MODEL SELECTION FOR NONNEGATIVE MATRIX FACTORIZATION BY SUPPORT UNION RECOVERY

Zhaoqiang Liu

Department of ECE, National University of Singapore

ABSTRACT

Nonnegative matrix factorization (NMF) has been widely used in machine learning and signal processing because of its non-subtractive, part-based property which enhances interpretability. It is often assumed that the latent dimensionality (or the number of components) is given. Despite the large amount of algorithms designed for NMF, there is little literature about automatic model selection for NMF with theoretical guarantees. In this paper, we propose an algorithm that first calculates an empirical second-order moment from the empirical fourth-order cumulant tensor, and then estimates the latent dimensionality by recovering the support union (the index set of non-zero rows) of a matrix related to the empirical second-order moment. By assuming a generative model of the data with additional mild conditions, our algorithm provably detects the true latent dimensionality. We show on synthetic examples that our proposed algorithm is able to find approximately correct number of components.

Index Terms— Nonnegative matrix factorization, Model selection, Tensor method, Multiple measurement vector, Support union recovery

1. INTRODUCTION

In a nonnegative matrix factorization (NMF) problem, we are given a data matrix $\mathbf{V} \in \mathbb{R}^{F\times N}$, and we seek non-negative factor matrices $\mathbf{W} \in \mathbb{R}^{F\times K}$, $\mathbf{H} \in \mathbb{R}^{K\times N}$ such that a certain distance between $\mathbf{V}$ and $\mathbf{WH}$ is minimized. To reduce the data dimension and for the purpose of efficient computation, the integer K, which is said to be the latent dimensionality or the number of components, is usually chosen such that $K(F+N) \ll FN$. Since the publication of the seminar paper [1] in 2000, NMF has been a popular topic in machine learning [2] and signal processing [3]. There are many fundamental algorithms to approximately solve the NMF problem [1, 4, 5] with the implicit assumption that an effective number of the latent dimensionality is known a priori.

Despite the practical success of these fundamental algorithms, the estimation of the latent dimensionality remains an important issue. For example, researchers may wonder whether we can achieve better approximation accuracy with significantly less running time by selecting a better K as the input of the algorithm. Unfortunately, there is generally little literature discussing the model selection problem for NMF. Moreover, the methods proposed in papers about detecting latent dimensionality for NMF [6–9] either lack theoretical guarantees or require rather stringent conditions on the generative model of data.

1.1. Main Contributions

We assume that each column $\mathbf{v}$ of the data matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N] \in \mathbb{R}^{F\times N}$ is sampled from the following generative model

$$\mathbf{v} = \mathbf{W}\mathbf{h} + \mathbf{z}, \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K] \in \mathbb{R}_+^{F\times K}$ is the mixing matrix (or the ground-truth non-negative dictionary matrix) and we assume that $\mathrm{rank}(\mathbf{W}) = K$. $\mathbf{h} \in \mathbb{R}^K$ is a latent random vector with independent coordinates[1], and $\mathbf{z} \in \mathbb{R}^F$ is a multivariate Gaussian random vector. $\mathbf{z}$ is assumed to be independent with $\mathbf{h}$. We write $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N] \in \mathbb{R}^{K\times N}$. In the context of this generative model, our goal is to find the number of columns of $\mathbf{W}$ from the observed matrix $\mathbf{V}$. This generative model can be viewed as a non-negative variant of that for independent component analysis (ICA) [11].

For the data matrix $\mathbf{V} \in \mathbb{R}^{F\times N}$ generated from the above model, we first calculate an empirical second-order moment, denoted as $\hat{\mathbf{M}}_2 \in \mathbb{R}^{F\times F}$, from the empirical fourth-order cumulant tensor. We prove that $\hat{\mathbf{M}}_2$ approximates its expectation, denoted as $\mathbf{M}_2$, well with high probability when N is sufficiently large. We also show that $\mathbf{M}_2$ can be written as $\mathbf{M}_2 = \mathbf{M}_2\mathbf{X}^*$, where $\mathbf{X}^* \in \mathbb{R}^{F\times F}$ contains exactly K non-zero rows. Finally, we prove that under certain conditions, an ℓ_1/ℓ_2 block norm minimization problem (cf. (27) to follow) over $\hat{\mathbf{M}}_2$ is able to detect the correct number of column of $\mathbf{W}$ from the recovery of a support union.

