Robust Asymmetric Nonnegative Matrix Factorization

Hyenkyun Woo and Haesun Park

Abstract

The problems that involve separation of grouped outliers and low rank part in a given data matrix have attracted a great attention in recent years in image analysis such as background modeling and face recognition. In this paper, we introduce a new formulation called ℓ_∞-norm based robust asymmetric nonnegative matrix factorization (RANMF) for the grouped outliers and low nonnegative rank separation problems. The main advantage of ℓ_∞-norm in RANMF is that we can control denseness of the low nonnegative rank factor matrices. However, we also need to control distinguishability of the column vectors in the low nonnegative rank factor matrices for stable basis. For this, we impose asymmetric constrains, i.e., denseness condition on the coefficient factor matrix only. As a byproduct, we can obtain a well-conditioned basis factor matrix. One of the advantages of the RANMF model, compared to the nuclear norm based low rank enforcing models, is that it is not sensitive to the nonnegative rank constraint parameter due to the proposed soft regularization method. This has a significant practical implication since the rank or nonnegative rank is difficult to compute and many existing methods are sensitive to the estimated rank. Numerical results show that the proposed RANMF outperforms the state-of-the-art robust principal component analysis (PCA) and other robust NMF models in many image analysis applications.

Index Terms

Nonnegative matrix factorization, Nonnegative rank, Nonnegative nuclear norm, Outliers, Robustness, Principal component analysis, Background modeling, Face recognition, Regularization

H. Woo is with School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Republic of Korea; e-mail: hyenkyun@kias.re.kr. and H. Park is with School of Computational Science and Engineering, Georgia Institute of Technology, USA; e-mail: hpark@cc.gatech.edu

I. INTRODUCTION

A high dimensional nonnegative data matrix, such as video image sequence or images of faces, generally resides in a low dimensional subspace but are often corrupted by grouped outliers, i.e., unavoidable artifacts such as moving objects in static background and shadows in face images under varying illumination condition. Therefore, it is important in various image analysis applications, such as background modeling and face recognition, to separate grouped outliers and the inherent nonnegative low rank structure of the given high dimensional image data.

Let $A = [a_1, ..., a_n]$ be a $m \times n$ nonnegative high dimensional image data matrix, where each column vector a_i corresponds to a column-wise stacked image frame. In this paper, we are interested in the problem of separating A into a low nonnegative rank matrix L and grouped outliers X:

$$A = L + X, \tag{1}$$

where $L = W\Lambda H^T \in \mathcal{L}_+(r) = \{W\Lambda H^T : W \in \mathbb{B}_2^{m\times r}, H \in \mathbb{B}_2^{n\times r}, \text{ and } \Lambda \in \mathbb{S}_{>0}^{r\times r}\}$ is a low nonnegative rank matrix, $W = [w_1, \ldots, w_r]$ is the nonnegative basis matrix of L, $H = [h_1, \ldots, h_r]$ is the associated nonnegative coefficients matrix of L, and Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i$ which we call as an asymmetric singular value of L. Here, $\mathbb{S}_{>0}^{r\times r} = \{\text{diag}(\lambda_1, \ldots, \lambda_r) \ : \ \lambda_i > 0 \ \}$ and $\mathbb{B}_2^{d\times r} = \{B = [b_1, \ldots, b_r] \ : \ b_i \in \mathbb{R}_+^d, \|b_i\|_2 = 1, \text{ and } \det(B^TB) \neq 0\}$. Note that $\mathbb{R}_+$ is the set of nonnegative real numbers, $\|x\|_p = (\sum_j x_i^p)^{1/p}$ for $p > 0$, and $\|x\|_0 = \#\{x \neq 0\}$ for $p = 0$. When Λ is subsumed into W or H, we get a typical nonnegative matrix factorization (NMF) formulation [1], [28]. We define (reduced) nonnegative rank as follows:

$$\text{rank}_+(L) = \arg\min_{\tau}\{\tau \ : \ L \in \mathcal{L}_+(\tau)\}. \tag{2}$$

Although (2) is a reduced version of the nonnegative rank, in this paper we call it as the nonnegative rank for simplicity. See [19], [42], [12], for more details on the nonnegative rank and its various theoretical properties. In general, the determination of the nonnegative rank of a matrix is a NP-hard problem [42]. Therefore, in many applications of NMF, the nonnegative rank parameter r is overestimated and is empirically selected based on the given data A. The main goal of this paper is to find a solution of (1) with overestimated r.

The following is a typical NMF model with ℓ_1-norm cost function (ℓ_1-NMF) [23], [25], [26], [43]:

$$\min_L\{ \ \|A - L\|_{\ell_1} \ : \ L \in \mathcal{L}_+(r) \ \}. \tag{3}$$

Note that we will use, for $Y \in \mathbb{R}^{m\times n}$, $||Y||_{l_p} = (\Sigma_{ij}|Y_{ij}|^p)^{1/p}$ with $p > 0$ and $||Y||_{l_0} = \#\{Y \neq 0\}$ with $p = 0$ throughout the paper to distinguish it from the conventional induced matrix norm[1]. The main advantage of ℓ_1-NMF (3) is that it is robust to outliers when compared to Frobenious norm based NMF [28]. However, ℓ_1-NMF (3) has an overfitting problem (e.g. see Figure 3) and an identifiability problem [2], [8], epecially when r is severely overestimated.

Note that matrix factorization based model [23] (without nonnegative constraints) is also suffer from an unavoidable overfitting problem and an identifiability problem [2], [8]. In various minimization problems with rank constraints, including matrix completion problem [3], [4], nuclear norm [14] is a typical regularization method to overcome overfitting and identifiability problems. The following is the nuclear norm based robust PCA (RPCA) model [2], [8] for sparse outliers and low rank separation problems.

$$(\hat{L}, \hat{X}) = \underset{L,X}{\arg\min} \ \{ \ \nu\|X\|_{\ell_1} + \|L\|_* \ : \ A = L + X \ \}, \tag{4}$$

where $\|L\|_* := \sum_i \sigma_i(L)$ denotes the nuclear norm of the matrix L and $\sigma_i(L)$ is the i-th largest singular value of the matrix L. Let us assume that $A = L_0 + X_0$ (L_0 is the true low rank solution and $X_0 = A - L_0$). The convexity of nuclear norm is fully utilized to guarantee exact separability, i.e. $\hat{L} = L_0$ and $\hat{X} = X_0$, under the incoherence condition (see Appendix A), although we fix the regularization parameter $\nu = 1/\sqrt{\max\{m, n\}}$ [2] or choose it in some specific region [8]. Here, sparse outliers X in (4) is assumed to have sufficient random structures or bounded sparsity. See Appendix A for more details.

A. Nonnegative Matrix Factorization for Column Outliers Detection

Let us start with the definition of structured grouped outliers, which we want to separate from the inherent low nonnegative rank matrix.

Definition I.1. *Let $X \in \mathbb{R}^{m\times n}$ be grouped outliers with limited row sparsity, i.e., $\max_i \|row_i(X)\|_0 < \zeta n$, then we call X as column outliers. Here, $0 < \zeta < \frac{1}{2}$ decides sparsity level of X in row direction.*

In video applications such as background modeling in Figure 7, moving objects correspond to column outliers with small ζ. Note that column outliers do not have sparsity limit in column direction and thus it covers dynamically changing outliers, such as in office scenario in Figure 7. Although RPCA (4) does not always guarantee separability between column outliers and low rank matrix (see Appendix A), when

[1] $\|Y\|_{p,q} = \max_{\|x\|_p=1} \|Yx\|_q$, where $p, q \geq 1$.

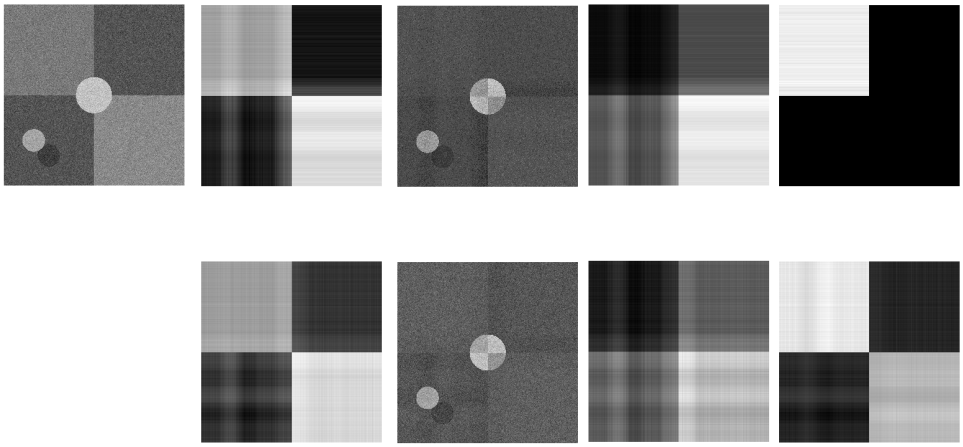

Fig. 1: Top row from left to right: the data matrix $A \in \mathbb{R}_+^{1000\times 1000}$, low nonnegative rank matrix L, column outliers X, the first rank one matrix ($w_1 h_1^T$), and the second rank one matrix ($w_2 h_2^T$) of the proposed robust asymmetric NMF model (5) (see also (22)) with ℓ_p-norm ($p = 0.65$) cost function. Bottom row from left to right: low rank matrix, column outliers, the first rank one matrix ($u_1 v_1^T$), and the second rank one matrix ($u_2 v_2^T$) of RPCA [2]. The vectors (w_2 and h_2) in the second rank matrix of robust asymmetric NMF is approximately $\frac{1}{\sqrt{k}}$-dense [22] with $k = 500$. Note that we tuned regularization parameter $\nu = \frac{0.15}{\sqrt{1000}}$ of RPCA (4) for best performance. However, we fixed rank parameter $r = 20$ of robust asymmetric NMF, although true (nonnegative) rank is two. Black is the minimum value and white is the maximum value.

we tune the parameter ν in (4), we still can separate column outliers as observed in Figure 1. See also [34] and Section IV for more details. However, when we have a priori information about the structure of outliers, such as some column vectors are totally corrupted then we can use $\ell_{2,1}$-norm [41], instead of simple ℓ_1-norm. In general, we do not have such a strong a priori outliers information, thus it is required to use other methods to separate column outliers.

Now, let us consider the following generic robust NMF formulation for separation problem between low nonnegative rank matrix L and column outliers X:

$$\min_{L,X}\{\ \Phi(X) + \frac{\alpha}{2}\|A - L - X\|_F^2\ :\ \mathcal{R}(L) \leq \tau\ \text{ and } 0 < L \leq B_L\ \}, \tag{5}$$

where $L = W\Sigma H^T \in \mathcal{L}_+(r)$, Φ is a sparsity enforcing function, such as ℓ_p-norm ($0 < p \leq 1$) or log-function [5], [30], $B_L = 255$ for image data, $\mathcal{R}(L)$ is a low rank$_+$ (i.e. nonnegative rank) enforcing

function such as nonnegative nuclear norm [12], and $\tau \in \mathbb{R}_+$ is a rank$_+$ constraint parameter.

In this paper, we mainly study the various properties of ℓ_∞-norm based rank$_+$ constraint $\mathcal{R}(L)$ for column outliers and low rank$_+$ separation problem. In Figure 1, we show a typical difference between rank$_+$ and rank for column outliers and low rank$_+$ (or rank) separation problems. All rank one matrices $u_i v_i^T$ of rank (Bottom row of Figure 1) can be dense. However, each factor w_i or h_i of rank$_+$ matrix can be $\frac{1}{\sqrt{k}}$-dense or k-subspace dense with minimum sparsity k. As commented in [22], a vector $v \in \mathbb{R}_+^d$ is $\frac{1}{\sqrt{k}}$-dense or k-subspace dense if $\|v\|_0 = k$ and $v = \frac{1}{\sqrt{k}}\iota_v$ with ι_v is an indicator vector of nonzero elements of v. Therefore, to separate column outliers and low rank$_+$, at least $\|h_i\|_0 > \zeta n$ should be satisfied. For instance, row factor vector h_2 of the second rank matrix $w_2 h_2^T$ in Figure 1 has $\|h_2\|_0 = 500$ and sparsity level of column outliers X is $\max_i \|row_i(X)\|_0 = 0.2n$.

