Non-negative Matrix Factorization under Heavy Noise

Chiranjib Bhattacharyya† CHIRU@CSA.IISC.ERNET.IN
Navin Goyal‡ NAVINGO@MICROSOFT.COM
Ravindran Kannan‡ KANNAN@MICROSOFT.COM
Jagdeep Pani† PANI.JAGDEEP@GMAIL.COM
† Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India
‡ Microsoft Research India

Abstract

The Noisy *Non-negative Matrix factorization* (NMF) is: given a data matrix A ($d \times n$), find non-negative matrices B, C ($d \times k$, $k \times n$ respy.) so that $A = BC + N$, where N is a noise matrix. Existing polynomial time algorithms with proven error guarantees require *each* column $N_{\cdot,j}$ to have l_1 norm much smaller than $||(BC)_{\cdot,j}||_1$, which could be very restrictive. In important applications of NMF such as Topic Modeling as well as theoretical noise models (eg. Gaussian with high σ), almost *every* column of $N_{\cdot j}$ violates this condition. We introduce the *heavy noise* model which only requires the average noise over large subsets of columns to be small. We initiate a study of Noisy NMF under the heavy noise model. We show that our noise model subsumes noise models of theoretical and practical interest (for eg. Gaussian noise of maximum possible σ). We then devise an algorithm **TSVDNMF** which under certain assumptions on B, C, solves the problem under heavy noise. Our error guarantees match those of previous algorithms. Our running time of $O((n+d)^2k)$ is substantially better than the $O(n^3d)$ for the previous best. Our assumption on B is weaker than the "Separability" assumption made by all previous results. We provide empirical justification for our assumptions on C. We also provide the first proof of identifiability (uniqueness of B) for noisy NMF which is not based on separability and does not use hard to check geometric conditions. Our algorithm outperforms earlier polynomial time algorithms both in time and error, particularly in the presence of high noise.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

1. Introduction

Let A be a $d \times n$ matrix (where each column, $A_{\cdot,j}$, is a d dimensional data-point) with non-negative real entries. *Exact NMF* is the problem of factoring A into the product BC of two non-negative matrices, with k columns in B. k is generally small. So, NMF would find the small number of "basis vectors" (columns of B) with each data point a non-negative combination of them. This has led to the applicability of NMF (Gillis, 2014).

There has been recent interest in developing polynomial time bounded algorithms with proven error bounds under specialized assumptions on the data (Arora et al., 2012; Gillis & Luce, 2014; Recht et al., 2012; Rong & Zou, 2015). All such algorithms require *separability* assumption, first introduced in (Donoho & Stodden, 2003). An NMF BC is *separable* if after a permutation of rows of A and B, the top k rows of B form a non-singular diagonal matrix D_0. Using separability (Donoho & Stodden, 2003) showed that B is essentially *identifiable* (i.e., unique) given A. (Arora et al., 2012) observed that if A has an exact separable NMF, and its rows are in general position, the rows of A corresponding to D_0 can be identified by solving d Linear Programs and then B can be found. (Recht et al., 2012) gave a clever reduction to a single linear program in n^2 variables. (Gillis & Luce, 2014) made important theoretical and practical improvements to the Linear Program. (Rong & Zou, 2015) introduces *subset separability*, a milder assumption than *separability*, but requires solving many convex programs. Algorithms based on Linear and Convex programs are hard to scale (Gillis, 2014) because of their high complexity. Besides separability, all algorithms make the stronger assumption of reasonably large diagonal entries in D_0.

In practice, A is not exactly factorable into BC, but $A = BC + N$, where, N is the noise matrix. (As before, B, C have non-negative entries, but N, A can have negative entries.) An algorithm for the *Noisy NMF* problem seeks to

compute a $d \times k$ matrix $\tilde{B}$ [1] for which

$$||B - \tilde{B}||_1 \leq \varepsilon, \quad (1)$$

where B is the true matrix. [Once $\tilde{B}$ is found, all algorithms find a $\tilde{C}$ by minimizing $||A - \tilde{B}\tilde{C}||_1$.] Existing separability-based algorithms rapidly deteriorate in the presence of noise. They first identify a set K of k special rows and columns of the data matrix A (which correspond to D_0) and use only K to find $\tilde{B}$. Noise in any one of the n data columns can change the choice of K and thus the answer. So, they all have to require that the l_1 norm of *each* column N be much smaller than the l_1 norm of the column of BC. Substantial noise in a SINGLE data-point can destroy the model hypothesis. But, in important applications like Topic Modeling, for almost all j, $||N_{\cdot,j}||_1 \approx ||(BC)_{\cdot,j}||_1$!