Complete proofs are presented in the extended version [12].

1.2. Notations

We use capital boldface letters to denote matrices and we use lower-case boldface letters to denote vectors. We use a_{ij}

The work of Z. Liu (a0109983@u.nus.edu, zqliu12@gmail.com) is supported by NUS IDS Grant R-263-000-C73-646.

[1] Similar to that in [10], we will not require $\mathbf{h}$ to be non-negative.

or $[\mathbf{A}]_{ij}$ to denote the (i,j)-th entry of $\mathbf{A}$. $[N]$ represents $\{1,2,\cdots,N\}$ for any positive integer N. For $\mathbf{X} \in \mathbb{R}^{L\times M}$ and any $l \in [L]$, $m \in [M]$, we use $\underline{\mathbf{x}}_l$, $\mathbf{x}_m$ to denote the l-th row and the m-th column of $\mathbf{X}$, respectively. We write $\underline{\mathbf{V}}_{\mathscr{K}} := \mathbf{V}(\mathscr{K},:)$ as the rows of $\mathbf{V}$ indexed by $\mathscr{K}$, and $\mathbf{V}_{\mathscr{K}} := \mathbf{V}(:,\mathscr{K})$ denotes the columns of $\mathbf{V}$ indexed by $\mathscr{K}$. $\|\mathbf{V}\|_1, \|\mathbf{V}\|_2, \|\mathbf{V}\|_\infty, \|\mathbf{V}\|_{\mathrm{F}}$ represents the 1-norm, the spectral norm, the infinity norm and the Frobenius norm of $\mathbf{V}$, respectively. Let $\mathbf{V}_1 \in \mathbb{R}^{F_1\times N}$ and $\mathbf{V}_2 \in \mathbb{R}^{F_2\times N}$. We denote by $[\mathbf{V}_1;\mathbf{V}_2]$ the vertical concatenation of the two matrices. $\mathrm{Diag}(\mathbf{w})$ represents the diagonal matrix whose diagonal entries are given by $\mathbf{w}$. The support of a vector $\mathbf{x}$ is denoted as $\mathrm{supp}(\mathbf{x}) := \{i : x_i \neq 0\}$. The support union of a matrix $\mathbf{X}$ with N columns is defined as $\mathrm{Supp}(\mathbf{X}) := \cup_{n=1}^N \mathrm{supp}(\mathbf{x}_n)$.

2. TENSOR METHODS

In this section, we calculate an empirical second moment $\hat{\mathbf{M}}_2$ using a tensor method, and we prove that the empirical second moment is close to its expectation $\mathbf{M}_2$ with high probability when the sample size N is sufficiently large.

2.1. The Derivation of $\mathbf{M}_2$ and $\hat{\mathbf{M}}_2$

Let $\mathbf{v}$ be a random vector corresponding to the generative model (1) with $\mathbb{E}[h_k] = 0$ and $\mathbb{E}[z_f] = 0$ for $k \in [K], f \in [F]$. We have the following lemma which says that $\mathbf{M}_2$ can be written in a nice form.

Lemma 1 ([13,14]) *Define*

$$\mathcal{M}_4 := \mathbb{E}[\mathbf{v}\otimes\mathbf{v}\otimes\mathbf{v}\otimes\mathbf{v}] - \mathcal{T}, \tag{2}$$

where for all $i,j,l,m \in [F]$, $[\mathbf{v}\otimes\mathbf{v}\otimes\mathbf{v}\otimes\mathbf{v}]_{ijlm} = v_i v_j v_l v_m$ and $\mathcal{T}$ is the fourth-order tensor with

$$[\mathcal{T}]_{ijlm} := \mathbb{E}[v_i v_j]\mathbb{E}[v_l v_m] + \mathbb{E}[v_i v_l]\mathbb{E}[v_j v_m] + \mathbb{E}[v_i v_m]\mathbb{E}[v_j v_l]. \tag{3}$$