Now, we have two important concerns for the robust NMF model (5). First, an appropriate regularization method is required to reduce the effect of overfitting problems. Second, we also need to generate well-conditioned k-subspace dense low rank$_+$ matrix with large k to separate column outliers efficiently. For these two problems, we will introduce ℓ_∞-norm based asymmetric nonnegative nuclear norm, which fully utilizes matrix factorization (or dictionary learning) structure for nonnegative data; the coefficient matrix H for denseness to separate column outliers and the basis matrix W for stability of the low rank$_+$ matrix L. We also introduce a soft regularization method to find a solution of the proposed asymmetric robust NMF model. The main advantage of the proposed low rank$_+$ enforcing soft regularization framework is that it is less sensitive to the rank selecting regularization parameter when compared to the conventional hard constraints; regularization parameter ν in nuclear norm based model (4) or rank selection parameter r in matrix factorization based model (3). We evaluate performance of the proposed method for the background modeling of video image sequence and removal of shadows and grossly corrupted artifacts in face images. Although we fix all parameters of the proposed model, the numerical results show that our proposed method outperforms the state-of-the-art nuclear norm based RPCA [2], [8], DECOLOR [44] and other matrix factorization based approaches such as ℓ_1-NMF (3). Moreover, the basis matrix W generated by the proposed method is more interpretable than that of the nuclear norm based model (4).

The paper is organized as follows. In Section II, we study the various properties of the proposed asymmetric nonnegative nuclear norm, such as stability, denseness, and distinguishability. In Section III, we introduce the proposed robust asymmetric nonnegative matrix factorization (RANMF) with soft regularization framework. In Section IV, we report our numerical results on background modeling of video image sequence and removal of grossly corrupted artifacts in face images under varying illumination condition. We give our conclusions in Section V. In Appendix, we review incoherence condition of RPCA

(4) and describe lower bound of the nonnegative rank.

II. Asymmetric Nonnegative Matrix Factorization

In this section, we introduce stability, denseness, and distinguishability of ℓ_∞-norm based asymmetric nonnegative nuclear norm.

First, let us consider the following sliced unit ℓ_2-norm sphere:

$$Z_\eta^d = \{v \in \mathbb{R}_+^d \ : \ \|v\|_2 = 1, \|v\|_\infty \leq \eta\}, \tag{6}$$

where $\eta \in [\frac{1}{\sqrt{d}}, 1]$. Note that if $\eta \approx \frac{1}{\sqrt{d}}$ then a vector $v \in Z_\eta^d$ is sufficiently away from the standard coordinate vector $e_i \in \mathbb{R}^d$ for all $i = 1, ..., d$. This sliced unit sphere (6) is a fundamental unit block of the proposed robust asymmetric NMF model. Now, we introduce asymmetric nonnegative nuclear norm[2] (ANN-norm) as a relaxation of rank_+:

Definition II.1. *For a given matrix $L \in \mathbb{R}_+^{m\times n}$ with $r = \text{rank}_+(L)$, we define asymmetric nonnegative nuclear norm of L as follows:*

$$\|L\|_\diamond = \arg\min\{ \sum_i \lambda_i \ : \ L = W\Lambda H^T \in \mathcal{L}_\mathbb{Z}(r) \ \}, \tag{7}$$

where

$$\mathcal{L}_\mathbb{Z}(r) = \{ \ W\Lambda H^T \ : \ W = [w_1, ..., w_r] \in \mathbb{Z}_{\eta_W}^{m\times r}, \ H = [h_1, ..., h_r] \in \mathbb{Z}_{\eta_H}^{n\times r}, \ \Lambda \in \mathbb{S}_\rightarrow^{r\times r} \ \}$$

with

- $\mathbb{Z}_\eta^{d\times r} = \{ \ V = (v_1, ..., v_r) \ : \ v_i \in Z_{\eta_i}^d, \ \eta = \max\{\eta_1, ..., \eta_r\}, \ det(V^TV) \neq 0, \ and \ \eta_i \in \mathbb{R}_{>0} \ \}$
- $\mathbb{S}_\rightarrow^{r\times r} = \{ \ \Lambda = diag(\lambda_1, ..., \lambda_r) \ : \ \lambda_i \geq \lambda_j > 0, \ for \ i < j \ \}$.

We simply call the following matrix factorization with respect to ANN-norm of L (7) as asymmetric NMF (ANMF):

$$L = W\Lambda H^T = \sum_{i=1}^r \lambda_i w_i h_i^T, \tag{8}$$

where $\lambda_i \geq \lambda_j > 0 \ (i < j), \ w_i \in Z_{\eta_{w_i}}^m, \ and \ h_i \in Z_{\eta_{h_i}}^n$.

Note that λ_i is an asymmetric singular value with respect to ANN-norm, and $\eta_{w_i} \in [\frac{1}{\sqrt{m}}, 1]$ and $\eta_{h_i} \in [\frac{1}{\sqrt{n}}, 1]$ are constants and control k-subspace denseness of basis and coefficient matrix. For robust

[2]Note that asymmetric nonnegative nuclear norm is not a matrix norm [1]. However, we called it as a norm since it is an extension of nuclear norm for the nonnegative rank.

NMF problem (5), small η_W and η_H are preferable for dense low rank$_+$ matrix. However it is not such a simple problem, because in this case, W and H will be ill-conditioned. The balance between denseness and stability should be decided for robust NMF (5). In addition, $\|L\|_\diamond$ is the Minkowski gauge function [1] of the set $\{wh^T \ : \ w \in Z_{\eta_w}^m, \ h \in Z_{\eta_h}^n\}$ and it thus is nonnegative, homogeneous, and convex with respect to rank one matrix. If we set $\eta_{w_i} = \eta_{h_i} = 1$ for all i, then $\|L\|_\diamond$ becomes the following nonnegative nuclear norm (9):

$$\|L\|_*^+ = \arg\min\{ \sum_{i=1}^{r} \lambda_i \ : \ L \in \mathcal{L}_+(r) \}. \tag{9}$$

Note that the nonnegative nuclear norm is a natural extension of the nuclear norm [14] for the nonnegative rank. However, it is not easy to find a global minimum of nonnegative nuclear norm [12], since we need to find a solution of completely positive matrix factorization of L which is another NP-hard problem [12]. Also, if we replace sliced unit sphere Z_η^d in (7) with unit ℓ_∞-norm ball $B_\infty(d) = \{v \in \mathbb{R}^d \ : \ \|v\|_\infty \leq 1\}$, then $\|L\|_\diamond$ is approximately equal to the γ_2-norm relaxation of rank:

$$\|L\|_{\gamma_2} \approx \arg\min\{\sum_i \lambda_i \ : \ L = \sum_i \lambda_i w_i h_i^T, \ w_i \in B_\infty(m), \ h_i \in B_\infty(n) \ \}, \tag{10}$$

where the approximation is bounded by Grothendieck constant [27]. It is well known that γ_2-norm based low rank approximation shows better performance in matrix completion problem though random sampling is inhomogeneous [38], [27].

Before we go further, we introduce a useful inverse square-root relation between ℓ_0-norm and ℓ_∞-norm in the sliced unit ℓ_2-norm sphere Z_η^d (6):

Lemma II.1. *Let $v \in Z_\eta^d$, then*

$$\frac{1}{\sqrt{\|v\|_0}} \leq \|v\|_\infty \leq \eta. \tag{11}$$

Also, we have $\mathbf{E}_0(v) \leq \eta$, where $\mathbf{E}_0(v)$ is a mean value of nonzero elements of v, i.e., $\frac{\|v\|_1}{\|v\|_0}$. Note that if $\|v\|_0 \in [\frac{1}{\eta^2}, 1 + \frac{1}{\eta^2})$, then v is the sparsest vector in Z_η^d. If $\|v\|_0 = \frac{1}{\eta^2}$ then v is η-dense or $\frac{1}{\eta^2}$-subspace dense.

Proof: For any $v \in Z_\eta^d$, we have the following inequalities

$$\langle v, \ \frac{1}{\sqrt{\|v\|_0}}\mathbf{1}\rangle \leq 1 \leq \langle v, \ \|v\|_\infty \mathbf{1}\rangle \leq \langle v, \ \eta\mathbf{1}\rangle,$$

where $\mathbf{1}$ is the all one vector in the dimension implied in the context. Note that we can replace $\mathbf{1}$ with ι_v. The first inequality follows from Cauchy-Schwartz inequality. The second is trivial, since $\forall v \in Z_\eta^d, v \leq$

$\|v\|_\infty \iota_v \leq \|v\|_\infty \mathbf{1}$ and the third follows by the definition of Z_η^d (6). Therefore, we have

$$\frac{1}{\sqrt{\|v\|_0}} \leq \|v\|_\infty \leq \eta,$$

since $\|v\|_1 > 0$ for any $v \in Z_\eta^d$. Also, from the Hölder inequality $\|v\|_1 \leq \|v\|_\infty \|v\|_0$, we get

$$\mathbf{E}_0(v) \leq \eta.$$

Since $\|v\|_0 \geq \frac{1}{\eta^2}$, there is a vector $v \in Z_\eta^d$ with the smallest sparsity $\|v\|_0 \in [\frac{1}{\eta^2}, 1 + \frac{1}{\eta^2})$. Note that such a vector v with the smallest sparsity is not always unique (see Theorem II.9). When $\frac{1}{\eta^2}$ is integer, $v \in Z_\eta^d$ is η-dense with $\eta = \frac{1}{\sqrt{\|v\|_0}}$. ■

In the following theorem, we show that an asymmetric singular value λ_i of ANMF (8) is roughly in the same order of square root of area of the corresponding rank one matrix.

Theorem II.2. *Let $L = \sum_{i=1}^r L_i$ be an ANMF* (8) *with $L_i = \lambda_i w_i h_i^T$. Then*

$$\mathbf{E}_0(L_i)\sqrt{\|L_i\|_{\ell_0}} \leq \lambda_i \leq B_{L_i}\sqrt{\|L_i\|_{\ell_0}}, \tag{12}$$

where $B_{L_i} = \|L_i\|_{\ell_\infty}$.

Proof: For the lower bound, since $\|L_i\|_{\ell_1} \leq \lambda_i \sqrt{\|L_i\|_{\ell_0}}$, we get

$$\mathbf{E}_0(L_i) = \frac{\|L_i\|_{\ell_1}}{\|L_i\|_{\ell_0}} \leq \frac{\lambda_i}{\sqrt{\|L_i\|_{\ell_0}}}.$$

For the upper bound, since $\|L_i\|_{\ell_\infty} \leq B_{L_i}$, we get

$$\lambda_i = \|L_i\|_F \leq B_{L_i}\sqrt{\|L_i\|_{\ell_0}}.$$

■

Note that λ_i is roughly in proportion to the square root of area of L_i. Therefore, it is natural to reorder each rank-one matrix L_i of ANMF (8) in descending order of λ_i-value to enforce dense low nonnegative rank structure. Based on this observation, we give two different lower bounds of $\text{rank}_+(L)$ in Appendix B.

A. Stability and denseness of the basis (coefficient) matrix of ANMF

In this section, we study stability, denseness, and δ-distinguishability of the basis/coefficient matrix of ANMF (8). Based on δ-distinguishability, we can clearly understand the relation between stability and denseness of the basis/coefficient matrix. Although we only consider the basis matrix $W \in \mathbb{Z}_{\eta_W}^{m\times r}$ of ANMF (8), the analysis of the coefficient matrix $H \in \mathbb{Z}_{\eta_H}^{n\times r}$ is same except dimension $n \times r$.

Definition II.2. *Let us assume that*

$$W \in \mathbb{Z}_{\eta_W}^{m\times r} = \{W = (w_1, ..., w_r) \ : \ w_i \in Z_{\eta_{w_i}}^m, \ \eta_W = \max\{\eta_1, ..., \eta_r\}, \ det(W^TW) \neq 0\}.$$

To measure stability and denseness of a matrix W, we define the following three parameters:

- *Stability (small is stable and large is unstable)*
  - $\|W\| = \sqrt{\rho_{max}(W^TW)} \in [1, \sqrt{r})$
  - $S(W) = \|W^TW - I\|_{\ell_\infty} \in [0, 1)$
- *Denseness (small is dense and large is sparse)*
  - $\eta_W = \|W\|_{\ell_\infty} \in (\frac{1}{\sqrt{m}}, 1]$

Here, $\rho(B) \in [\rho_{min}(B), \rho_{max}(B)]$ is the eigenvalue of B. The range of $S(W)$ and η_W is obvious. We describe the range of $\|W\|$ in the following Lemma II.3.

