# A PROVABLE NONCONVEX MODEL FOR FACTORING NONNEGATIVE MATRICES

*Dung N. Tran, Sang P. Chin* Trac D. Tran†*

Johns Hopkins University
Department of Electrical and Computer Engineering

## ABSTRACT

We study the Nonnegative Matrix Factorization problem which approximates a nonnegative matrix by a low-rank factorization. This problem is particularly important in Machine Learning, and finds itself in a large number of applications. Unfortunately, the original formulation is ill-posed and NP-hard. In this paper, we propose a row sparse model based on Row Entropy Minimization to solve the NMF problem under separable assumption which states that each data point is a convex combination of a few distinct data columns. We utilize the concentration of the entropy function and the $\ell_\infty$ norm to concentrate the energy on the least number of latent variables. We prove that under the separability assumption, our proposed model robustly recovers data columns that generate the dataset, even when the data is corrupted by noise. We empirically justify the robustness of the proposed model and show that it is significantly more robust than the state-of-the-art separable NMF algorithms.

***Index Terms—*** separable nonnegative matrix factorization, sparse representation, row entropy minimization

## 1. INTRODUCTION

The *Nonnegative Matrix Factorization* (NMF) problem [1] aims to express an $n \times m$ nonnegative matrix $\boldsymbol{Y}$ as $\boldsymbol{DX}$, where $\boldsymbol{D}$ and $\boldsymbol{X}$ are nonnegative matrices of size $n \times s$ and $s \times m$, respectively, for some positive integer $s < \min\{n, m\}$. This problem finds itself in enormous number of applications in various fields such as blind source separation, topic modeling, recommendation systems, and clustering to name a few. Unfortunately, the original formulation is ill–posed [2] and NP–hard [3]. Most traditional methods rely on solving a non-convex optimization problem which lacks of optimality guarantee [4]. Therefore, provable algorithms for computing NMF under appropriate assumptions are of particular interest.

**Notation.** We use bold uppercase letters for matrices, and bold lowercase letters for column vectors. The notations $\mathbf{0}_k$ and $\mathbf{1}_k$ denote the all-zero and all-one vectors of length $k$, respectively. We let $\boldsymbol{I}_k$ be the identity matrix in $\mathbb{R}^{k\times k}$. Without subscripts, the sizes of these vectors and matrices should be inferred from the context. Given a matrix $\boldsymbol{Y}$, we let $\boldsymbol{y}_i$, $\boldsymbol{y}^j$, and $y_{i,j}$ denote its $i$-th column, $j$-th row and $(i, j)$ element, respectively. For an index set $\mathcal{S}$, the matrix $\boldsymbol{Y}_\mathcal{S}$ consists of the columns of $\boldsymbol{Y}$ whose indices supported by $\mathcal{S}$. The notation $\mathbb{R}_+$ denotes nonnegative numbers. Similar notations are used for higher dimensional vector spaces.

Recently, several provable algorithms have been proposed in literature based on the separable assumption, which allows the NMF problem to admit a unique solution [2]:

**Definition 1** (Separable NMF)**.** *A data matrix $\boldsymbol{Y}$ is $s$–separable if there exists a cone generated by a few columns of $\boldsymbol{Y}$ that contains the entire dataset.*

In this paper, we consider a variant of the separability which also offers a unique factorization up to some permutation of the data points.

**Assumption 2** (Convex hull assumption)**.** *Given a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ and some positive integer $s < \min\{n, m\}$, there exists an index set $\mathcal{S}$ of cardinality $s$ such that $\boldsymbol{Y} = \boldsymbol{Y}_\mathcal{S}\boldsymbol{Z}$, for some $\boldsymbol{Z} \in \mathbb{R}_+^{s\times m}$ satisfying $\mathbf{1}^\mathsf{T}\boldsymbol{Z} = \mathbf{1}^\mathsf{T}$. Without loss of generality (WLOG), we assume that $\mathcal{S} = \{1, ..., s\}$.*

That is, the entire dataset is contained in a convex hull generated by $s$ columns of the data matrix. This assumption was justified in several applications such as text modeling, hyperspectral unmixing, and blind source separation [5, 6, 7, 8, 9]. Throughout the paper, we assume that the vertices of this convex hull are *distinct*.