Let[2] $\kappa_k = \mathbb{E}[h_k^4] - 3\mathbb{E}[h_k^2]$ *for each $k \in [K]$. Then*

$$\mathcal{M}_4 = \sum_{k=1}^K \kappa_k \mathbf{w}_k\otimes\mathbf{w}_k\otimes\mathbf{w}_k\otimes\mathbf{w}_k. \tag{4}$$

In addition, we have that for any $\mathbf{s},\mathbf{t} \in \mathbb{R}^F$,

$$\mathbf{M}_2 := \mathcal{M}_4(\mathbf{I},\mathbf{I},\mathbf{s},\mathbf{t}) = \sum_{k=1}^K \kappa_k(\mathbf{s}^T\mathbf{w}_k)(\mathbf{t}^T\mathbf{w}_k)\mathbf{w}_k\mathbf{w}_k^T, \tag{5}$$

where for matrices $\mathbf{V}_1 \in \mathbb{R}^{F\times F_1}, \mathbf{V}_2 \in \mathbb{R}^{F\times F_2}, \mathbf{V}_3 \in \mathbb{R}^{F\times F_3}, \mathbf{V}_4 \in \mathbb{R}^{F\times F_4}$, $\mathcal{M}_4(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4)$ is defined as the tensor whose (i_1,i_2,i_3,i_4)-th entry is

$$\sum_{j_1,j_2,j_3,j_4\in[F]} [\mathcal{M}_4]_{j_1,j_2,j_3,j_4}[\mathbf{V}_1]_{j_1,i_1}[\mathbf{V}_2]_{j_2,i_2}[\mathbf{V}_3]_{j_3,i_3}[\mathbf{V}_4]_{j_4,i_4}.$$

[2] It is implicitly assumed in [14] that $\mathrm{Var}[h_k] = 1$, and thus $\kappa_k = \mathbb{E}[h_k^4] - 3$.

We calculate $\hat{\mathbf{M}}_2$ from the sample matrix $\mathbf{V}$. Let

$$\hat{\mathcal{M}}_4 := \frac{\sum_{n=1}^N \mathbf{v}_n\otimes\mathbf{v}_n\otimes\mathbf{v}_n\otimes\mathbf{v}_n}{N} - \hat{\mathcal{T}}, \tag{6}$$

where $\hat{\mathcal{T}}$ is the empirical approximation tensor for $\mathcal{T}$. Denoting $\hat{\mathbf{M}}_2$ as

$$\hat{\mathbf{M}}_2 = \hat{\mathcal{M}}_4(\mathbf{I},\mathbf{I},\mathbf{s},\mathbf{t}). \tag{7}$$

We have that $\mathbb{E}[\hat{\mathbf{M}}_2] = \mathbf{M}_2$. For simplicity, we take $\mathbf{s} = \mathbf{t} = \mathbf{e} \in \mathbb{R}^F$, where $\mathbf{e}$ is the vector of all ones. For any $k \in [K]$, because $\mathbf{w}_k \neq \mathbf{0}$, we have that $\mathbf{e}^T\mathbf{w}_k > 0$. In addition, if $\kappa_k \neq 0$, let $\alpha_k = \kappa_k(\mathbf{e}^T\mathbf{w}_k)^2$, we have $\alpha_k \neq 0$ and

$$\mathbf{M}_2 = \sum_{k=1}^K \alpha_k \mathbf{w}_k\mathbf{w}_k^T. \tag{8}$$

Moreover, now we have that for $i,j \in [F]$,

$$[\hat{\mathbf{M}}_2]_{ij} = \sum_{l=1}^F\sum_{m=1}^F [\hat{\mathcal{M}}_4]_{ijlm}. \tag{9}$$