In this paper we introduce the following noise model, which we will refer to as Heavy noise model in the sequel.

$$\forall T \subseteq [n] \ , \ \text{ with } |T| \geq \varepsilon n \ , \ \frac{1}{|T|} \left|\left| \sum_{j \in T} N_{\cdot,j} \right|\right|_1 \leq \varepsilon. \quad (2)$$

We assume without loss of generality that each column of B sums to 1 and each column of C sums to at most 1. (See Remark 1). We require that the noise be small only when averaged over $\Omega(\varepsilon n)$ columns of N (instead of for each column).

We study *Noisy NMF* under the Heavy noise condition. More formally we pose the following problem.

Problem Definition: If A satisfies $A = BC + N$ find $\tilde{B}$ which can satisfy (1), under Heavy noise (2).

Contributions: As mentioned earlier existing separability based algorithms will not be able to solve the problem of interest. We devise algorithm TSVDNMF which solves the Noisy NMF problem under heavy noise.

- **Heavy Noise Model:** Since the noise condition (2) only bounds noise in an average of many columns, it can hold even when the condition $||N_{\cdot,j}||_1 \leq \varepsilon$ (imposed by earlier papers) is violated for almost all j. Indeed, we prove that our noise model (2) subsumes several models of theoretical and practical interest, including spherical Gaussian noise (with the amplitude of the noise much greater than the data), general Gaussian noise and multinomial noise (arising in Topic Modeling) among others. The proof of these is based on recent results bounding the eigenvalues of random matrices (with independent vector-valued random variables as columns) and some standard tools from Probability - concentration inequalities (Höffding-Azuma). The word "heavy" refers to the fact that noise can be as large as or even larger than data in individual columns.

- **Dominant NMF** and **TSVDNMF** Like previous papers, we make assumptions on B as well as C. On B, our assumptions weaken separability in two ways. Instead of a diagonal matrix D_0, we require only a diagonally dominant matrix - the off-diagonal entries are to be smaller, not necessarily 0. Also instead of a single row $i(l)$ for each column l of B with $B_{i(l),l} \neq 0$, we allow a set S_l of rows. In the context of Topic Modeling, this would be multiple "anchor words", an issue which has been an open question in earlier papers. This is the Dominant Feature Assumption (D1) spelt out later.

  We make two assumptions on C. Each column must have a dominant entry. We call this the Dominant Basis Vector Assumption (D2). Furthermore, there must be a (small) fraction of "nearly pure records", namely, for each $l = 1, 2, \ldots k$, there are at least $\varepsilon_0 n$ columns of C with $C_{lj} \geq 1 - \varepsilon$, see Assumption D3 later. If a matrix pair B, C satisfies assumptions D1-3 then it is said to be Dominant NMF family.

  Dominant NMFs can be recovered in presence of heavy noise by **TSVDNMF** devised in this paper. **TSVDNMF** is based on three crucial steps: (i) Thresholding, (ii) Clustering based on SVD, and (iii) identifying dominant basis vectors and then Dominant Features and then nearly pure records. Our main contribution is to show that **TSVDNMF** recovers the correct NMF if data satisfies the Dominant NMF assumption. The overall complexity is $O((n+d)^2 k)$ (compared to the $O(n^3 d + n^{2.5} d^{1.5})$ of best previous algorithm). Note that the SPA algorithm of Gillis and Vavasis (Gillis & Vavasis, 2014) has a better complexity than $O(n^3 d)$, however they do not solve our problem; they only guarantee that the 2-norm of $\tilde{B}_{\cdot,l} - B_{\cdot,l}$ is small, not the 1-norm. Both Dominant NMF and **TSVDNMF** are inspired from (Bansal et al., 2014) but have crucial differences.

- **Identifiability** One subtle but important remark about our algorithmic contribution above is that it seems to suggest the uniqueness of the factorization B because of (1). This, however, is only implied for B, C satisfying the Dominant NMF model, and does not follow for unrestricted NMFs. We prove that in fact such a uniqueness result holds even for the unrestricted NMF. This is clearly an important and desirable characteris-

[1] For a matrix M, $||M||_1$ denotes the maximum l_1 norm of a column of M. In the introduction, we use ε to denote a generic small quantity and each instance of it may have different values.

tic of the problem making it well-posed.

(Donoho & Stodden, 2003) first argued the importance of Identifiability (uniqueness of B), introduced the notion of separability and showed that it implies identifiability. Indeed, under separability, since, every row of C has a scaled copy of it in A (and if the rows of A are in general position), then, these rows are the extreme rays of the cone spanned by all the rows. So, by Linear Programming, we can identify them. With our noise model, since single columns/rows can be extremely corrupted, this argument does not work anymore. We supply the first proof of (approximate) identifiability (Theorem 2) under a set of assumptions on B, C similar to the ones we use for computation.

2. Heavy Noise subsumes several noise models

As opposed to previous noise models, ours subsumes several noise models.