Lemma II.3. *Let $W \in \mathbb{Z}_1^{m\times r}$, then $\rho_{max}(W^TW) \in [1, 1 + (r-1)S(W)]$ and $\rho_{min}(W^TW) \in (0, 1]$. In addition, if $\rho_{max}(W^TW) = 1$ (or $\rho_{min}(W^TW) = 1$) then $S(W) = 0$.*

Proof: Since $W \in \mathbb{Z}_1^{m\times r}$, the gram matrix W^TW is positive definite and $Tr(W^TW) = r$. For the lower bound, let $\rho_{max}(W^TW) < 1$ then $Tr(W^TW) < r$. It contradicts $Tr(W^TW) = r$. Hence $\rho_{max}(W^TW) \geq 1$. For the upper bound, we use the Gershgorin circle theorem [20]:

$$\rho_{max}(W^TW) \in \cup_{i=1}^r D_i$$

where $D_i = \{b \in \mathbb{R}_+ \ : \ |b-1| \leq \sum_{j\neq i} w_i^Tw_j\}$. Since $w_i^Tw_j \leq S(W) < 1$ for all $i \neq j$, we get

$$\rho_{max}(W^TW) \leq 1 + (r-1)S(W).$$

In addition, since $Tr(W^TW) = r$, when $\rho_{max}(W^TW) = 1$, we get $\rho_{min}(W^TW) = 1$, i.e., all eigenvalues are one. Therefore, since $\|W^TW\|_F^2 = r$ and $w_i^Tw_j \geq 0$, we get $S(W) = 0$.

Note that we could show $\rho_{min}(W^TW) \leq 1$ in the same way. The lower bound $\rho_{min}(W^TW) > 0$ follows by definition of $\mathbb{Z}_1^{m\times n}$. Also, $\rho_{min}(W^TW) = 1 \rightarrow S(W) = 0$ is trivial, since $\rho_{min}(W^TW) = 1$ means $\rho_{max}(W^TW) = 1$. ■

As observed in the Lemma II.3, we could use $\rho_{min}(W^TW)$ as a measure of stability of W. However, $\rho_{max}(W^TW)$ is more appropriate, since it reveals the relation between stability of W and rank parameter r explicitly as expressed in the following Lemma II.4 and Remark II.5. Note that, regarding $\|W\|$, stability of W corresponds to one side of condition number $cond(W) = \sqrt{\frac{\rho_{max}(W^TW)}{\rho_{min}(W^TW)}}$.

Lemma II.4. *Let $W = [w_1, ..., w_r] \in \mathbb{Z}_{\eta_W}^{m\times r}$ and $w_i^T w_j = S(W)$ for all $i \neq j$, then there are only two eigenvalues of W^TW:*

$$\rho(W^TW) = \{1 - S(W), 1 + (r-1)S(W)\}.$$

Proof: As observed in [16, Lemma 5.11], due to the symmetry, it is easy to check that all one vector $\mathbf{1}$ is the eigenvector of the gram matrix W^TW and the corresponding eigenvalue is $1+(r-1)S(W)$. Also, $r-1$ linearly independent vectors $[1, -1, ..., 0]^T, ..., [1, 0, ..., 0, -1]^T$ are eigenvectors for the eigenvalue $1 - S(W)$. ■

For general case, the estimation of the largest eigenvalue of W^TW are done by many researchers [31]. The following is a typical example of the estimation of $\|W\|$, which is useful for the analysis of our model.

Remark II.5. *As observed in [31, Theorem 1.4, Page 30], by using Perron-Frobenious Theorem with the assumption that $\epsilon_W = \min_{i,j} w_i^T w_j > 0$ for all i, j, we have tight bound of $\|W\|$ as follows:*

$$\sqrt{\min_i c_i + \epsilon_W \left(\frac{1}{a} - 1\right)} \leq \|W\| \leq \sqrt{\max_i c_i - \epsilon_W(1-a)}, \tag{13}$$

where c_i is the sum of i-th column elements of W^TW and $a = \sqrt{\frac{\min_i c_i - \epsilon_W}{\max_i c_i - \epsilon_W}}$. When $\epsilon_W = S(W)$, we get $\|W\| = \sqrt{1 + (r-1)S(W)}$, which is a boundary of $\|W\|$ in Lemma II.3. Note that stability of W is defined by $\|W\|$ or $S(W)$; see Definition II.2. However, $\|W\|$ is more robust measure of stability of W, since it is related to all elements of W by (13)*. On the contrary, $S(W) = \|W^TW - I\|_{\ell_\infty}$ depends on one maximum value and thus it is sensitive to noise.*

Note that if $S(W) = 0$, i.e., $\|W\| = 1$, then we get a sufficiently stable orthonormal basis nonnegative matrix. However, denseness of W is limited.

Lemma II.6. *Let $W \in \mathbb{Z}_1^{m\times r}$ be an orthonormal matrix. Then $\|W\|_{\ell_0} \leq m$ and $\min_i \|w_i\|_0 \leq m/r$. In addition, $\sqrt{\frac{r}{m}} \leq \|W\|_{\ell_\infty} \leq 1$.*

Proof: Since $w_i \in \mathbb{R}_+^m$ and $w_i^T w_j = 0$ for all $i \neq j$, $\iota_{w_i} \cap \iota_{w_j} = \emptyset$. Therefore, we have $\sum_{i=1}^r \|w_i\|_0 \leq m$ and thus $\min_i \|w_i\|_0 \leq m/r$. With Lemma II.1, we have

$$\sqrt{\frac{r}{m}} \leq \frac{1}{\sqrt{\min_i \|w_i\|_0}} \leq \|W\|_{\ell_\infty} \leq 1.$$

Note that upper bound is obtained when there is w_i such that $\|w_i\|_0 = 1$. ■

Example II.7. *Let* $W = [w_1, \ldots, w_r] \in \mathbb{R}_+^{m\times r}$ *with* $w_i \in Z^m_{\frac{1}{\sqrt{m}}}$ *then we get* $w_i = \frac{1}{\sqrt{m}}\mathbf{1}$ *for all* $i = 1, ..., r$ *and* $\|W\| = r$ *(see Lemma II.4).*

As observed in Example II.7, in a nonnegative set, we cannot obtain stable and fully dense basis matrix simultaneously. If we put stronger dense condition on W, then the basis matrix W becomes more unstable, i.e., all column vectors are not distinguishable. Now, we introduce δ-distinguishability [17] of Z^m_η (6). Based on δ-distinguishability, we can clearly understand the relation between stability, (i.e., $S(W)$ and $\|W\|$) and denseness, (i.e., $\|W\|_{\ell_\infty}$).

Definition II.3. *Let* Z^m_η *be a set in* (6)*. If there exist* $w^i, w^j \in Z^m_\eta$ *such that*

$$\|w^i - w^j\|_2 \geq \delta,$$

then we say that Z^m_η *is* δ*-distinguishable [17]. Also, it relates to the* $\frac{\delta}{2}$*-net (here, we mean a* $\frac{\delta}{2}$*-radius ball whose center is in* Z^m_η *and it does not intersect with other* $\frac{\delta}{2}$*-radius balls) and the cardinality of* $\frac{\delta}{2}$*-net of* Z^m_η *is defined as* $N(\frac{\delta}{2}, Z^m_\eta)$*. Therefore,* Z^m_η *is* δ*-distinguishable if and only if* $N(\frac{\delta}{2}, Z^m_\eta) \geq 2$*. In addition, we say that a basis matrix* $W \in \mathbb{Z}^{m\times r}_{\eta_W}$ *in ANMF* (8) *is a* δ*-distinguishable matrix when* $\|w_i - w_j\|_2 \geq \delta$ *for all* $w_i, w_j \in W = [w_1, ..., w_r]$ *with* $i \neq j$*.*

Lemma II.8. *Let* $W \in \mathbb{Z}^{m\times r}_{\eta_W}$ *be a* δ*-distinguishable matrix. Then we get*

$$0 < \delta \leq \sqrt{2(1 - S(W))}.$$

Therefore, when $\delta > \sqrt{2}$*,* $\mathbb{Z}^{m\times r}_{\eta_W}$ *does not have any* δ*-distinguishable matrix.*

Proof: Since $W = [w_1, w_2, ..., w_r] \in \mathbb{Z}^{m\times r}_{\eta_W}$ is δ-distinguishable matrix, we get

$$S(W) = \max_{i\neq j} w_i^T w_j = 1 - \frac{1}{2}\|w_i - w_j\|_2^2 \leq 1 - \frac{\delta^2}{2}.$$

When $\delta > \sqrt{2}$, there is not any δ-distinguishable matrix in $\mathbb{Z}^{m\times r}_{\eta_W}$ for all η_W. Note that $Z^m_{\frac{1}{\sqrt{m}}}$ is not δ-distinguishable for any $\delta > 0$, since $Z^m_{\frac{1}{\sqrt{m}}}$ has only one element $\{\ \frac{1}{\sqrt{m}}\mathbf{1}\ \}$; see also Example II.7. ■

In the following theorem, we show the relation between δ-distinguishability and k-subspace denseness of $Z^m_{\eta_W}$. That is, we give a lower bound of $N(\frac{\delta}{2}, Z^m_{\frac{1}{\sqrt{k}}})$ [17].

Theorem II.9. *For* $1 \leq k \leq m$ *and* $0 < \delta \leq \sqrt{\frac{2}{k}}$*, we get*

$$N(\frac{\delta}{2}, Z^m_{\frac{1}{\sqrt{k}}}) \geq \begin{pmatrix} m \\ k \end{pmatrix}, \tag{14}$$

where $N(\varepsilon, Z^m_{\eta_W})$ *is the cardinality of a minimal* ε*-net of* $Z^m_{\eta_W}$*.*

Proof: By Lemma II.1, for any $x \in Z^m_{\eta_W}$, we have $\frac{1}{\sqrt{\|x\|_0}} \leq \eta_W$. Therefore, it is reasonable to consider $\frac{1}{\sqrt{k}}$-dense subspace

$$V_k = \{x = (x_1, ..., x_m)^T \in \mathbb{R}^m_+ \ : \ \|x\|_2 = 1, \|x\|_0 = k, \text{ and } x_i \in \{0, \frac{1}{\sqrt{k}}\}\} \subseteq Z^m_{\frac{1}{\sqrt{k}}}. \tag{15}$$

Here, the minimum distance between $v, w \in V_k$ is

$$\min_{v,w \in V_k} \|v - w\|_2 = \sqrt{\frac{2}{k}} \geq \delta.$$

Then, we get

$$N(\frac{\delta}{2}, Z^m_{\frac{1}{\sqrt{k}}}) \geq \text{Card } V_k = \begin{pmatrix} m \\ k \end{pmatrix},$$

where CardV_k is the number of elements in V_k. ■

Therefore, if we want more stable basis (i.e., large δ-distinguishable), then we need to sacrifice denseness of the basis matrix, i.e., $\frac{1}{\sqrt{k}}$-denseness with small k. However, we have enough potential candidate for basis matrix $W \in \mathbb{Z}^{m \times r}_{\eta_W}$, since $r \ll m$. In the following Example, we analyze the nontrivial worst case stability of basis matrix W.

Example II.10. *Let $W = [w_1, ..., w_r] \in \mathbb{Z}^{m \times r}_{\frac{1}{\sqrt{k}}}$ with $w_i \in V_k$ (15) for all $i = 1, ..., r$. Furthermore, we assume that $k < m$ and $\iota^T_{w_i} \iota_{w_j} = k - 1$ for all $i \neq j$. Then, by Lemma II.4, we have*

$$\rho(W^T W) = \{\frac{1}{k}, r - \frac{r-1}{k}\}$$

and

$$\sqrt{1 - S(W)} = \sqrt{\rho_{min}(W^T W)} = \|W\|_{\ell_\infty} = \frac{1}{\sqrt{k}}.$$

In addition, the condition number of W becomes

$$cond(W) = \sqrt{\frac{\rho_{max}(W^T W)}{\rho_{min}(W^T W)}} = \sqrt{rk - r + 1}.$$

That is, stability of W depends on rank parameter r and sparsity of each column vector w_i.

B. Asymmetric Incoherence Criterion

In this section, we introduce asymmetric incoherence criterion to measure how well NMF is constructed for outlier detection problems.