The equation in Assumption 2 can be rewritten as $\boldsymbol{Y} = \boldsymbol{YX}$, where $\boldsymbol{X} \in \mathbb{R}^{m\times m}$ such that each column of $\boldsymbol{X}$ sums to one, and at most $s$ rows of $\boldsymbol{X}$ are nonzero. These rows of $\boldsymbol{X}$ are supported by $\mathcal{S}$. Then solving the NMF problem under this assumption becomes finding the $s$ nonzero rows of $\boldsymbol{X}$ satisfying these above constraints. In the language of sparse representation, the problem can be modeled as

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_{\text{row},0} \quad \text{s.t.} \quad \boldsymbol{YX} = \boldsymbol{Y},\ \boldsymbol{X} \geq \mathbf{0},\ \mathbf{1}^\mathsf{T}\boldsymbol{X} = \mathbf{1}^\mathsf{T}, \quad (1)$$

where $\|\boldsymbol{X}\|_{\text{row},0}$ counts the number of nonzero rows of $\boldsymbol{X}$. The $distinct$ vertices of the dataset can then be identified by

*Thanks to NSF-DMS-1222567 and Air Force Office of Scientific Research (FA9550-12-1-0136) for funding.

†This work is partially supported by National Science Foundation under Grants ECCS-1443936 and CCF-1422995.

extracting the nonzero rows of an optimal solution returned by this row sparse problem.

Unfortunately, this problem is intractable and NP-Hard. Several algorithms have been introduced in literature [8, 10, 11, 12, 13, 14, 15, 16, 17] which aim to recover the vertex set by either solving easier linear programming or convex optimization problems, or by adopting a greedy pursuit method. In this paper, we address the intractability of (1) by proposing a non-convex relaxation to this problem. Specifically, we introduce a row sparsity measure based on the entropy function over the rows of the coefficient matrix. We demonstrate rigorously that by minimizing this measure under separability, one can robustly recover the vertices even when the data is corrupted by noise. As we will illustrate in the experiment section, our algorithm is remarkably more robust than the state-of-the-art algorithms for solving separable NMF.

## 2. ROW ENTROPY MINIMIZATION

We solve the NMF problem under Assumption 2 by considering the row sparse problem (1). To begin, for any matrix $\boldsymbol{X} \in \mathbb{R}^{m\times m}$, define $\boldsymbol{\nu}(\boldsymbol{X}) = [\|\boldsymbol{x}^1\|_\infty, ..., \|\boldsymbol{x}^m\|_\infty]^\mathsf{T}$. Then the sparsity of $\boldsymbol{\nu}(\boldsymbol{X})$ and the row sparsity of $\boldsymbol{X}$ are equivalent. To overcome the NP-hardness of (1), we propose to solve the following optimization problem named *Row Entropy Minimization (REM)*:

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_{h,\infty} \quad \text{s.t.} \quad \boldsymbol{Y}\boldsymbol{X} = \boldsymbol{Y},\ \boldsymbol{X} \geq \boldsymbol{0},\ \boldsymbol{1}^\mathsf{T}\boldsymbol{X} = \boldsymbol{1}^\mathsf{T}, \quad (2)$$

where $\|\boldsymbol{X}\|_{\mathrm{h},\infty} = h(\boldsymbol{\nu}(\boldsymbol{X}))$. Here, the *entropy function* $h(\cdot)$ is defined as

$$h(\boldsymbol{z}) = -\sum_i \frac{|z_i|}{\|\boldsymbol{z}\|_1} \log \frac{|z_i|}{\|\boldsymbol{z}\|_1}, \quad (3)$$

for any vector $\boldsymbol{z} \in \mathbb{R}^m$. We adopt the common convention that $0 \log 0 = 0$ and $h(0) = 0$. It was argued in [18, 19] that this function promotes the sparsity of its argument by skewing the signal energy towards a few of its elements. Therefore, a small value of the row entropy term $\|\boldsymbol{X}\|_{h,\infty}$ induces the row sparsity of $\boldsymbol{X}$.