2.2. Bounding the Distance between $\mathbf{M}_2$ and $\hat{\mathbf{M}}_2$

Let $\mathbf{E} = \hat{\mathbf{M}}_2 - \mathbf{M}_2$, and assume that all the coordinates of $\mathbf{h}$ are identically and independently distributed with $m_p := \mathbb{E}[h_k^p]$ for all $k \in [K], p \in \mathbb{N}$. In particular, we assume that $m_1 = 0$ and $m_4 \neq 3m_2$. Let $M_8 = m_8 + m_7m_1 + m_6m_2 + m_6m_1^2 + \ldots + m_2^4 + m_2^3m_1^2 + m_2^2m_1^4 + m_2m_1^6 + m_1^8$, $M_4 = m_4 + m_3m_1 + m_2^2 + m_2m_1^2 + m_1^4$ and $M = \max\{M_8, M_4^2\}$. Denote $W_{\max}$ as $W_{\max} := \max_{f,k} w_{fk}$. Suppose that $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$ and let $\Delta = \max\{\sigma,1\}$. From the following lemma, we can see that if N is sufficiently large, the distance between $\mathbf{M}_2$ and $\hat{\mathbf{M}}_2$ (with respect to Frobenius norm) is sufficiently small with high probability.

Lemma 2 *For any $\delta \in (0,1)$, we have that with probability at least $1-\delta$,*

$$\|\mathbf{E}\|_{\mathrm{F}} < \frac{117\sqrt{70M}W_{\max}^4K^4\Delta^4F^3}{\sqrt{\delta N}}. \tag{10}$$

3. SUPPORT UNION RECOVERY

In this section, we first show that $\mathbf{M}_2$ can also be written as $\mathbf{M}_2 = \mathbf{M}_2\mathbf{X}^*$, where the cardinality of the support union of $\mathbf{X}^*$ is $|\mathrm{Supp}(\mathbf{X}^*)| = K$. This motivates us to consider approaches for support union recovery or multiple measurement vectors [15–19]. We then present theoretical guarantees for support union recovery for an ℓ_1/ℓ_2 block norm minimization problem (cf. (20) to follow).

3.1. Another Formulation of $\mathbf{M}_2$

Recall that from (8), we have

$$\mathbf{M}_2 = \sum_{k=1}^K \alpha_k\mathbf{w}_k\mathbf{w}_k^T = \mathbf{W}\mathrm{Diag}(\boldsymbol{\alpha})\mathbf{W}^T, \tag{11}$$

where $\boldsymbol{\alpha} := [\alpha_1; \ldots; \alpha_K] \in \mathbb{R}^K$. We know that $\boldsymbol{\alpha}$ contains all non-zero entries if $m_4 \neq 3m_2$. Because we assume that $\operatorname{rank}(\mathbf{W}) = K$, there exists an index set $\mathscr{K}$ for rows of $\mathbf{W}$ such that $|\mathscr{K}| = K$ and $\operatorname{rank}(\underline{\mathbf{W}}_{\mathscr{K}}) = K$. Let $\mathbf{R} \in \mathbb{R}^{K\times(F-K)}$ be the matrix such that

$$\mathbf{R}^T = \underline{\mathbf{W}}_{\mathscr{K}^c}(\underline{\mathbf{W}}_{\mathscr{K}})^{-1}. \tag{12}$$

Let $\boldsymbol{\Pi}$ be the permutation matrix corresponding to the index set $\mathscr{K}$. We have that

$$\mathbf{M}_2 = \mathbf{W}\mathrm{Diag}(\boldsymbol{\alpha})\mathbf{W}^T = \mathbf{W}\mathrm{Diag}(\boldsymbol{\alpha})[\underline{\mathbf{W}}_{\mathscr{K}}^T, \underline{\mathbf{W}}_{\mathscr{K}^c}^T]\boldsymbol{\Pi} \tag{13}$$

$$= \mathbf{W}\mathrm{Diag}(\boldsymbol{\alpha})[\underline{\mathbf{W}}_{\mathscr{K}}^T, \underline{\mathbf{W}}_{\mathscr{K}}^T\mathbf{R}]\boldsymbol{\Pi} \tag{14}$$

$$= \mathbf{W}\mathrm{Diag}(\boldsymbol{\alpha})\underline{\mathbf{W}}_{\mathscr{K}}^T[\mathbf{I}, \mathbf{R}]\boldsymbol{\Pi} \tag{15}$$

$$= \mathbf{M}_2\boldsymbol{\Pi}\begin{bmatrix}\mathbf{I} & \mathbf{R}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\boldsymbol{\Pi} = \mathbf{M}_2\mathbf{X}^*, \tag{16}$$

where $\mathbf{X}^* := \boldsymbol{\Pi}\begin{bmatrix}\mathbf{I} & \mathbf{R}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\boldsymbol{\Pi}$. Note that the number of non-zero rows in $\mathbf{X}^*$ is exactly K, i.e., $|\mathrm{Supp}(\mathbf{X}^*)| = |\mathscr{K}| = K$.