As observed in Section II-A, the basis matrix W or coefficient matrix H of ANMF (8) cannot be both dense and stable. Indeed, a balance between denseness and stability is strongly required. What's worse, we need to separate structured grouped outliers, i.e., column outliers, which frequently appear

in image analysis problems, such as background modeling and face recognition problems. That is, if objects stay in long time in small area in background modeling problems then it should be included into background objects (i.e., low rank$_+$ matrix). For instance, see Figure 9. Due to the inherent nonnegative constraints of NMF, these complicated conditions appear. However, surprisingly, by simply keeping ℓ_∞-norm constraints on the coefficient matrix H only, we could relax denseness and stability dilemma of NMF and solve column outliers and low rank$_+$ separation problems.

Now, let us assume that $\eta_H \ll 1$ of the coefficient matrix $H = [h_1, ..., h_r] \in \mathbb{Z}_{\eta_H}^{n\times r}$ for denseness of H and $\eta_W = 1$ of the basis matrix $W \in \mathbb{Z}_1^{m\times r}$. With this asymmetric constraints, we have the following properties of asymmetric NMF (8) for the low rank$_+$ and outliers separation problem (5):

- Stability of ANMF (8) depends on stability of W (i.e., $\|W\|$ or $S(W)$):

$$E_{ij} \leq w_i^T w_j, \quad \forall i \neq j,$$

  where $E_{ij} = \langle w_i h_i^T, w_j h_j^T \rangle$. Note that ANMF, $L = W\Lambda H^T$, is stable means that $\|E\|$ (or $S(E)$) is small.

- Denseness of each rank one matrix $w_i h_i^T$ of ANMF (8) depends on denseness of H (i.e., η_H):

$$\|w_i h_i^T\|_{\ell_0} = \|w_i\|_0 \|h_i\|_0 \geq \frac{1}{\eta_H^2},$$

  which follow from (11). Since $w_i \in Z_1^m$, a rank one matrix $w_i h_i^T$ can be thin structure in row direction (i.e., in h_i direction).

- Column outliers X need to satisfy the following condition

$$\max_i \|\text{row}_i(X)\|_0 \leq \zeta n < \frac{1}{\eta_H^2} \leq \min_i \|h_i\|_0.$$

  That is, at least $\eta_H < \frac{1}{\sqrt{\zeta n}}$ should be satisfied to separate column outliers and low rank$_+$ matrix. Therefore, small η_H is preferable.

Now, we introduce asymmetric incoherence criterion for ANMF (8) to see how well ANMF is constructed with stable basis W and dense coefficient H. Let us start with the well-known incoherence condition [3] (see also Appendix A):

$$\max_i \|P_H e_i\|_2 \leq \varepsilon_{inc_+}, \tag{16}$$

where $H = [h_1, ..., h_r]$ is a coefficient matrix of ANMF (8) and $P_H = H(H^T H)^{-1} H^T$ is a projection operator on subspace generated by H. Also, we get

$$\varepsilon_{inc_+} \in [\sqrt{\frac{r}{n}}, 1]. \tag{17}$$

This is equal to the incoherence condition in (29) of Appendix A. Note that r of (17) is $\text{rank}_+(L)$, however r of (29) is $\text{rank}(L)$. Therefore, since $\text{rank}_+(L) \geq \text{rank}(L)$, ANMF (8) is always worse in terms of incoherence condition (29). Note that if H is orthonormal matrix then it is stable $S(H) = 0$ with sparsity $\|H\|_{\ell_0} \leq n$.

Lemma II.11. *Let $H \in \mathbb{Z}_1^{n\times r}$ and $S(H) = 0$. Then*

$$\max_i \|P_H e_i\|_2 = \|H\|_{\ell_\infty}, \tag{18}$$

where $\|H\|_{\ell_\infty} \in \left[\sqrt{\frac{r}{n}}, 1\right]$. In this case, the size of column outliers is limited by rank parameter as $\zeta < \frac{1}{r}$.

Proof: Under the orthonormal conditions, we get $P_H = HH^T$ and $\iota_{h_i} \cap \iota_{h_j} = \emptyset$ for all $h_i, h_j \in H$ with $i \neq j$. Therefore,

$$\max_i \|P_H e_i\|_2 = \max_i \|H^T e_i\|_2 = \|H\|_{\ell_\infty}.$$

Also, $\|H\|_{\ell_\infty} \in [\sqrt{\frac{r}{n}}, 1]$ from Lemma II.6 and we get

$$\zeta < \frac{1}{n\|H\|_{\ell_\infty}^2} \leq \frac{1}{r}.$$

■

Under the orthonormal condition $S(H) = 0$, we recover incoherence condition (16) with $\|H\|_{\ell_\infty}$ as observed in Lemma II.11. However, it has a strong disadvantage for column outliers separation problems. For instance, if we set $r = 20$ then we can not separate column outliers in Figure 1, since $\zeta = 0.2$. That is, we cannot keep orthonormal constraint in nonnegative set and need to consider stability parameter for incoherence condition. Now, we introduce a new incoherence condition for the coefficient matrix H in the following.

Definition II.4. *For $H \in \mathbb{Z}_{\eta_H}^{n\times r}$, let*

$$\Xi(H) = \frac{\|H\|_{\ell_\infty}}{\|H\|}, \tag{19}$$

then we get $\Xi(H) \in (\frac{1}{\sqrt{rn}}, 1]$. It decides stability and denseness of a matrix H. If $\Xi(H)$ is large (close to one), then H is stable but sparse. However, if $\Xi(H)$ is small (close to $1/\sqrt{rn}$), then H is dense but unstable.

The condition (19) can be applied to the basis matrix W with different dimension parameter m. See Appendix A for the incoherence condition of RPCA (4). Note that, instead of $\|H\|$, we can use $S(H)$ to measure stability. However, $\|H\|$ is more appropriate parameter for incoherence condition $\Xi(H)$ (19), since $\|H\|$ is less sensitive to noise; see Remark II.5 and Table I for more details.

In the following, we introduce a measure of goodness of ANMF (8) for column outliers separation problem.

Definition II.5. *Let $L = \sum_{i=1}^{r} \lambda_i w_i h_i^T$ be an ANMF* (8) *with basis matrix $W = [w_1, ..., w_r]$ and coefficient matrix $H = [h_1, ..., h_r]$. We define asymmetric incoherence criterion of L as follows:*

$$\frac{1}{\sqrt{rn}} < aINC(L) = \frac{\Xi(H)}{\Xi(W)} < \sqrt{rm} \tag{20}$$

If H is dense and W is stable then $aINC(L)$ is relatively small and column outliers are not included into low rank$_+$ matrix L. However, if column outliers are in low rank$_+$ matrix L then W becomes dense and H becomes sparse. Empirically, we observe that this asymmetric incoherence criterion is well matched with numerical experiments. See Table II and Figure 6 for more details.

Example II.12. *Let $L = W\Lambda H^T = \sum_{i=1}^{r} \lambda_i w_i h_i^T$ be an ANMF* (8) *with ideal $\frac{1}{\sqrt{k}}$-dense vectors. That is, $W \in \mathbb{Z}^{m\times r}_{\frac{1}{\sqrt{k_W}}}$, $H \in \mathbb{Z}^{n\times r}_{\frac{1}{\sqrt{k_H}}}$, $\|w_i\|_0 = k_W$, $\|h_i\|_0 = k_H$ for all $i = 1, ..., r$, $k_W < m$, and $k_H < n$. Let us assume that $\iota_{w_i}^T \iota_{w_j} = k_W - 1$ and $\iota_{h_i}^T \iota_{h_j} = h_H - 1$ for worst case separability. Then, by the results in Example II.10, we get*

$$\Xi(W) = \frac{1}{\sqrt{rk_W - r + 1}} = \frac{1}{cond(W)} \quad \text{and} \quad \Xi(H) = \frac{1}{\sqrt{rk_H - r + 1}} = \frac{1}{cond(H)}.$$

Therefore, the asymmetric incoherence criterion (20) *of L becomes*

$$aINC(L) = \frac{cond(W)}{cond(H)} = \sqrt{\frac{rk_W - r + 1}{rk_H - r + 1}} \approx \sqrt{\frac{k_W}{k_H}} = \frac{\eta_H}{\eta_W}. \tag{21}$$

This result is interesting since the asymmetric incoherence criterion of $L = W\Lambda H^T$ is just the ratio of condition number of W and H. Also, it is approximately the ratio of ℓ_∞-norm bound of basis matrix W and coefficient matrix H, i.e., η_H/η_W. In other words, small $aINC(L)$ means that W stable and sparse, H unstable and dense, and thus we can separate column outliers well. That is, $\zeta \lesssim \frac{1}{nn_W^2 aINC(L)^2}$.

III. Robust Asymmetric NMF with soft regularization

In this section, we describe robust asymmetric NMF with soft regularization method and the connection to the foreground detection problem.

Now, we propose robust asymmetric NMF (RANMF) for column outliers and low rank$_+$ separation problems:

$$\min_{L,X}\{\ \Phi(X) + \frac{\alpha}{2}\|A - X - L\|_F^2\ :\ L = W\Lambda H^T, W \in \mathbb{Z}_1^{m\times r}, H \in \mathbb{Z}_{\eta_H}^{n\times r}, \Lambda \in \mathbb{S}_{\rightarrow}^{r\times r}\ \}, \tag{22}$$

where $\Phi(X) = \sum_{i,j} |X_{i,j}|^p,\ (0 < p \leq 1)$ or $\sum_{i,j} \log(\varepsilon + |X_{i,j}|)$ with $\varepsilon > 0$. We call this robust ANMF model (22) as RANMF; ℓ_p-RANMF if $\Phi(X)$ is ℓ_p-norm and log-RANMF if $\Phi(X)$ is log function. As observed in various areas, such as compressed sensing [5], [30], due to $(\cdot)^p$ and $\log(\cdot)$ function, grossly corrupted error are less penalized and thus the recovered low rank$_+$ matrix of the given image data is expected to be robust to column outliers. Also, η_H should be sufficiently small to guarantee the separability of column outliers, i.e., $\eta_H < \frac{1}{\sqrt{\zeta n}}$. However, it is not easy to choose appropriate $\eta_H = \max_i \eta_{h_i}$, since each η_{h_i} depends on sparsity level of each row factor vector, h_i, i.e., $\|h_i\|_0 \geq \frac{1}{\eta_{h_i}^2}$. For instance, see Figure 8. Therefore, we indirectly find a solution of (22) with the following soft regularization method.

A. Soft Regularization

In this section, we introduce an algorithmic description of a soft regularization method.

As observed in [1], we have various equivalent models of (22) due to the freedom in formulation of asymmetric singular value Λ. For the proposed soft regularization method, let $\bar{W}\bar{H} = W\Lambda H^T$ and we minimize (22) with respect to $\bar{W}$ and $\bar{H}$. Here, Λ is subsumed into $\bar{W}$ and $\bar{H}$ alternatively. We fully utilize this alternative selection of Λ into the soft regularization alternating minimization (SRAM) (23).

SRAM (Soft regularized asymmetric alternating minimization):

Given $\bar{W}^0$, $\bar{H}^0$, $\bar{X}^0$. Choose $\alpha > 0$. For $t = 0, 1, 2, ...,$

STEP 1: $X^{t+1} = \underset{X}{\operatorname{argmin}}\ \Phi(X) + \frac{\alpha}{2}\|A - X - \bar{W}^t(\bar{H}^t)^T\|_F^2$

STEP 2: $(\bar{W}^{t+\frac{1}{2}}, \bar{H}^{t+\frac{1}{2}}) = \underset{\bar{W}\geq 0, \bar{H}\geq 0}{\operatorname{argmin}}\ \{\|A - X^{t+1} - \bar{W}(\bar{H})^T\|_F^2\ :\ \|\bar{H}\|_{\ell_\infty} \leq B_H\}$

STEP 3: $(\bar{W}^{t+1}, \bar{H}^{t+1}) = BASIS(\bar{W}^{t+\frac{1}{2}}, \bar{H}^{t+\frac{1}{2}})$

with

$$(W, H\Lambda) = BASIS(\bar{W}, \bar{H})$$

satisfies the following conditions:

$$\begin{cases} W\Lambda H^T = \bar{W}\bar{H}^T \\ W \in \mathbb{Z}_1^{m\times r}, H \in \mathbb{Z}_{\eta_H}^{n\times r}, \Lambda \in \mathbb{S}_{\rightarrow}^{r\times r} \end{cases} \tag{23}$$

In the above SRAM (23), we use a reweighted iterative thresholding [5], [30] to find a solution X^{t+1} in STEP 1. For STEP 2, we modified 2r-BCD (block coordinate descent) framework [24]. It is also known as HALS (hierarchical alternating least square) [9]. Note that, in STEP 2, we put $\|\bar{H}\|_{\ell_\infty} \leq B_H$ with $B_H < \frac{1}{\sqrt{n}}$ then Λ moves into $\bar{W}^{t+\frac{1}{2}}$. In STEP 3, we move Λ back into $\bar{H}^{t+1}$ and reorder each rank one matrix by $\lambda_i \geq \lambda_j$ for $i < j$. It helps to generate dense nonnegative rank one matrix with

large asymmetric singular value. Note that the ℓ_∞-norm bound, η_H, is indirectly controlled with upper bound B_H in STEP 2. The main advantage of this approach is that η_H is automatically decided based on the recovered sparsity pattern of H, i.e., $\eta_H \geq \max_i 1/(\|h_i\|_0)^2$. Note that the result generated by $BASIS$ function is similar to that of SVD. The only difference is that each column vector of W and H is nonnegative unit ℓ_2-norm but not necessarily orthogonal.