More importantly, in practice, data is often contaminated by noise. In this case, we consider the following noisy model

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{N} = \boldsymbol{Y}_{\mathcal{S}}\boldsymbol{Z} + \boldsymbol{N}, \quad (4)$$

where $\boldsymbol{Y}, \boldsymbol{Y}_{\mathcal{S}}$ and $\boldsymbol{Z}$ are defined in Assumption 2, and $\boldsymbol{N} \in \mathbb{R}^{n\times m}$ is a bounded noise matrix. Here, each column of the noise matrix $\boldsymbol{N}$ is assumed to be bounded, i.e., $\|\boldsymbol{n}_j\|_2 \leq \epsilon$, for some small positive number $\epsilon$, and for every column $\boldsymbol{n}_j$ of $\boldsymbol{N}$. We thus find the vertices in noisy settings by first solving the following robust variant of REM:

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \|\boldsymbol{X}\|_{h,\infty} \quad (5) \\ \text{s.t.} \quad & \|\tilde{\boldsymbol{y}}_j - \tilde{\boldsymbol{Y}}\boldsymbol{x}_j\|_2 \leq 2\epsilon, \forall j = 1, ..., m, \\ & \boldsymbol{X} \geq \boldsymbol{0},\ \boldsymbol{1}^\mathsf{T}\boldsymbol{X} = \boldsymbol{1}^\mathsf{T}. \end{aligned}$$

The vertices can then be identified from the dominant rows of the optimal solution of this optimization problem. This procedure is summarized in Algorithm 1.

**Algorithm 1** Robust REM for Vertex Identification

**input:** Noisy data matrix $\tilde{\boldsymbol{Y}}$, the noise level $\epsilon$.
**output:** The estimated vertex set $\hat{\mathcal{S}}$ of the original data matrix.

1. Find the optimal solution $\boldsymbol{X}_*$ of the optimization problem (5).
2. Let $\hat{\mathcal{S}}$ be the index set corresponding to the $s$ rows of $\boldsymbol{X}_*$ with the largest $\ell_\infty$ norm.

In the next section, we will show that under the convex hull assumption, Algorithm 1 is guaranteed to exactly identify the vertices of the corrupted data matrix, provided that the noise power is relatively small. Before continuing, we would like to point out that the row entropy term $\|\cdot\|_{h,\infty}$ is not a norm as it does not satisfy the triangle inequality.

## 3. THEORETICAL GUARANTEES

In this section, we prove that REM is robust under small perturbation. To simplify the analysis, the columns of $\boldsymbol{Y}$ are assumed to be distinct. Also, let us define the margin parameter

$$\rho = \min_{j\notin\mathcal{S}, k\in\mathcal{S}} \|\boldsymbol{y}_j - \boldsymbol{y}_k\|_2, \quad (6)$$

which characterizes the isolation of the vertices. We assume that $\rho > 0$, meaning that the vertices are separated enough from the non-vertex data points. Furthermore, let

$$\gamma = \min_{k\in\mathcal{S}} \min_{\alpha\geq 0, \boldsymbol{1}^\mathsf{T}\alpha = 1} \|\boldsymbol{y}_k - \boldsymbol{Y}_{\mathcal{S}\setminus k}\alpha\|_2, \quad (7)$$

which bounds from below the distance from a vertex to the convex hull generated by the other vertices. In some sense, this parameter characterize the fatness of the polytope generated by the data vertices. Intuitively, large values of $\rho$ and $\gamma$ make the isolation of the vertices and the shape of the data polytope more robust to noise. This in turn makes it easier to identify the vertices. Finally, we assume that the data is bounded by a finite number defined by

$$\kappa = \max_{j=1,...,m} \|y_j\|_2. \quad (8)$$

We are now ready to state our main result.

**Theorem 3.** *Let $\boldsymbol{Y}$ be a data matrix satisfying Assumption 2. Suppose the data is corrupted by bounded noise, i.e., $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{N}$, where $\|\boldsymbol{n}_j\|_2 \leq \epsilon, \forall j = 1, ..., m$. If*

$$\epsilon < \frac{\rho\gamma}{8\kappa(s+1)}, \quad (9)$$

*then Algorithm 1 identifies the vertices of $\boldsymbol{Y}$ exactly.*

This result is consistent with the intuition discussed above. If $\rho$ or $\gamma$ is small, meaning that the vertices are less isolated from the others, or the data polytope is thin, the noise power must be small to retain the separability of the data. Furthermore, when the signal space is bounded, if the number of vertices $s$ is large, the data points tend to be close to each other. Therefore, the noise level must be small to guarantee vertex exact recovery.