3.2. Theoretical Results for Support Union Recovery

For $1 \le a \le b < \infty$ and any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the ℓ_a/ℓ_b block norm of $\mathbf{A}$ is defined as follows:

$$\|\mathbf{A}\|_{\ell_a/\ell_b} = \left(\sum_{i=1}^m \|\underline{\mathbf{a}}_i\|_b^a\right)^{1/a}, \tag{17}$$

where $\underline{\mathbf{a}}_i$ is the i-th row of $\mathbf{A}$. In particular, we define

$$\|\mathbf{A}\|_{\ell_\infty/\ell_2} = \max_{i\in[m]} \|\underline{\mathbf{a}}_i\|_2. \tag{18}$$

Assume that an observed data matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ can be written as

$$\mathbf{Y} = \mathbf{A}\mathbf{B}^* + \mathbf{L}, \tag{19}$$

where $\mathbf{A} \in \mathbb{R}^{m\times p}$ is the dictionary matrix, $\mathbf{B}^* \in \mathbb{R}^{p\times n}$ is block sparse. Let $\underline{\mathbf{b}}_i^*$ be the i-th row of $\mathbf{B}^*$, we write the support union of $\mathbf{B}^*$ as $\mathcal{S} := \mathrm{Supp}(\mathbf{B}^*)$. Considering the following ℓ_1/ℓ_2 block norm minimization problem,

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{Y} - \mathbf{A}\mathbf{B}\|_\mathrm{F}^2 + \lambda\|\mathbf{B}\|_{\ell_1/\ell_2}. \tag{20}$$

Let $b_{\min}^* = \min_{i\in\mathcal{S}} \|\underline{\mathbf{b}}_i^*\|_2$. According to Lemma 2 in [18], we can prove the following lemma which ensures the recovery of support union under certain conditions.

Lemma 3 *Assume that $\mathbf{A}_\mathcal{S}^T\mathbf{A}_\mathcal{S}$ is invertible and let $D_{\max} = \|\left(\mathbf{A}_\mathcal{S}^T\mathbf{A}_\mathcal{S}\right)^{-1}\|_\infty > 0$. If there exists a fixed parameter $\gamma \in (0, 1]$, such that*

$$\|\mathbf{A}_{\mathcal{S}^c}\mathbf{A}_\mathcal{S}\left(\mathbf{A}_\mathcal{S}^T\mathbf{A}_\mathcal{S}\right)^{-1}\|_\infty \le 1 - \gamma, \tag{21}$$

and $D_{\max}(\lambda + \|\mathbf{A}_\mathcal{S}\|_1\|\mathbf{L}\|_{\ell_\infty/\ell_2}) \le \frac{1}{2}b_{\min}^*$, $\|\mathbf{A}_{\mathcal{S}^c}\|_1\|\mathbf{L}\|_{\ell_\infty/\ell_2} \le \frac{\lambda\gamma}{2}$, then there is a unique optimal solution $\hat{\mathbf{B}}$ for (20) such that $\mathrm{Supp}(\hat{\mathbf{B}}) = \mathcal{S}$. Moreover, $\hat{\mathbf{B}}$ satisfies the bound

$$\|\hat{\mathbf{B}} - \mathbf{B}^*\|_{\ell_\infty/\ell_2} \le D_{\max}(\lambda + \|\mathbf{A}_\mathcal{S}\|_1\|\mathbf{L}\|_{\ell_\infty/\ell_2}) \le \frac{1}{2}b_{\min}^*. \tag{22}$$