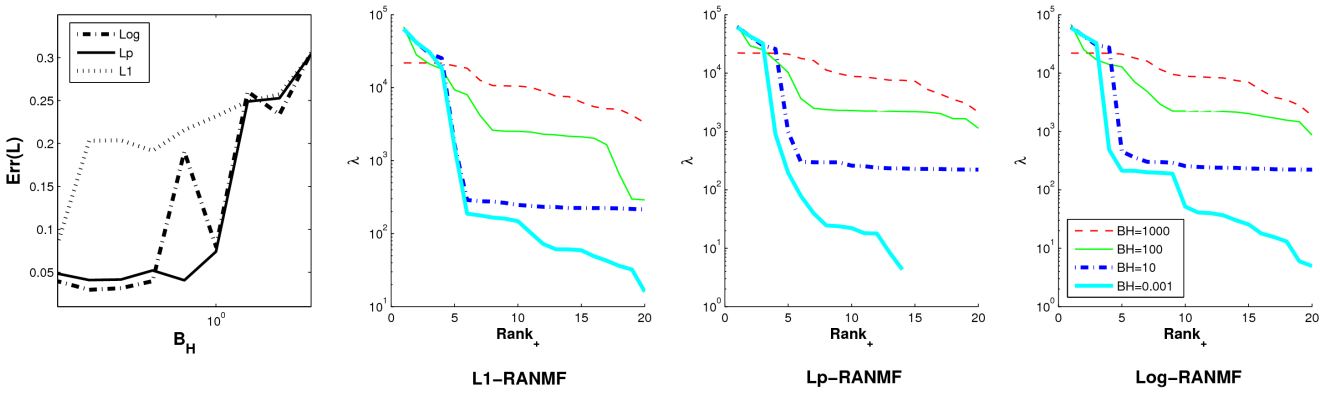

Fig. 2: From left to right: $err(L)$ (24) vs. B_H, diag(Λ) vs. B_H for ℓ_1-norm, ℓ_p-norm ($p = 0.65$), and log function. As B_H decreases, $diag(\Lambda)$ is getting more sharp and error decreases. That is, we minimize (22) indirect way with SRAM (23). The proposed ℓ_p-RANMF model shows best performance. Note that we use synthetic image in Figure 1. The rank parameter is fixed as $r = 20$. Although the given rank is relatively larger than the ground truth rank $r = 2$, the dominant rank is usually less then 5 when $B_H < \frac{1}{\sqrt{n}}$. The first graph shows that the performance is not sensitive to B_H, especially for ℓ_p-RANMF or log-RANMF.

In Figure 2, we analyze the effect of B_H on the performance of recovery of the low rank$_+$ matrix. As we decrease B_H, the recovered low rank$_+$ matrix is getting close to the ground truth and the graph of diag(Λ) is getting sharp. In other words, we minimize asymmetric nonnegative nuclear norm (7) indirect way to find a solution of (22). Note that, we can easily remove meaningless rank$_+$ matrix by simple thresholding. Figure 2 also shows that the selection of B_H is not sensitive. To escape from being stuck into the unwanted zero set, $B_H < \frac{1}{\sqrt{n}}$ is necessary, since $\max_i \|h_i\|_\infty \in [\frac{1}{\sqrt{n}}, 1]$ for the coefficient matrix $H = [h_1, ..., h_r]$. Note that the recovery error is

$$err(L) = \|L - L_0^+\|_F / \|L_0^+\|_F \tag{24}$$

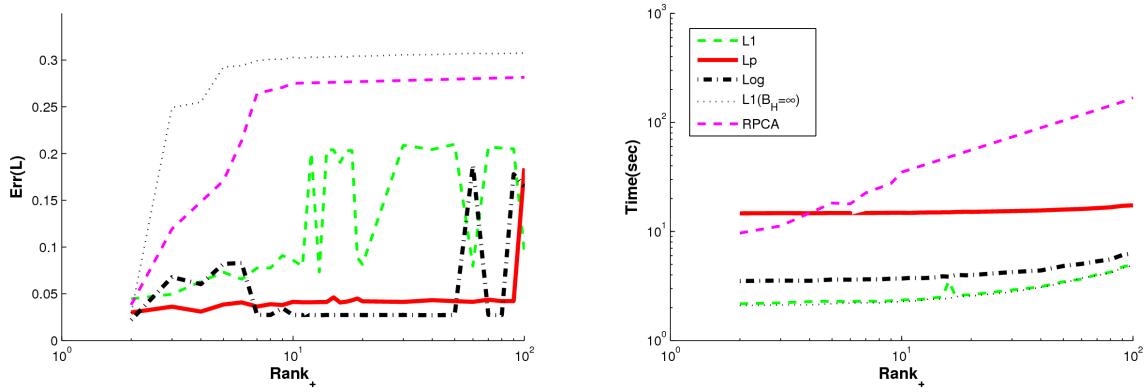

Fig. 3: A performance comparison for different choice of rank parameter. The proposed ℓ_p-RANMF model with $p = 0.65$ shows best performance irrespective of the choice of rank parameter r up to $r = 100$. But computational cost is relatively lager than other models except the RPCA model. Synthetic image in Figure 1 is used and $B_H = \frac{0.1}{\sqrt{n}}$ for RANMF. Note that we use the error function in (24).

for all models versus rank parameter r.

In Figure 3, we show the effect of rank parameter r for various model. The parameter ν in RPCA (4) is tuned to generate the corresponding rank. Since ℓ_1-NMF (3) does not have a regularization method for the nonnegative rank, it shows poor performance, except when $r = \text{rank}_+(L_0^+)$. Note that the performance of RPCA (4) is also sensitive to the choice of the parameter ν. The proposed ℓ_p-RANMF (22) is robust irrespective of choice of rank parameter r (up to $r = 100$). As we can see in Figure 2 and Figure 3, among various penalty functions ℓ_1-norm, ℓ_p-norm ($p < 1$) and log functions, ℓ_p-norm penalty function is stable and well recover low rank_+ matrix than any other models.

B. Connection to the Background modeling

In this subsection, we extend the proposed RANMF model (22) for foreground detection problems. Since the separated grouped outliers is not foreground mask and can be very noisy, e.g. tree scenario in Figure 7, we may need additional denoising/segmentation process to detect foreground mask. Among the various denoising process, we propose to use total variation (TV) [37]. The following is the general framework for foreground detection problem [44]:

$$\min_{X,L,\phi} \Phi(X) + \frac{\alpha}{2}\|A - X - L\|_F^2 + \beta(\Psi(X,\phi) + \gamma TV(Q(\phi))), \tag{25}$$

where Q is the reshaping operator from $2D$ to $3D$ and $TV(\phi) = \langle \sqrt{(\nabla_x \phi)^2 + (\nabla_y \phi)^2 + (\nabla_t \phi)^2}, \mathbf{1} \rangle$ is the 3D TV. For simple denoising based approach, we can consider $\Psi(X, \phi) = \| |X| - \phi \|_2^2$ with simple thresholding. For more sophisticated segmentation, we can consider the following Chan-Vese model [7]

$$\Psi(X, \phi) = \langle (|X| - c_+)^2 - (|X| - c_-)^2, \, \phi \rangle,$$

where $0 \leq \phi \leq 1$, $c_+ = \langle \phi, \, |X| \rangle / \langle \phi, \, \mathbf{1} \rangle$, and $c_- = \langle \phi, \, |X| \rangle / \langle \mathbf{1} - \phi, \, \mathbf{1} \rangle$. When β is sufficiently small, we do not need to consider the additional terms $\beta \Psi(X, \phi)$, when we minimize with respect to X and L in (25). Therefore, we only need additional one iteration to generate foreground mask ϕ:

$$\begin{cases} (\hat{X}, \hat{L}) = SRAM(A) \\ \hat{\phi} = \arg\min_{\phi} \Psi(\hat{X}, \phi) + \gamma TV(Q(\phi)), \end{cases} \tag{26}$$

where $\hat{L} = W^* \Lambda^* (H^*)^T$ and $\hat{X}$ is a solution of RANMF (22).

Note that, for foreground detection problem, Zhou et. al. [44] proposed DECOLOR (detecting contiguous outliers in the low-rank representation) model:

$$\min_{L,Z} \ \frac{1}{2} \|P_{Z^\perp}(A - L)\|_F^2 + \alpha \|L\|_* + \beta \|Z\|_1 + \gamma \|T(Q(Z))\|_1 \tag{27}$$

to detect foreground object (i.e., outliers) in noisy video image sequence. Here $Z \in \{0, 1\}$, T is the node-edge incidence operator in markov random fields. For more details, see [44]. This model is useful when we want to know only the location of outliers from noisy data.

IV. Numerical Experiments

In this section, we compare the performance of the proposed RANMF model (22) using ℓ_p-norm and $\log$ function with other outlier detection models; RPCA (4), DECOLOR (27), and ℓ_1-NMF (3). Note that all models are implemented in Matlab (version 7.10). RPCA (4) uses a special SVD library, PROPACK[3]. All runs are performed on a laptop with an Intel i7-2720QM CPU (2.20GHz) and 16GB Memory. The Operating System is 64bit Windows.

For RANMF, we use the proposed SRAM algorithm (23). For column outlier separation, we use two different sparsity detection functions; ℓ_p-norm ($p = 0.65$) and $\log$ function. We set $B_H = \frac{0.1}{\sqrt{n}}$, where n is the size of row dimension of each data matrix. As observed in Figure 3, we can choose sufficiently large rank parameter r, but computational cost is also increased and can be unstable (see Example II.10). Therefore, based on our empirical observation, we fix $r = 20$ for all experiments. α is tuned for best

[3]http://sun.stanford.edu/ rmunk/PROPACK/

Fig. 4: A comparison of the performance of the proposed RANMF (22) using ℓ_p-norm and $\log$ function with RPCA (4), ℓ_1-NMF (3) for removal of shadows, specularities, and gross error in images of faces. From left to right, the data matrix A, the low rank/rank$_+$ matrix L of RPCA, ℓ_1-NMF, ℓ_1-RANMF, ℓ_p-RANMF ($p = 0.65$), and $\log$-RANMF. The proposed RANMF model (22) well recover low rank$_+$ face images than any other models. Note that YaleB face database [18] is used.

performance; $\alpha = 0.02$ for ℓ_p-RANMF, 0.0022 for $\log$-RANMF, 0.1 for ℓ_1-NMF. We initialize W^0 with $\mathbf{1}/\sqrt{mn}$ and H^0 with B_H for SRAM algorithm (i.e., for ℓ_p-RANMF, $\log$-RANMF). Other models are initialized with random positive matrices. For ℓ_p-RANMF, we fix $p = 0.65$. The stopping criterion for the outer iteration is

$$\|A - X^t - \bar{W}^t\bar{H}^t\|_F/\|A\|_F \leq 10^{-2} \tag{28}$$

and the maximum number of iterations is set to 20. For SRAM based model, we modified $2r$-BCD [24] to find a solution of the ℓ_2-NMF subproblems in (23). Note that the number of iterations of $2r$-BCD is fixed as five. For ℓ_1-NMF (3), we use the same framework of RANMF with $B_H = \infty$, since the algorithm in [43] is the case of one iteration of $2r$-BCD (without $BASIS$ and B_H) and the algorithm of [43] is a little bit algorithmically unstable to be used for equal comparison.