The proof for the main theorem is given at the end of this section. The main ingredient of our analysis is the *concentration* property of the entropy function. The following lemmas formalize this important property.

**Lemma 4.** *Let $\boldsymbol{x} \in \mathbb{R}_+^m$ such that $0 \leq x_i \leq 1, \forall 1 \leq i \leq m$. Let $k$ and $l$ be two arbitrary distinct indices satisfying $x_k \geq x_l$. Define $\boldsymbol{x}(\delta) \coloneqq \tilde{\boldsymbol{x}}$ as*

$$\tilde{x}_i = x_i, \ \forall i \neq k, l; \quad \tilde{x}_k = x_k + \delta, \quad \tilde{x}_l = x_l - \delta,$$

*where $\delta$ is a small positive number such that $0 \leq \tilde{x}_k, \tilde{x}_l \leq 1$. Then $h(\boldsymbol{x}) > h(\boldsymbol{x}(\delta))$.*

In other words, concentrating signal energy on significant elements while dispersing energy from small elements decreases the value of the entropy function.

**Lemma 5.** *Let $\boldsymbol{x} = \begin{pmatrix}\mathbf{1}_k^\mathsf{T} & \mathbf{0}_{m-k}^\mathsf{T}\end{pmatrix}^\mathsf{T}$, for some $1 \leq k \leq m-1$, and $\boldsymbol{x}(\alpha) = \begin{pmatrix}\mathbf{1}_k^\mathsf{T} & \alpha & \mathbf{0}_{m-k-1}^\mathsf{T}\end{pmatrix}^\mathsf{T}$, for some $0 < \alpha \leq 1$. It follows that $h(\boldsymbol{x}) < h(\boldsymbol{x}(\alpha))$.*

Intuitively, the aforementioned lemmas suggest that when the vector elements are bounded from above, solutions of entropy function minimization tend to concentrate the energy on the least number of elements, resulting in a sparser solution. This is formalized in the lemma below. Its proof can be obtained by iteratively applying Lemma 4 and Lemma 5.

**Lemma 6.** *Let $\boldsymbol{x} = \begin{pmatrix}\mathbf{1}_k^\mathsf{T} & \mathbf{0}_{m-k}^\mathsf{T}\end{pmatrix}^\mathsf{T}$, and $\tilde{\boldsymbol{x}} = \begin{pmatrix}\mathbf{1}_k^\mathsf{T} & \boldsymbol{\alpha}^\mathsf{T}\end{pmatrix}^\mathsf{T}$, where $\boldsymbol{\alpha} \in \mathbb{R}_+^{m-k}$. If $\boldsymbol{\alpha}$ is nonzero, then $h(\boldsymbol{x}) < h(\tilde{\boldsymbol{x}})$.*

We are now ready to provide the proof sketch for Theorem 3.

**Proof sketch for Theorem 3.** Consider the noisy model (4), where $\boldsymbol{Y}$ satisfies Assumption 2, and $\boldsymbol{N}$ is a bounded noise matrix whose column energy is bounded by $\epsilon > 0$. Let $\boldsymbol{X}$ be a feasible solution of (5). Similar to the proof of Theorem 1 in [20], we can show that

$$\|\boldsymbol{x}^k\|_\infty \geq x_{kk} \geq 1 - \frac{8\epsilon\kappa}{\rho\gamma}, \forall k \in \mathcal{S}. \tag{10}$$

Let $\boldsymbol{Z}$ be the coefficient matrix defined in Assumption 2, then $\bar{\boldsymbol{Z}} = \begin{bmatrix}\boldsymbol{Z}^\mathsf{T} & \mathbf{0}\end{bmatrix}^\mathsf{T}$ is a feasible solution of (5). Let $\boldsymbol{X}_*$ be the optimal solution of (5). It follows that $\|\boldsymbol{X}_*\|_{h,\infty} \leq \|\bar{\boldsymbol{Z}}\|_{h,\infty}$. Iteratively applying Lemmas 4, 5, and 6, this implies, $\forall j \notin \mathcal{S}$,

$$\|\boldsymbol{x}_*^j\|_\infty \leq \sum_{j \notin \mathcal{S}} \|\boldsymbol{x}_*^j\|_\infty \leq s - \sum_{k \in \mathcal{S}} \|\boldsymbol{x}_*^k\|_\infty \leq \frac{8\epsilon\kappa}{\rho\gamma} s. \tag{11}$$