4. THE MAIN THEOREM

Recall that from Section 3.1, we obtain

$$\|\mathbf{X}^*\|_{\ell_\infty/\ell_2} = \sqrt{1 + \|\mathbf{R}\|_{\ell_\infty/\ell_2}^2} = \sqrt{1 + r_{\max}^2}, \tag{23}$$

where $r_{\max} = \max_k \|\underline{\mathbf{r}}_k\|_2$ with $\underline{\mathbf{r}}_k$ being the k-th row of $\mathbf{R}$. In addition, let $r_{\min} = \min_k \|\underline{\mathbf{r}}_k\|_2$. We have

$$\min_{i\in\mathscr{K}} \|\underline{\mathbf{x}}_i^*\|_2 = \sqrt{1 + r_{\min}^2}, \tag{24}$$

where $\underline{\mathbf{x}}_i^*$ is the i-th row of $\mathbf{X}^*$. Let

$$\mathbf{W}_1 := \underline{\mathbf{W}}_{\mathscr{K}}\mathrm{Diag}(\boldsymbol{\alpha})\mathbf{W}^T\mathbf{W}\mathrm{Diag}(\boldsymbol{\alpha})\underline{\mathbf{W}}_{\mathscr{K}}^T, \tag{25}$$

$$\mathbf{W}_2 := \underline{\mathbf{W}}_{\mathscr{K}^c}\mathrm{Diag}(\boldsymbol{\alpha})\mathbf{W}^T\mathbf{W}\mathrm{Diag}(\boldsymbol{\alpha})\underline{\mathbf{W}}_{\mathscr{K}}^T. \tag{26}$$

We consider the ℓ_1/ℓ_2 block norm minimization problem over $\hat{\mathbf{M}}_2$.

$$\min_{\mathbf{X}\in\mathbb{R}^{F\times F}} \frac{1}{2}\|\hat{\mathbf{M}}_2 - \hat{\mathbf{M}}_2\mathbf{X}\|_\mathrm{F}^2 + \lambda\|\mathbf{X}\|_{\ell_1/\ell_2}. \tag{27}$$

We have the following main theorem which guarantees the discovering of the correct K.

Theorem 4 *Let $\lambda_{\min}(\mathbf{W}_1) := C_{\min} > 0$, where $\lambda_{\min}(\mathbf{W}_1)$ is the the minimal eigenvalue of $\mathbf{W}_1$. Let $D_{\max} := \|\mathbf{W}_1\|_\infty > 0$. Suppose there is a $\gamma \in (0, 1]$ such that*

$$\|\mathbf{W}_2\mathbf{W}_1^{-1}\|_\infty \le 1 - \gamma. \tag{28}$$

Let[3]

$$\zeta = \min\{\frac{\sqrt{C_{\min}}}{2}, \frac{\|\mathbf{M}_2\|_1}{\sqrt{F}}, \zeta_1, \zeta_2\}. \tag{29}$$

For any $\delta \in (0, 1)$, let

$$l_\lambda = \frac{936\sqrt{70M}W_{\max}^4K^4\Delta^4F^3}{\sqrt{\delta N}}\frac{\|\mathbf{M}_2\|_1\left(1 + \sqrt{F(1 + r_{\max}^2)}\right)}{\gamma} \tag{30}$$

and

$$u_\lambda = \frac{2\sqrt{1 + r_{\min}^2}}{(4 + \gamma)\left(D_{\max} + 6\|\mathbf{M}_2\|_1^2D_{\max}^2\right)}. \tag{31}$$

Then if

$$N \ge \frac{958230MW_{\max}^2K^8\Delta^8F^6}{\delta\zeta^2}, \tag{32}$$

[3] $\zeta_1 = \gamma/\left(6\sqrt{F}\|\mathbf{M}_2\|_1D_{\max}(1 + 8\|\mathbf{M}_2\|_1^2D_{\max})\right)$,
$\zeta_2 = \frac{\gamma\sqrt{1+r_{\min}^2}}{4(4+\gamma)\|\mathbf{M}_2\|_1\left(1+\sqrt{F(1+r_{\max}^2)}\right)\left(D_{\max}+6\|\mathbf{M}_2\|_1^2D_{\max}^2\right)}$.