For RPCA (4), we use IALM [29][4]. The stopping condition is

$$\|A - X^t - L^t\|_F/\|A\|_F \leq 10^{-7}.$$

[4]http://perception.csl.uiuc.edu/matrix-rank/ Files/inexact_alm_rpca.zip

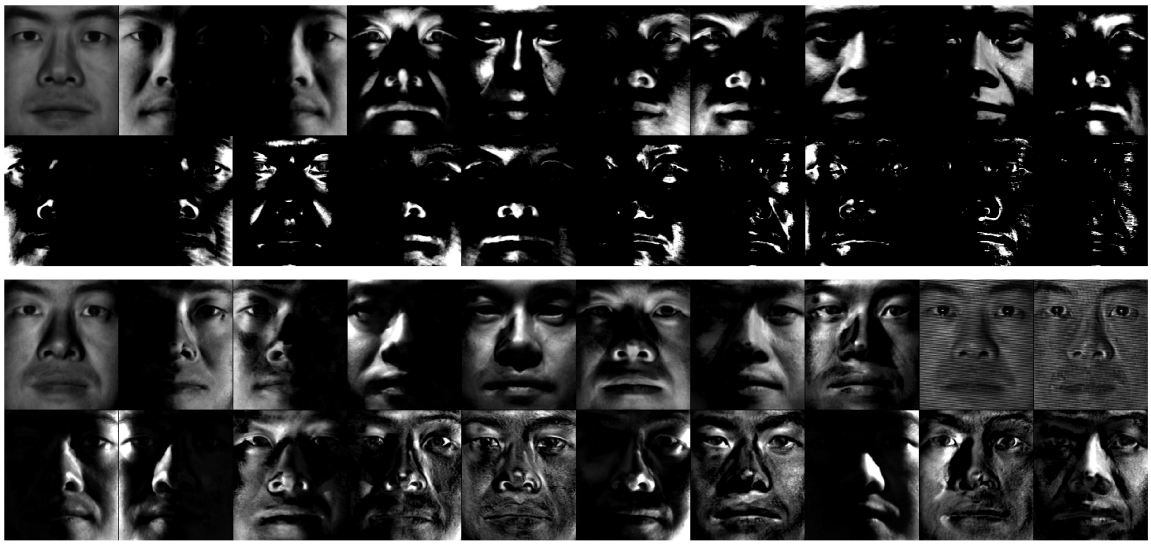

Fig. 5: Twenty column vector w_i (reshaped as a matrix) of basis matrix W of the ℓ_p-RANMF (the first row) and ℓ_1-NMF (the second row) in Figure 4. The W matrix of the proposed RANMF model shows more interpretable basis matrix than that of ℓ_1-NMF. Each basis matrix w_i generated by the proposed RANMF (the first row) shows that it is related to specific light direction. However, the basis generated by ℓ_1-NMF do not have such an interpretation. Note that the basis matrix W of ℓ_p-RANMF is much sparser than that of ℓ_1-NMF, i.e. more stable.

The maximum number of iterations for IALM is set to 100. For more details, see [29]. We use $\nu = \frac{1}{\sqrt{m}}$ (recommended in [2]) for Figure 4 only. In general, we need to tune this parameter for best performance, especially when column outliers do not have random sparse structures. For instance, see Figure 3 for performance variation of the RPCA model (4) when we choose different regularization parameter (i.e. different choice of rank). For the background modeling problems in Figure 7, we tuned ν for best performance. For escalator scenario (the first row image sequence), we set $\nu = \frac{0.8}{\sqrt{m}}$, for tree scenario (the second row image sequence) and the office scenario (the third row image sequence), we set $\nu = \frac{0.47}{\sqrt{m}}$. For Figure 1, we set $\nu = \frac{0.15}{\sqrt{m}}$. For DECOLOR (27), we use the recommend parameters in [44].

In Figure 4, we evaluate the performance of the proposed RANMF model with the RPCA model (4) and the ℓ_1-NMF model (3). Here, shadow and gross error are considered as grouped outliers. We use Yale B[5]

[5]http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

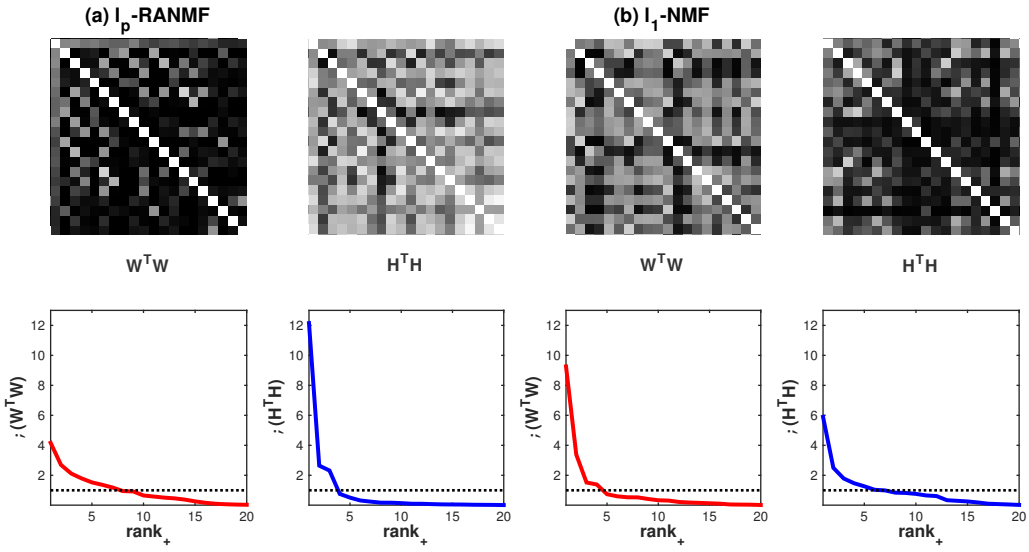

Fig. 6: A comparison of stability of (a) ℓ_p-RANMF model and (b) ℓ_1-NMF model for face image in Figure 4 and Figure 5. It shows the gram matrix W^TW and H^TH and the corresponding eigenvalues $\rho(W^TW)$ and $\rho(H^TH)$. In (a) ℓ_p-RANMF, we have stable (sparse) basis matrix W and dense (unstable) coefficient matrix H; see also Table I. In (b) ℓ_1-NMF, as observed in Figure 5, we have dense (unstable) basis matrix W. The reason is that since $r = 20$ is relatively large compared to $n = 58$ and ℓ_1-NMF does not have any regularization for rank_+, each image (corresponding to column outliers) of A is included into a rank and therefore the basis matrix W of ℓ_1-NMF is more denser and H is more sparser than that of RANMF; see also Figure 5.

face database [18]. Each person has 64 different images captured under different illumination condition, however in our experiments we use 58 image by removing extremely low illumination condition. The size of face images is 192×168. Therefore, the data matrix is $A \in [0, 255]^{32760\times 58}$. The proposed RANMF model does better separate shadow and gross error from low rank_+ face matrix than other models. Figure 5 shows twenty different basis matrix for W matrix of the ℓ_p-RANMF model in Figure 4. The W matrix of the proposed RANMF model shows more interpretable basis matrix than that of ℓ_1-NMF. That is, each basis vector (matrix after reshaping) w_i generated by the proposed RANMF (the first row) shows that it is related to specific light direction. However, the basis generated by ℓ_1-NMF do not have such an interpretation. Also, since $r = 20$ is relatively large compared to $n = 58$, the basis matrix W of ℓ_1-NMF

Fig. 7: A comparison of performance of the proposed RANMF (22) with RPCA (4) and DECOLOR (27) for column outliers separation/foreground detection problem. From left to right, the input image data, the results of RPCA (background and outliers), RANMF (background, outliers, and foreground), and DECOLOR (background and foreground). The proposed RANMF models separate column outliers (foreground objects) from the low rank background objects better than any other models. For the first row image sequences (escalator scenario), we used $\log$-RANMF model and for the second and the third row image sequences (waving tree and office scenario), we used ℓ_p-RANMF model with $p = 0.65$. Note that the regularization parameter ν in RPCA (4) is tuned for best performance. The rank parameter of the proposed RANMF (22) is fixed $r = 20$ for all experiments. The second and third row image sequences are from Wallflower [40]. The first row image sequences are from [21].

include each image frame as a rank and thus it cause dense basis matrix W and sparse coefficient matrix H. In more details, Figure 6 shows gram matrix of W and H and their eigenvalues of ℓ_p-RANMF and ℓ_1-NMF. As expected, W of ℓ_p-RANMF is stable (but sparse) and H of ℓ_p-RANMF is unstable (but

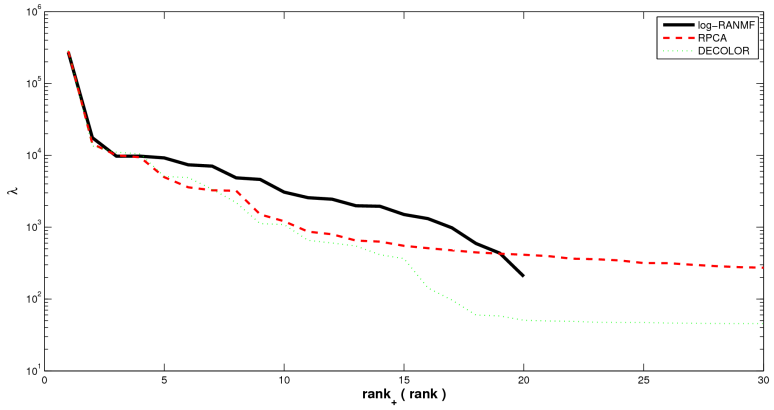

Fig. 8: The (asymmetric) singular values of the proposed log-RANMF (22) and RPCA (4) and DECOLOR (27) for escalator scenario in Figure. 7. Although asymmetric singular values of log-RANMF is larger than that of nuclear norm based models, the asymmetric singular values of log-RANMF are comparable to singular values of RPCA.

dense). For denseness condition, see Table I (face scenario); $\frac{1}{\sqrt{m}} = 0.0055$ and $\frac{1}{\sqrt{n}} = 0.1313$.

In Figure 7, we evaluate the performance of the proposed RANMF model (22) with SRAM algorithm (23) for background modeling and foreground detection. Due to the high correlation in consecutive video frames, which is captured by a fixed camera, it is natural to consider background objects as a low rank_+ matrix for the given nonnegative video data. Background objects can be static objects, such as walls and doors, or non-static objects, such as waving trees and moving escalators. Moving objects, such as pedestrians, are considered as column outliers. Here, we use the Wallflower [40] test images[6] for the second and third row test image sequences (waving tree and office scenarios). The size of each image of Wallflower is 160×120. We select 400 frames for the office scenario and use 287 frames for the waving tree scenario. Since we column-wise stacked each image frame, the data is $A \in [0, 255]^{19200 \times 400}$ for the office scenario and $A \in [0, 255]^{19200 \times 287}$ for the waving tree scenario. For the first row escalators scenario [21], the size of each image is 160×130 and we select 200 frames.

The first row of Figure 7 is the moving escalator background scenario. The second row of Figure 7 is the waving tree scenario; the background objects include the waving tree. These two scenarios show that the background objects are not static, but include periodic moving objects, such as waving tree and

[6]http://research.microsoft.com/en-us/um/people/jckrumm/wallflower/testimages.htm.