Therefore, if $\frac{8\epsilon\kappa}{\rho\gamma} s < 1 - \frac{8\epsilon\kappa}{\rho\gamma}$, or equivalently, $\epsilon < \frac{\rho\gamma}{8\kappa(s+1)}$, then $\|\boldsymbol{x}_*^j\|_\infty < \|\boldsymbol{x}_*^k\|_\infty, \forall j \notin \mathcal{S}, k \in \mathcal{S}$. In other words, the $s$ rows of $\boldsymbol{X}_*$ with the largest $\ell_\infty$ norm correspond to the vertices of $\boldsymbol{Y}$. This completes the proof for the theorem. □

## 4. ITERATIVE ALGORITHMS FOR REM

As we will show in the experiment section, solving robust REM leads to better solutions comparing to the state-of-the-art separable NMF algorithms. The main issue here is that the row entropy objective $\|\cdot\|_{h,\infty}$ in REM is nonconvex. We thus approximate the objective function by its first order approximation and utilize an iterative algorithm to solve a series of easier subproblems.

In a simplified setting, we denote $\boldsymbol{\nu} = \boldsymbol{\nu}(\boldsymbol{X})$ and $\boldsymbol{\nu}^t = \boldsymbol{\nu}(\boldsymbol{X}^t)$. Let $\boldsymbol{X}^t$ be the solution estimate at iteration $t$ of the algorithm, then the first order approximation of the objective function in REM is given by

$$\begin{aligned}\|\boldsymbol{X}\|_{h,\infty} = h(\boldsymbol{\nu}) &\approx h(\boldsymbol{\nu}^t) + \nabla h(\boldsymbol{\nu}^t)^\mathsf{T}(\boldsymbol{\nu} - \boldsymbol{\nu}^t) \\ &= \sum_i [\nabla h(\boldsymbol{\nu}^t)]_i \nu_i + h(\boldsymbol{\nu}^t) - \nabla h(\boldsymbol{\nu}^t)^\mathsf{T} \boldsymbol{\nu}^t.\end{aligned} \tag{12}$$

Recall that $\nu_i = \|\boldsymbol{x}^i\|_\infty, \forall 1 \leq i \leq m$, the next solution estimate can thus be obtained by solving

$$\min_{\boldsymbol{X}} \sum_i w_i^t \|\boldsymbol{x}^i\|_\infty \quad \text{s.t.} \quad \boldsymbol{Y}\boldsymbol{X} = \boldsymbol{Y}, \ \boldsymbol{X} \geq \mathbf{0}, \ \mathbf{1}^\mathsf{T}\boldsymbol{X} = \mathbf{1}^\mathsf{T}, \tag{13}$$

where $w_i^t = [\nabla h(\boldsymbol{\nu}^t)]_i, \forall 1 \leq i \leq m$. The following proposition shows that the weights have closed form and are easily computable [18].

**Proposition 7.** *Let $h$ be the entropy function defined in (3), and let $\boldsymbol{\nu}$ be a nonzero nonnegative vector, then*

$$\frac{\partial h(\boldsymbol{\nu})}{\partial \nu_i} = -\frac{\log \nu_i}{\|\boldsymbol{\nu}\|_1} + \frac{\sum_j \nu_j \log \nu_j}{\|\boldsymbol{\nu}\|_1^2}. \tag{14}$$

As a consequence,

$$w_i^t = -\frac{\log \nu_i^t}{\|\boldsymbol{\nu}^t\|_1} + \frac{\sum_j \nu_j^t \log \nu_j^t}{\|\boldsymbol{\nu}^t\|_1^2}, \tag{15}$$

for $\nu_i^t > 0$. When $\nu_i^t = 0$, we let $w_i^t = +\infty$. Moreover, the weights are dictated by the concentration behavior of the entropy function minimization. The following corollary summarizes this insight. It follows by the fact that $0 \leq \nu_i^t \leq 1$, $\forall 1 \leq i \leq m$.

**Corollary 8.** *If $\nu_i^t < \nu_k^t$, then $w_i^t > w_k^t$.*

In other words, small energy rows are given large weights at the next iteration, and are thus further suppressed. Therefore, at the end of the algorithm energy is concentrated only on a small subset of rows.

