. Zhang, "Flexible component analysis for sparse, smooth, nonnegative coding or representation," in *Proc. Neural Inf. Process.*, 2008, pp. 811–820.
[11] M. Heiler and C. Schnorr, "Learning sparse representations by non-negative matrix factorization and sequential cone programming," *J. Mach. Learn. Res.*, vol. 7, pp. 1385–1407, Jul. 2006.
[12] K. Stadlthanner, D. Lutter, F. J. Theis, E. W. Lang, A. M. Tome, P. Georgieva, and C. G. Puntonet, "Sparse nonnegative matrix factorization with genetic algorithms for microarray analysis," in *Proc. Int. Joint Conf. Neural Netw.*, Orlando, FL, Aug. 2007, pp. 294–299.

[13] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 4. Jul. 2004, pp. 2529–2533.
[14] F. J. Theis, K. Stadlthanner, and T. Tanaka, "First results on uniqueness of sparse non-negative matrix," in *Proc. 13th Eur. Signal Process. Conf.*, 2005, pp. 1–4.
[15] H. Laurberg and L. K. Hansen, "On affine non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Honolulu, HI, Apr. 2007, pp. 653–656.
[16] P. O. Hoyer, "Non-negative sparse coding," in *Proc. 12th IEEE Workshop Neural Netw. Signal Process.*, Nov. 2002, pp. 557–565.
[17] L. Weixiang, Z. Nanning, and L. Xiaofeng, "Non-negative matrix factorization for visual coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3. Apr. 2003, pp. 293–296.
[18] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.
[19] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, May 1994.
[20] J. M. P. Nascimento and J. M. B. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, Apr. 2005.
[21] R. Schachtner, G. Popel, and E. W. Lang, "Toward unique solutions of non-negative matrix factorization problems by a determinant criterion," *Digital Signal Process.*, vol. 21, no. 4, pp. 528–534, Jul. 2011.
[22] G. Zhou, Z. Yang, S. Xie, J.-M. Yang, "Online blind source separation using incremental nonnegative matrix factorization with volume constraint," *IEEE Trans. Neural Netw.*, vol. 22, no. 4, pp. 550–560, Apr. 2011.
[23] T. H. Chan, C. Y. Chi, Y. M. Huang, and W. K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418–4432, Nov. 2009.
[24] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. New York: Wiley, 2003.
[25] V. Zarzoso and P. Comon, "Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size," *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 248–261, Feb. 2010.
[26] Y. Washizawa, Y. Yamashita, T. Tanaka, and A. Cichocki, "Blind extraction of global signal from multi-channel noisy observations," *IEEE Trans. Neural Netw.*, vol. 21, no. 9, pp. 1472–1481, Sep. 2010.
[27] G. X. Zhou, S. L. Xie, Z. Y. Yang, and J. Zhang, "Nonorthogonal approximate joint diagonalization with well-conditioned diagonalizers," *IEEE Trans. Neural Netw.*, vol. 20, no. 11, pp. 1810–1819, Nov. 2009.
[28] G. X. Zhou, Z. Y. Yang, S. L. Xie, and J. M. Yang, "Mixing matrix estimation from sparse mixtures with unknown number of sources," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 211–221, Feb. 2011.
[29] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. 13*, 2000, pp. 556–562.
[30] S. A. Robila and L. G. Maciak, "Considerations on parallelizing non-negative matrix factorization for hyperspectral data unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 57–61, Jan. 2009.
[31] R. Schachtner, G. Pöppel, A. Tomé, and E. Lang, "Minimum determinant constraint for non-negative matrix factorization," in *Proc. Independ. Comp. Anal. Signal Separat.*, 2009, pp. 106–113.
[32] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.
[33] B. Nils, "The gramian and K-volume in N-space: Some classical results in linear algebra," *J. Young Investigat.*, vol. 2, no. 1, pp. 1–4, 1999.
[34] A. Cichocki, R. Zdunek, and S. I. Amari, "Nonnegative matrix and tensor factorization [lecture notes]," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 142–145, Jan. 2008.
[35] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, Mar. 1998.
[36] C. J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, Nov. 2007.
[37] R. Badeau, N. Bertin, and E. Vincent, "Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 12, pp. 1869–1881, Dec. 2010.
[38] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization," *Signal Process.*, vol. 87, no. 8, pp. 1904–1916, Aug. 2007.
[39] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognit.*, vol. 41, no. 4, pp. 1350–1362, 2008.
[40] Y. Xue, C. S. Tong, Y. Chen, and W. S. Chen, "Clustering-based initialization for non-negative matrix factorization," *Appl. Math. Comput.*, vol. 205, no. 2, pp. 525–536, Nov. 2008.
[41] 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.
[42] M. D. Plumbley and E. Oja, "A 'nonnegative PCA' algorithm for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 66–76, Jan. 2004.
[43] P. Hoyer. (2004). *NMFpack (1.1st ed.)* [Online]. Available: http://www.cs.helsinki.fi/patrik.hoyer
[44] A. Cichocki and R. Zdunek. (2006). *NMFLAB for Signal Processing NMFLAB - MATLAB Toolbox for Non-Negative Matrix Factorization* [Online]. Available: http://www.bsp.brain.riken.jp/ICALAB/nmflab.html
[45] A. Hyvarinen, "The fixed-point algorithm and maximum likelihood estimation forindependent component analysis," *Neural Process. Lett.*, vol. 10, no. 1, pp. 1–5, Aug. 1999.
[46] H. Shen, M. Kleinsteuber, and K. Huper, "Local convergence analysis of fastICA and related algorithms," *IEEE Trans. Neural Netw.*, vol. 19, no. 6, pp. 1022–1032, Jun. 2008.
[47] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, Dec. 1996.