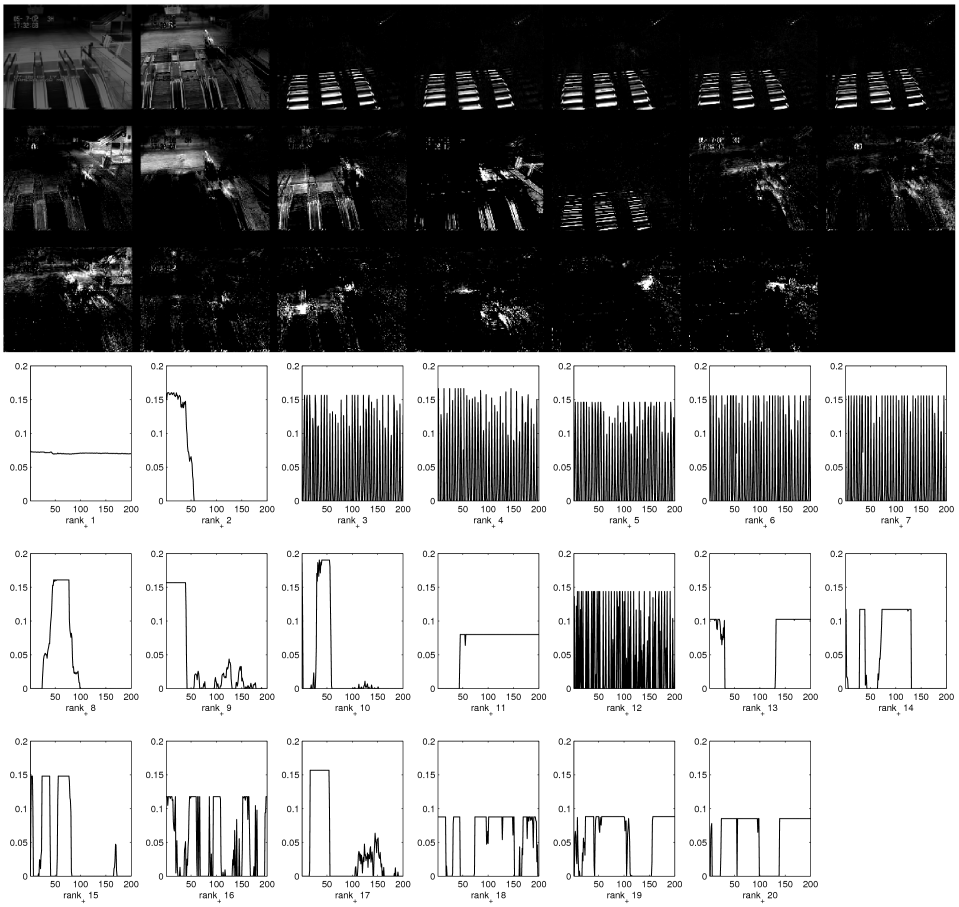

Fig. 9: The basis matrix W and the coefficient matrix H of the proposed $\log$-RANMF model (22) for escalator scenario in Figure 7. As expected, the basis matrix W is sparse and try to characterize inherent nonnegative rank structure. For instance, the first basis w_1 and the corresponding coefficient h_1 show the global basis structure. The basis $w_2, w_8 - w_{11}$ correspond to the global illumination change for specific period as shown in $h_2, h_8 - h_{11}$. Also, $w_3 - w_7, w_{12}$ is related to the movement of the escalator. See also the corresponding coefficient vectors $h_3 - h_7, h_{12}$.

moving escalator. The third row of Figure 7 is office scenario; a person switches off the light. There is an abrupt change of the background objects from bright to dark. Also the size of that person (i.e., column outliers) changes dynamically. In all cases, the proposed RANMF model well separates the moving objects (column outliers) from the various background objects (low-rank$_+$ approximation). On the contrary, the state-of-the-art RPCA (4) sometimes does not perform well although we tuned parameter. Especially, for

the office scenario, RPCA (4) does not clearly separate outliers from the background. The DECOLOR model sometimes fails to detect foreground or detect too large area than the moving objects itself. On the contrary, the proposed RANMF model with TV regularization in (25) well detect foreground objects with sharp boundary. For the first row image sequence, we use $\log$-RANMF model and for the second and the third row image sequence, we used the ℓ_p-RANMF model.

In Figure 8, we show asymmetric singular values of the proposed $\log$-RANMF (22) and singular values of RPCA (4) and DECOLOR (27) of escalator scenario in Figure 7. Although we indirectly minimize ANN-norm (7) via SRAM (23), the low rank$_+$ structure (i.e., via asymmetric singular value) of RANMF is comparable to that of nuclear norm based model. This is the main advantage of the proposed approach for column outliers separation problems. Also, see Figure 2 for asymmetric singular values of RANMF for synthetic data case. In Figure 9, we show basis matrix W and coefficient matrix H of the $\log$-RANMF model (22) for escalator scenario. As expected, basis matrix W is sparse and try to characterize inherent nonnegative rank structure. For instance, the first basis vector w_1 (i.e., matrix after reshaping) and the corresponding coefficient h_1 show the global basis structure. The basis $w_2, w_8 - w_{11}$ correspond to the global illumination change for specific period as shown in $h_2, h_8 - h_{11}$. Also, $w_3 - w_7, w_{12}$ catch the periodic movement of the escalator. See also the corresponding coefficient vectors $h_3 - h_7, h_{12}$.

In Table I, we compare the stability and denseness parameters of RANMF (22) and ℓ_1-NMF (3). Note that, for office scenario, $S(W)$ of RANMF is larger than that of ℓ_1-NMF, although $\|W\|$ shows that RANMF is more stable than that of ℓ_1-NMF. The reason of this is the following. Since the background of office scenario is in principle is two (light on and light off) and the coefficient matrix H bounded by η_H, the other many basis vectors of W are contributed randomly with constrained coefficient matrix H. Therefore, it has a chance to be very close each other. In this case, we get large $S(W)$. However, $\|W\|$ depends on the sum of all column of W^TW as noticed in (13). Therefore, the change of $\|W\|$ is more robust to the noise than $S(W)$ which depends on maximum value of W^TW matrix. In Table I, we can see clearly the relation between stability $\|Y\|$ vs. denseness $\|Y\|_{\ell_\infty}$ with $Y = W, H$. The basis matrix W of RANMF is clearly more stable (but sparse) than that of ℓ_1-NMF. On the contrary, the coefficient matrix H of RANMF is more dense (but unstable) than that of ℓ_1-NMF. It is well matched with the theoretical analysis in Section II-A.

In Table II, we compare the asymmetric incoherence criterion (20) of RANMF (22) and ℓ_1-NMF (3). It shows that $aINC$ of the proposed RANMF model is always sufficiently lower than that of ℓ_1-NMF. That is, $aINC$ well characterize a model with stable basis matrix W and dense coefficient matrix H for column outliers separation problem.

		$S(Y)$				$\|Y\|$				$\|Y\|_{\ell_\infty}$			
W/H	**Model**	Face	Escalator	Tree	Office	Face	Escalator	Tree	Office	Face	Escalator	Tree	Office
W	ℓ_p-RANMF	0.7606	0.6530	0.3916	0.8489	2.0230	1.7637	1.8071	2.1130	0.1114	0.4579	0.2198	0.1786
	log-RANMF	0.8246	0.6554	0.3892	0.8792	2.0582	1.8176	1.8179	1.9610	0.1126	0.3714	0.2450	0.6103
	ℓ_1-RANMF	0.7503	0.6635	0.4769	0.9817	2.0922	1.8176	1.7308	2.3443	0.1378	0.7277	0.4537	0.1552
	ℓ_1-NMF	0.8533	0.8649	0.8639	0.7320	2.9485	2.9701	3.5258	3.0229	0.0448	0.1035	0.0775	0.0616
H	ℓ_p-RANMF	0.9653	0.9452	0.8090	0.9960	3.5406	3.1849	3.1639	3.7671	0.2221	0.1608	0.1307	0.2105
	log-RANMF	0.9672	0.9490	0.7833	1.0000	3.6174	3.0326	3.1967	3.1428	0.2320	0.1704	0.1333	0.2045
	ℓ_1-RANMF	0.9312	0.9498	0.9720	0.9985	3.4077	3.1478	3.3717	3.6533	0.2314	0.1828	0.1325	0.2272
	ℓ_1-NMF	0.7453	0.8095	0.7093	0.8656	2.3602	2.2079	2.3800	2.0932	0.8935	0.5058	0.5630	0.9998

TABLE I: A comparison of stability $S(Y), \|Y\|$ and denseness $\|Y\|_{\ell_\infty}$ $(Y = W, H)$ of four different models; RANMF with ℓ_p-norm (p=0.65,1) and log-function and ℓ_1-NMF. Here, we use four different test image sequences; face, escalator, waving tree, and office in Figure 4 and Figure 7. For office scenario, due to noise, $S(W)$ of RANMF is larger than that of ℓ_1-NMF. However, since $\|W\|$ is related to whole column values of W^TW (e.g. (13)), it is more robust to noise. Note that $\|Y\|_\infty$ depends on its size of data; $\|W\|_\infty \in [\frac{1}{\sqrt{m}}, 1]$ and $\|H\|_\infty \in [\frac{1}{\sqrt{n}}, 1]$.

	$\Xi(W)$				$\Xi(H)$				$aINC(L)$			
Model	Face	Escalator	Tree	Office	Face	Escalator	Tree	Office	Face	Escalator	Tree	Office
ℓ_p-RANMF	0.0551	0.2596	0.1216	0.0845	0.0627	0.0505	0.0413	0.0559	1.1394	0.1944	0.3397	0.6611
log-RANMF	0.0547	0.2043	0.1348	0.3112	0.0641	0.0562	0.0417	0.0651	1.1726	0.2750	0.3095	0.2091
ℓ_1-RANMF	0.0658	0.4004	0.2621	0.0662	0.0679	0.0581	0.0393	0.0622	1.0313	0.1450	0.1499	0.9393
ℓ_1-NMF	0.0152	0.0348	0.0220	0.0204	0.3786	0.2291	0.2365	0.4777	24.9004	6.5751	10.7567	23.4457

TABLE II: A comparison of $aINC(L)$ of four different models; RANMF with ℓ_p-norm (p=0.65,1) and log-function and ℓ_1-NMF. Here, we use four different test image sequences; face, escalator, waving tree, and office. We can see the significant different incoherent parameter $aINC(L)$ between the proposed RANMF model and the conventional ℓ_1-NMF model for all test image sequences. That is, $aINC(L)$ well characterize stable (and sparse) basis matrix with dense (and unstable) coefficient matrix for column outliers separation problem.

V. Conclusion

In this paper, we propose the robust asymmetric NMF model with soft regularized asymmetric alternating minimization algorithm to remove column outliers, while obtaining the inherent low nonnegative rank structure of the given high dimensional image data. The numerical results, for background modeling

in video image sequence and removal of gross error in images of faces, show that our proposed robust asymmetric NMF models with ℓ_p-norm or $\log$ cost function do better recover the inherent low nonnegative rank structure than the state-of-the-art nuclear norm based robust PCA and DECOLOR and other robust NMF models. The main advantage of the proposed robust asymmetric NMF model is that it does not sensitive to the choice of the rank parameter.

VI. Acknowledgement

The authors would like to thank Nicolas Gillis for useful discussions. Hyenkyun Woo was supported by the Basic Science Research Program (2010-0510-1-3) through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology. Haesun Park was supported in part by NSF Grant CCF-1348152 and DARPA XDATA Grant FA8750-12-2-0309. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Appendix A
Incoherence Condition of RPCA (4)

The incoherence condition [3] of a matrix $Y \in \mathbb{R}^{m\times r}$ is defined as

$$ic(Y) = \max_i \|P_Y e_i\|_2, \tag{29}$$

where P_Y is a projection operator onto the subspace generated by Y, $e_i \in \mathbb{R}^m$ is a standard coordinate unit vector. Let us assume that rank$(Y) = r$, then $ic(Y) \in [\sqrt{\frac{r}{m}}, 1]$. Furthermore, let

$$\text{inc}(L_0) = \max\{ic(U), ic(V)\}, \tag{30}$$

where $SVD(L_0) = U\Sigma V^T$ and U, V are the left and right singular vectors of L_0. Note that (30) is incoherence condition of row and column subspace of L_0 [8]. As discussed in [2], the incoherence condition asserts that for small value of inc(L_0), the singular vectors of L_0 are sufficiently dense and away from standard coordinate axes.

In addition, Chandrasekaran et.al. [8] introduce slightly different incoherence condition for RPCA (4). To measure denseness of a matrix L_0, they introduce

$$\xi(L_0) := \max_{N\in T(L_0), \|N\|\le 1} \|N\|_{\ell_\infty}, \tag{31}$$

where $\|N\| = \max_{\|y\|_2 \leq 1} \|Ny\|_2$ and $T(L_0) = \{UC^T + DV^T \ : \ C \in \mathbb{R}^{n\times r}, D \in \mathbb{R}^{m\times r}\}$. The small value of $\xi(L_0)$ means that the elements of the tangent space $T(L_0)$ are not too sparse. Actually, they showed that the following relation between (30) and (31):

$$\text{inc}(L_0) \leq \xi(L_0) \leq 2\text{inc}(L_0). \tag{32}$$

To measure sparsity pattern of the given matrix X_0, Chandrasekaran et.al. [8] introduced an additional criterion:

$$\mu(X_0) := \max_{N\in\Omega(X_0), \|N\|_{\ell_\infty}\leq 1} \|N\|, \tag{33}$$

where $\Omega(X_0) = \{N \in \mathbb{R}^{m\times n} \ : \ \iota_N \subseteq \iota_{X_0}\}$ and ι_Y is an indicator matrix of nonzero elements of Y. Note that a matrix X_0 with bounded degree (i.e., limited sparsity pattern) has small $\mu(X_0)$.

$$\text{deg}_{min}(X_0) \leq \mu(X_0) \leq \text{deg}_{max}(X_0), \tag{34}$$

where $X_0 \in \mathbb{R}^{m\times n}$, $\text{deg}_{min}(Y) = \arg\min\{\min_i \|row_i(Y)\|_0, \ \min_i \|column_i(Y)\|_0\}$, and $\text{deg}_{max}(Y) = \arg\max\{\max_i \|row_i(Y)\|_0, \ \max_i \|column_i(Y)\|_0\}$. Here, $row_i(Y)/column_i(Y)$ are the i-th row/column vector of Y. The sparsity pattern of a matrix X_0 determines the value of $\mu(X_0)$. Note that, to get a true solution $\{\hat{L} = L_0, \hat{X} = X_0\}$, at least $\mu(X_0)\xi(L_0) < 1$ should be satisfied. From (32) and (34) with $\text{inc}(L_0) \geq \sqrt{\frac{r}{\min\{m,n\}}}$, we get $\text{deg}_{max}(X_0) < \sqrt{\frac{\min\{m,n\}}{4r}}$. It means that, when $r = 1$, we cannot recover outliers, which is larger than $\sqrt{\min\{m,n\}}/2$ only in one direction. For instance, for $A = 1000 \times 1000$ matrix, 16×1 size outliers X_0 cannot be guaranteed to be separated from low rank matrix L_0. However, we empirically tune regularization parameter ν of RPCA (4) for the separation between column outliers and low rank matrix; see Figure 1.