In noisy settings, the subproblems in robust REM can be written as

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \lambda \sum_i w_i^t \|\boldsymbol{x}^i\|_\infty \\ \text{s.t.} \quad & \|\tilde{\boldsymbol{y}}_j - \tilde{\boldsymbol{Y}}\boldsymbol{x}_j\|_2 \le 2\epsilon, \forall j = 1, ..., m, \\ & \boldsymbol{X} \ge \boldsymbol{0},\ \boldsymbol{1}^\mathsf{T}\boldsymbol{X} = \boldsymbol{1}^\mathsf{T}. \end{aligned} \tag{16}$$

Therefore, at each iteration of REM and its robust variant, we solve a weighted $\ell_{1,\infty}$ subproblem under the same constraints as the original problem. Problems (13) and (16) can be solved efficiently by an Alternating Direction Method of Multipliers (ADMM) approach [21]. The main steps of this iterative algorithm are summarized in Algorithm 2.

**Algorithm 2** Iterative Algorithms for Solving Robust REM

**input:** data matrix $\boldsymbol{Y}$, the noise level $\epsilon$.
**initialization:** $\boldsymbol{X}^0$.
**while** not converged **do**
1. *Update the weights*:

$$w_i^t = -\frac{\log \nu_i^t}{\|\boldsymbol{\nu}^t\|_1} + \frac{\sum_j \nu_j^t \log \nu_j^t}{\|\boldsymbol{\nu}^t\|_1^2},\ i = 1, ..., m. \tag{17}$$

2. *Update the estimate*: Set $\boldsymbol{X}^{t+1}$ to be the optimal solution of (16).
**end while**
**output:** Estimated solution $\boldsymbol{X}_* = \boldsymbol{X}^t$.

## 5. EXPERIMENTAL RESULTS

This section presents experimental results for REM algorithm. For the benchmarked algorithms, we use the implementations on the author's websites. We test the robustness of our proposed algorithm against noise on synthetic dataset. The experiment setting is similar to that in [15]. For each simulation, the data is generated as follows. Elements of each column of the vertex matrix $\boldsymbol{Y}_S \in \mathbb{R}^{n\times s}$ are sampled from a uniform distribution on $[0, 1]$. The coefficient matrix $\boldsymbol{Z} \in \mathbb{R}^{s\times m}$ has the form of $[\boldsymbol{I}_s, \boldsymbol{Z}']$ where $\boldsymbol{I}_s \in \mathbb{R}^{s\times s}$ is the identity matrix, and each column of $\boldsymbol{Z}' \in \mathbb{R}_+^{m\times(m-s)}$ follows from a Dirichlet distribution whose parameters are chosen from a uniform distribution on $[0, 1]$. The data matrix is generated by $\boldsymbol{Y} = \boldsymbol{Y}_S\boldsymbol{Z} + \boldsymbol{N}$ where each element of the noise matrix $\boldsymbol{N}$ is drawn from a Normal distribution, then is multiplied by some parameter $\beta$. In the experiments, we let $n = 5, m = 25$, and $s = 5$. The number of trials is 100.

Figure 1 shows the $\ell_\infty$ norm of the rows of typical solutions of REM when the data is corrupted by moderate and large noise. It can be seen that the energy of the solutions concentrates on the rows corresponding to the vertices, which is consistent with Theorem 3.

We next compare our proposed algorithm with the state-of-the-art near-separable NMF algorithms: XRAY [8], SPA [14], SNPA [15], and GVP [9]. Figure 2 shows the exact recovery rates of the algorithms. It can be seen that REM is significantly more robust than the others.