Guoxu Zhou was born in Hubei, China, in 1977. He received the Ph.D. degree in intelligent signal and information processing from the South China University of Technology, Guangzhou, China, in 2010.

His current research interests include machine learning, intelligent information processing, blind signal processing, and tensor analysis.

Shengli Xie (M'01–SM'02) was born in Hubei, China, in 1958. He received the M.S. degree in mathematics from Central China Normal University, Wuhan, China, in 1992, and the Ph.D. degree in control theory and applications from the South China University of Technology, Guangzhou, China, in 1997.

He is currently a Full Professor with the South China University of Technology and a Vice-Head of the Institute of Automation and Radio Engineering. He is the author or co-author of two books and more than 80 scientific papers published in journals and conference proceedings. His current research interests include automatic control, blind signal processing, and machine learning.

Prof. Xie won the Second Prize in the National Natural Science Award in 2009.

Zuyuan Yang (M'92) was born in Hubei, China, in 1982. He received the Ph.D. degree in signal and information processing from the South China University of Technology, Guangzhou, China, in 2010.

He is currently a Post-Doctoral Fellow with the School of Electronics and Information Engineering, South China University of Technology. His current research interests include machine learning, blind signal processing, image processing, and information security.

Jun-Mei Yang was born in Shandong, China, in 1979. She received the M.S. degree in cybernetics from the Chinese Academy of Sciences, Beijing, China, in 2005, and the Ph.D. degree in systems science from the Graduate School of Informatics, Kyoto University, Kyoto, Japan, in 2008.

She is currently a Lecturer with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. Her current research interests include system identification, blind signal processing, and adaptive noise cancellation.

Zhaoshui He was born in Hunan, China. He received the B.A. degree in applied mathematics from Hunan Normal University, Hunan, in 2000, and the Ph.D. degree in electronics and information engineering from the South China University of Technology, Guangzhou, China, in 2005.

He is currently a Faculty Member with the Laboratory for Intelligent Information Processing (LIIP), Guangdong University of Technology, Guangzhou. Before joining LIIP, he was a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan. His current research interests include blind signal processing, sparse representation, model selection, and clustering and their application.