APPENDIX B
LOWER BOUND OF THE NONNEGATIVE RANK

In this section, we introduce two different forms of lower bound of the nonnegative rank. One is related to nuclear norm and the other is related to combinatorial ℓ_0-norm.

Since we set ℓ_2-norm and ℓ_∞-norm (i.e. ℓ_0-norm by Lemma II.1) constraints on the set Z_η^d (6), the lower bound of rank_+ also has a connection to the norm based lower bound [12], [13] and area based lower bound [15], [36]. In the following theorem, we give an explanation of the proposed lower bounds of rank_+.

Theorem B.1. *Let $L = \sum_{i=1}^{r} L_i \in \mathbb{R}_+^{m\times n}$ be an ANMF* (8) *with $L_i = \lambda_i w_i h_i^T$, then we have two different types of lower bounds of rank$_+(L)$:*

$$rank_+(L) \geq \frac{\|L\|_\diamond}{\max_i \lambda_i} \quad \textit{and} \quad rank_+(L) \geq \frac{\|L\|_A}{\max_i \sqrt{\|L_i\|_{\ell_0}}}, \tag{35}$$

where $\|L\|_\diamond = \sum_{i=1}^{r} \lambda_i$ and $\|L\|_A = \sum_{i=1}^{r} \sqrt{\|L_i\|_{\ell_0}}$.

Proof: Since $L = \sum_{i=1}^{r} \lambda_i w_i h_i^T$, we have

- norm-based:

$$\|L\|_{\ell_1} \leq \sum_{i=1}^{r} \lambda_i \sqrt{\|L_i\|_{\ell_0}} \leq \max_j \sqrt{\|L_j\|_{\ell_0}} \sum_{i=1}^{r} \lambda_i \leq r \max_i \lambda_i \max_j \sqrt{\|L_j\|_{\ell_0}},$$

- area-based:

$$\|L\|_{\ell_1} \leq \sum_{i=1}^{r} \lambda_i \sqrt{\|L_i\|_{\ell_0}} \leq \max_j \lambda_j \sum_{i=1}^{r} \sqrt{\|L_i\|_{\ell_0}} \leq r \max_j \lambda_j \max_i \sqrt{\|L_i\|_{\ell_0}}.$$

The first inequality follows from Cauchy-Schwarz inequality and $\|w_i h_i^T\|_{\ell_2} = 1$, since $\|w_i\|_2 = \|h_i\|_2 = 1$. ■

Recently, Fawzi and Parrilo [13] introduced general lower bound of rank$_+(L)$ of $L \in \mathbb{R}_+^{m\times n}$:

$$\text{rank}_+(L) \geq \frac{\mathcal{N}^*(L)}{\mathcal{N}(L)}, \tag{36}$$

where $\mathcal{N}(L)$ is positively homogeneous and monotone, that is,

$$\mathcal{N}(aL) = a\mathcal{N}(L) \quad \text{and} \quad L \leq M \to \mathcal{N}(L) \leq \mathcal{N}(M),$$

where $a \geq 0$ and $L, M, M - L \in \mathbb{R}_+^{m\times n}$. Also, they define $\mathcal{N}^*(L) = \min\{\, t > 0 \;:\; L \in t\,\text{conv}(\mathcal{A}_\mathcal{N}) \,\}$ where $\mathcal{A}_\mathcal{N} = \{B \in \mathbb{R}_+^{m\times n} \;:\; \text{rank}(B) \leq 1 \;\text{ and }\; \mathcal{N}(B) \leq 1\}$ and conv(C) is a convex hull of C. Note that, if we set $\mathcal{N}(L) = \|L\|$ then we get

$$\text{rank}_+(L) \geq \frac{\|L\|_+^*}{\|L\|}, \tag{37}$$

where $\|L\| = \max_{\|v\|_2=1} \|Lv\|_2$. Notice that since we do not use ℓ_∞-norm bound, the lower bound in Theorem B.1 also apply to the general NMF with nonnegative nuclear norm $\|L\|_*^+$ (9). Hence, if $\max_i \lambda_i = \|L\|$ for the first equation in (35), then we get the same lower bound (37). Moreover, it is not unnatural to assume $L \in \mathbb{R}_{>0}^{m\times n}$ for image data. Then, by Perron-Frobenius Theorem [31], we get $\max_i \sqrt{\|L_i\|_{\ell_0}} = \sqrt{mn}$ for $L \in \mathbb{R}_{>0}^{m\times n}$ when $\max_i \lambda_i = \|L\|$. Therefore, the second inequality in (35) becomes

$$\text{rank}_+(L) \geq \|L\|_A,$$

where $\|L\|_A$ is simply normalized version with $\sqrt{mn}$.

REFERENCES

[1] Bach F. (2013) Convex relaxations of structured matrix factorization *arXiv preprint*

[2] Candés E. J., Li X., Ma Y., and Wright J. (2010) Robust principal component analysis? *Journal of ACM*, 58:1-37.

[3] Candès E. and Recht B. (2009) Exact matrix completion via convex optimization *Found. Comput. Math.*, 9:717-772.

[4] Candès E. and Tao T. (2010) The power of convex relaxation: Near-optimal matrix completion *IEEE Trans. Info. Theory*, 56:2053-2080.

[5] Candès E., Wakin M., and Boyd S. (2008) Enhancing sparsity by reweighed ℓ_1 minimization *J. Fourier and Anal. Appl.*, 14:877-905.

[6] Chan R., Ho C.-W., and Nikolova M. (2005) Salt-and-pepper noise removal by median-type noise detectors and datail-preserving regularization *IEEE Trans. Image Processing*, 14:1479-1485.

[7] Chan T.F., Vese L. A. (2001) Active contour without edges *IEEE Trans. on Image Processing*, 10:266-277.

[8] Chandrasekaran V., Sanghavi S., Parrilo P.A., and Willsky A.S. (2011) Rank-Sparsity incoherence for matrix decomposition *SIAM J. Opt.*, 21:572-596

[9] Cichocki A. and Phan A. (2007) Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization, *Independent Component Analysis and Signal Separation*, 169-176.

[10] Cichocki A., Zdunek R., Phan A.H., and Amari S.-i. (2009) Nonnegative matrix and tensor factorizations *John Wiley & Sons.*

[11] Epstein R., Hallinan P., and Yuille A. (1995) 5 ± 2 Eigneimages Suffice: An empirical investigation of low-dimensional lighting models *Proc. IEEE Workshop Physics-based modeling in computer vision*, 108-116.

[12] Fawzi H. and Parrilo P.A. (2012) New lower bounds on nonnegative rank using conic programming *arXiv preprint*

[13] Fawzi H. and Parrilo P.A. (2014) Self-scaled bounds for atomic cone ranks: applications to nonnegative rank and cp-rank *arXiv preprint*

[14] Fazel M. (2002) Matrix rank minimization with applications Ph.D. Thesis, Stanford University.

[15] Fiorini S., Kaibel V., Pashkovich K., Theis D.O. (2013) Combinatorial bounds on nonnegative rank and extended formulations *Discrete Mathematics*, 313:67-83.

[16] Foucart S. and Rauhut H. (2013) A mathematical introduction to compressive sensing *Birkhäuser*

[17] Garnaev A.Y. and Gluskin E.D. (1984) On widths of the Euclidean ball *Sov. Math. Dokl.*, 30:200-204.

[18] Georghiades A., Belhumeur P., and Kriegman D. (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:643-660.

[19] Gillis N. and Glineur F. (2012) On the geometric interpretation of the nonnegative rank *Linear Algebra and its Applications*, 437:2685-2712.

[20] Golub G. and Van Loan C. F. (2013) Matrix Computations (4e) *Johns Hopkins Press*.

[21] Huang Li L., Gu W.,, Tian I. Q. (2004) Statistical modeling of complex backgrounds for foreground object detection *IEEE Transaction on Image Processing*, 13:14591472.

[22] Kelner J. (2014) Recovering hidden sparsity via sum of squares *http://simons.berkeley.edu/talks/jonathan-kelner-2014-09-23*

[23] Ke Q. and Kanade T. (2005) Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming, *IEEE CVPR*.

[24] Kim J., He Y., and Park H. (2014) Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework, *J. Glob. Optim.*, 58:285-319.

[25] Kong D., Ding C., and Huang H. (2011) Robust nonnegative matrix factorization using L21-norm, *ACM Int. Conf. on Info. Know. and Manage.*.

[26] Lam E. Y. (2008) Non-negative matrix factorization for images with Laplacian noise, *IEEE Asia Pacific Conf. on Circuits and Systems*.

[27] Lee J., Recht B., Salakhutdinov R., Srebro N., and Tropp J.A. (2010) Practical Large-Scale Optimization for max-norm regularization *NIPS*

[28] Lee D. and Seung H. (2001) Algorithms for non-negative matrix factorization, *Adv. in NIPS*.

[29] Lin Z., Chen M., Wu L., and Ma Y. (2009) The augmented Lagrange multiplier method for exact recovery of a corrupted low-rank matrices, *Preprint*.

[30] Ling Q., Wen Z., and Yin W. (2013) Decentralized jointly sparse optimization by reweighed ℓ_q minimization *IEEE Trans. Signal Processing*, 61:1165-1170.

[31] Minc H. (1988) Nonnegative Matrices *John Wiley & Sons*.

[32] Pompili F., Gillis N., Absil P.-A., and Glineur F. (2014) Two algorithms for orthogonal nonnegative matrix factorization with application to clustering *Neurocomputing*, 141:15-25

[33] Ramamoorthi R. (2002) Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object *IEEE Trans. PAMI*, 24:1-12.

[34] Ramirez I. and Sapiro G. (2012) An MDL framework for sparse coding and dictionary learning *IEEE Trans. on Signal Processing*, 60:2913-2927.

[35] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, 1970.

[36] Rothvoss T. The matching polytope has exponential extension complexity *STOC 2014*

[37] Rudin L., Osher S., and Fatemi E. (1992) Nonlinear total variation based noise removal algorithms *Phys. D.*, 60:259-268.

[38] Srebro N. and Shraibman A. (2005) Rank, Trace-Norm and Max-Norm *Lecture Notes in Comp. Sci.; Learning Theory*, 3559:545-560

[39] Studer C., Goldstein T., Yin W., and Baraniuk R. G. (2014) Democratic Representations *arXiv preprint*

[40] Toyama K., Krumm J., Brumitt B., and Meyers B. (1999) Wallflower: principles and practice of background maintenance, *IEEE ICCV*.

[41] Xu H., Caramanis C., and Sanghavi S. (2010) Robust PCA via Outlier Pursuit *Advances in Neural Information Processing Systems*.

[42] Vavasis S. (2009) On the complexity of nonnegative factorization *SIAM J. on Opitmization*, 20:1364-1377.

[43] Zhang L., Chen Z., Zheng M., and He X. (2011) Robust non-negative matrix factorization, *Fron. Electr. Electron. Eng. China*, 6:192-200.

[44] Zhou X., Yang C., and Yu W. Moving object detection by detecting contiguous outliers in the low-rank representation *IEEE Trans. PAMI*, 35:597-610.