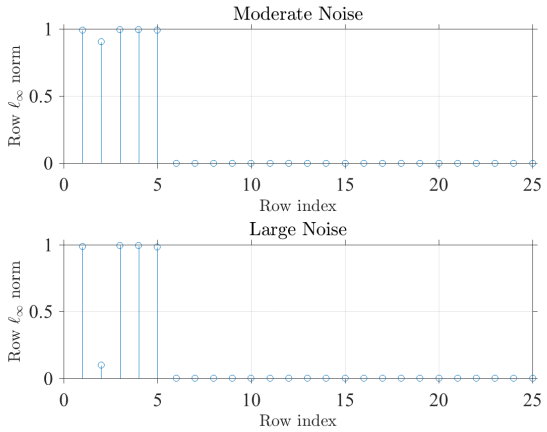


**Fig. 1**. Row $\ell_\infty$ norm of typical solutions of REM. *Top*: moderate noise. *Bottom*: large noise.

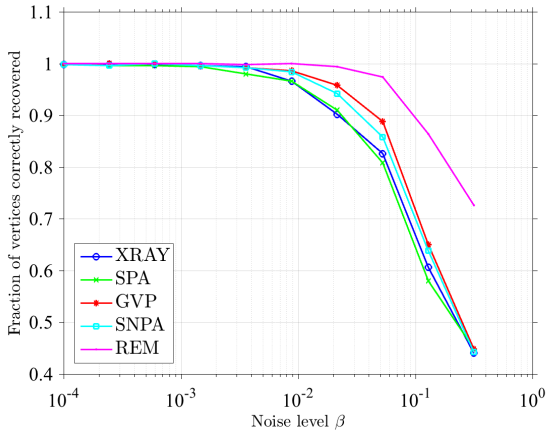


**Fig. 2**. Robustness comparison on synthetic data.

## 6. CONCLUSION

In this paper, we propose a row sparse model, namely Row Entropy Minimization, based on the entropy function to solve the separable NMF problem. We prove rigorously that, under separability , REM robustly recovers the vertices generating data. We propose an iterative algorithm to efficiently solve REM, which consists of a series of weighted $\ell_{1,\infty}$ subproblems. Finally, we show empirical evidences supporting our theoretical analysis. We show that REM is remarkably more robust than the state-of-the-art separable NMF algorithms.

## 7. REFERENCES


[1] D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] David Donoho and Victoria Stodden, "When does non-negative matrix factorization give correct decomposition into parts?," in *NIPS*. 2003, MIT Press.

[3] S. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM J. on Optimization*, vol. 20, pp. 1364–1377, 2009.

[4] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *NIPS*. 2000, pp. 556–562, MIT Press.

[5] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *ICML*, 2009, vol. 28, pp. 280–288.

[6] T.H. Chan, W.K. Ma, C.Y. Chi, and Y. Wang, "A convex analysis framework for blind separation of non-negative sources," *IEEE Trans. on Signal Processing*, vol. 56, pp. 5120–5134, 2008.

[7] J. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, pp. 354–379, 2012.

[8] A. Kumar, V. Sindhwani, and P. Kambadur, "Fast conical hull algorithms for near-separable non-negative matrix factorization," in *ICML*, 2012, vol. 28, pp. 231–239.

[9] Dung N. Tran, Tao Xiong, Sang P. Chin, and Trac D. Tran, "Nonnegative matrix factorization with gradient vertex pursuit," in *ICASSP*, 2015.

[10] S. Arora, R. Ge, R. Kannan, and A. Moitra, "Computing a nonnegative matrix factorization provably," in *Proceedings of the 44th symposium on Theory of Computing*, 2012, pp. 145–162.

[11] V. Bittorf, B. Recht, C. Re, and J.A. Tropp, "Factoring nonnegative matrices with linear programs," in *NIPS*, 2012, pp. 1223–1231.

[12] A. Kumar and V. Sindhwani, "Near-separable nonnegative matrix factorization with $\ell_1$ and bregman loss functions," in *arXiv:1312.7167*, 2013.

[13] N. Gillis and R. Luce, "Robust near-separable nonnegative matrix factorization using linear optimization," *Journal of Machine Learning Research*, vol. 15, pp. 1249–1280, Apr 2014.

[14] N. Gillis and S.A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 698–714, 2014.

[15] N. Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation," *SIAM J. on Imaging Sciences*, vol. 7, pp. 1420–1450, 2014.

[16] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Transactions on Image Processing*, vol. 21, pp. 3239–3252, 2012.

[17] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[18] Dung N. Tran, Shuai Huang, Sang P. Chin, and Trac D. Tran, "Low-rank matrices recovery via entropy function," in *ICASSP*, 2016.

[19] Shuai Huang, Dung N. Tran, and Trac D. Tran, "Sparse signal recovery based on nonconvex entropy minimization," in *ICIP*, 2016.

[20] Wing-Kin Ma Xiao Fu, "Robustness analysis of structured matrix factorization via self-dictionary mixed-norm optimization," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 60–64, 2016.

[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2011.