Algorithm 1 Model selection for NMF by support union recovery

Input: Data matrix $\mathbf{V} \in \mathbb{R}^{F\times N}$, $\lambda > 0$, $\epsilon > 0$
Output: The estimated value of K, denoted as $\hat{K}$
1) Calculate $\hat{\mathbf{M}}_2$ from (35).
2) Obtain $\hat{\mathbf{X}}$ by optimizing (27).
3) $\hat{K} := |\{f \in [F] : \|\hat{\underline{\mathbf{x}}}_f\|_2 > \epsilon\}|$.

and

$$l_\lambda \leq \lambda \leq u_\lambda, \tag{33}$$

we have that with probability at least $1-\delta$*, there exists a unique optimal solution* $\hat{\mathbf{X}}$ *for* (27) *such that* $\mathrm{Supp}(\hat{\mathbf{X}}) = \mathscr{K}$*. In addition, we have the error bound*

$$\|\mathbf{X}^* - \hat{\mathbf{X}}\|_{\ell_\infty/\ell_2} \leq \frac{\sqrt{1+r_{\min}^2}}{2}. \tag{34}$$

If the conditions of Theorem 4 are satisfied, the optimal solution $\hat{\mathbf{X}}$ for (27) satisfies that $|\mathrm{Supp}(\hat{\mathbf{X}})| = K$, and thus we can count the number of non-zero rows of $\hat{\mathbf{X}}$ to obtain the true K. The whole procedure of our algorithm is summarized in Algorithm 1.

5. NUMERICAL RESULTS

To demonstrate the efficacy of Algorithm 1 for estimating K, we perform numerical simulations on synthetic datasets. We need to obtain $\hat{\mathbf{M}}_2$ in the first step of Algorithm 1. The time complexity of calculating $\hat{\mathbf{M}}_4$ is $O(F^4N)$. However, note that we do not need to calculate $\hat{\mathbf{M}}_4$ explicitly before calculating $\hat{\mathbf{M}}_2$. Let $p_n = \sum_{f=1}^F v_{f,n}$ and $q_n = p_n^2$ for $n \in [N]$. We can show that [12]

$$\hat{\mathbf{M}}_2 = \frac{\mathbf{V}\mathrm{Diag}(\mathbf{q})\mathbf{V}^T}{N} - \left(\frac{\sum_{n=1}^N q_n}{N^2}\mathbf{V}\mathbf{V}^T + 2\frac{\mathbf{V}\mathbf{p}\mathbf{p}^T\mathbf{V}^T}{N^2}\right). \tag{35}$$

The time complexity for calculating $\hat{\mathbf{M}}_2$ is reduced to $O(F^2N)$. In the second step of Algorithm 1, we use CVX [20] to obtain a solution $\hat{\mathbf{X}}$ for (27). ϵ is set to be 10^{-6} in the third step of Algorithm 1.

5.1. Synthetic Datasets

We fix $K = 10$ and vary F between 20 and 50. We vary the number of samples N from 100 to 10000. We set the dictionary matrix $\mathbf{W} \in \mathbb{R}^{F\times K}$ as $[\mathbf{I}_K; \tau\mathbf{W}_c]$, where $\mathbf{I}_K$ is the identity matrix in $\mathbb{R}^{K\times K}$ and $\mathbf{W}_c \in \mathbb{R}^{(F-K)\times K}$ is a random non-negative matrix generated from the command `rand(F-K,K)` in Matlab. $\tau > 0$ is properly chosen such that $\|\mathbf{W}_2\mathbf{W}_1^{-1}\|_\infty < 1$ (cf. (28)). Each entry h of $\mathbf{H}$ is generated from an exponential distribution[4] $\mathrm{Exp}(u)$ with parameter $u = 1$, and then is centralized by $h \leftarrow h - \frac{1}{u}$.

[4] $\mathrm{Exp}(u)$ is the function $x \mapsto u\exp(-ux)1\{x \geq 0\}$.

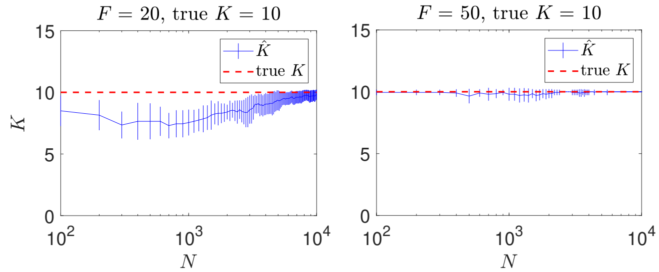

Fig. 1. Estimated number of components K with different F for exponential distribution. The error bars denote one standard deviation away from the mean.

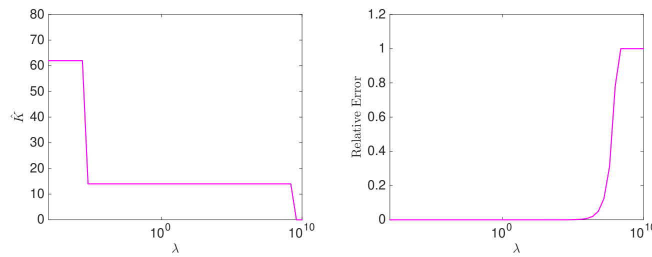

Fig. 2. Estimated number of components and relative error with different λ for the swimmer dataset.

The regularization parameter λ is set to be 10. The data matrix $\mathbf{V} = \mathbf{W}\mathbf{H} + \mathbf{Z}$ and each entry of the noise matrix $\mathbf{Z}$ is sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.01$. For each setting of the parameters, we generate 20 data matrices $\mathbf{V}$ independently. From Fig. 1, we observe that when $F = 20$, the algorithm cannot detect the true K until N is sufficiently large (e.g., $N \geq 6 \times 10^3$). When $F = 50$, we need a smaller τ such that $\|\mathbf{W}_2\mathbf{W}_1^{-1}\|_\infty < 1$, and the algorithm works well even when the sample size is relatively small.

5.2. The Swimmer Dataset

We perform experiments on the well-known swimmer [21] dataset, which is widely used for benchmarking NMF algorithms. The swimmer dataset we use contains 256 binary images (20-by-11 pixels) which depict figures with four limbs, each can be in four different positions. The latent dimensionality of the corresponding data matrix is 16. From the regularization path for this dataset presented in Fig. 2, we observe that the estimated latent dimensionality $\hat{K}$ is always 14 when $10^{-5} \leq \lambda \leq 10^9$. In addition, the relative error $\frac{\|\hat{\mathbf{M}}_2 - \hat{\mathbf{M}}_2\hat{\mathbf{X}}\|_{\mathrm{F}}}{\|\hat{\mathbf{M}}_2\|_{\mathrm{F}}}$ is close to 0 when $\lambda \leq 10^4$ and becomes intolerably large (larger than 0.75) when $\lambda \geq 10^8$. Therefore, a reasonable estimate for the latent dimensionality is 14, which is close to the true latent dimensionality.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2001, pp. 556–562.

[2] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Non-negative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons, 2009.

[3] I. Buciu, "Non-negative matrix factorization, a new tool for feature extraction: Theory and applications," *Int. J. Comput. Commun.*, vol. 3, no. 3, pp. 67–74, 2008.

[4] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, 2008.

[5] A. Cichocki, R. Zdunek, and S. Amari, "Hierarchical als algorithms for nonnegative matrix and 3D tensor factorization," in *ICA*. Springer, 2007, pp. 169–176.

[6] V. Y. F. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1592–1605, 2013.

[7] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, 2009.

[8] O. Winther and K. B. Petersen, "Bayesian independent component analysis: Variational methods and non-negative decompositions," *Digit. Signal Process.*, vol. 17, no. 5, pp. 858–872, 2007.

[9] Z. Liu and V. Y. F. Tan, "Rank-one NMF-based initialization for NMF and relative error bounds under a geometric assumption," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4717–4731, 2017.

[10] Y. Li, Y. Liang, and A. Risteski, "Recovery guarantee of non-negative matrix factorization via alternating updates," in *NIPS*, 2016, pp. 4987–4995.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, vol. 46, John Wiley & Sons, 2004.

[12] Z. Liu, "Model selection for nonnegative matrix factorization by support union recovery," *arXiv preprint arXiv:1810.10078*, 2018.

[13] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

[14] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, 2014.

[15] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1135–1146, 2012.

[16] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 340–354, 2013.

[17] M. M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $\ell_{2,0}$ norm approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4646–4655, 2010.

[18] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Ann. Stat.*, vol. 39, no. 1, pp. 1–47, 2011.

[19] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[20] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," `http://cvxr.com/cvx`, Mar. 2014.

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