Independent Gaussian Noise: Suppose the entries of N are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. We prove the following:

Lemma 1. Informal Statement

- *If $\sigma \in O(1/\sqrt{d})$, then, (2) holds whp.*
- *If $\sigma \geq \Omega(1/\sqrt{d})$, the MLE of B violates (1) (even with $k = 1$).*
- *If $\sigma \geq \Omega(1/d)$, then, for MOST j, $||N_{\cdot,j}||_1 \geq c_5$.*

Since MLE is the "best possible" the second part is an impossibility result - noise greater than $\Omega(1/\sqrt{d})$ cannot be tolerated. Lemma 1 and Lemma 2 in Supplement specify precisely what is inside O, Ω . The proof of the first part draws on Random Matrix Theory, in particular, bounds on the largest eigenvalue. The other parts are simple.

General Correlated Noise: While noise in different data points may be independent, noise in different coordinates of the same data point need not be. We can model this more general case by having $N_{\cdot,j}$ be independent, not neccessarily identical, vector-valued random variables. Suppose Σ_j is the covarience matrix of $N_{\cdot,j}$ for $j = 1, 2, \ldots, n$. We prove (see Supplement) :[2]

Lemma 2. *If $||\Sigma_j||_2 \leq O(1/\sqrt{d})$ then (2) holds whp.*

This is tight by the impossibility result earlier.

Heavy Noise In the above two cases, noise was bounded by $O(1/\sqrt{d})$ in *each* direction. Here, we will only assume $||N_{\cdot,j}|| \in O(1)$ (comparable to data). But l_2 length won't do (as is seen from part 2 of Lemma 1).

[2]For a matrix M, $||M||_2$ is the spectral norm.

Lemma 3. *Suppose $||C_{\cdot,j}||_1 = ||A_{\cdot,j}||_1 = 1$ and suppose each column $N_{\cdot,j}$ of N is the average of $m \geq 8c_0^2/\varepsilon^4$ independent zero-mean vector-valued random variables $N_{\cdot,j}^{(1)}, N_{\cdot,j}^{(2)}, \ldots, N_{\cdot,j}^{(m)}$- with $||N_{\cdot,j}^{(t)}||_1 \leq c_0$, c_0 any constant. Then, (2) is satisfied whp for large enough n.*

The proof (Supplement) is based on the bounded difference inequality for Martingales. (See for example (McDiarmid, 1989)). This still allows $\|N_{\cdot,j}\|_1$ to be $\Omega(1)$. See example under Multinomial.

Multinomial Noise (A special case of Heavy noise): Assume $||C_{\cdot,j}||_1 = ||A_{\cdot,j}||_1 = 1$ (in addition to $||B_{\cdot,j}||_1 = 1$). Also assume there is $d \times n$ matrix P with non-negative entries and column sums 1. P could be just BC. $N_{\cdot,j}$ is the average of m i.i.d. Multinomial trials, each of which picks a unit vector e_i with probabilities proportional to $P_{i,j}$. Subtract the mean to make $E(N_{ij}) = 0$. So,

$$E(N_{ij}) = 0 \; ; \; \text{Var}(N_{ij}) = \frac{P_{ij}}{m}.$$

In Topic Modeling, the process above has picked m words in each document. Usually, $m << n, d$. Consider the following example, for any fixed j, if $d/2$ of the P_{ij}'s are $\frac{1}{4d}$ each, $N_{\cdot j}$ is likely to have $\frac{1}{m} - \frac{1}{4d}$ in $m/2$ of these coordinates and so $\|N_{\cdot j}\|_1 \geq 0.5$.

Adversarial Noise: All the above noise models are stochastic. A non-stochastic noise model which is also a special case of our noise model:

Lemma 4. *Suppose we have adversarial (not random) corruption (up to l_1 norm 1 each) of at most an ε^2 fraction of all data points. Then, (2) is satisfied.*

For example in Image Processing, this allows for an ε^2 fraction of images to have large corruption which can be used to model occlusion (Oh et al., 2008).

3. Dominant NMF model

We say that A, B, C, N with $A = BC + N$ and B, C nonnegative, satisfy the dominant NMF model if they satisfy the following four conditions.

We call the rows of A (indexed by i) "features" and the columns (indexed by j) "records." The columns (indexed by ℓ) of B are called "basis vectors."

Remark 1. *We may assume without loss of generality that each column of B sums to 1 and each column of C sums to at most 1. We can divide each column of B by its l_1 length (ensuring $||B_{\cdot,j}||_1 = 1$) and multiply the corresponding row of C by the same amount and preserve BC. Then, we can scale each column of C (and correspondingly each column of A, N) so as to ensure $||C||_1 \leq 1$. Note that this only scales the entire error.*

In the following, $0 < \epsilon, p_0, \gamma, w_0, \rho, \alpha, \beta < 1$ etc. are constants; ϵ's should be thought of as being small.

D1. Dominant Features: There are k disjoint sets of features, $S_1, S_2, \ldots, S_k \subset [d]$, such that (a) $\forall i \in S_l,\ \forall l' \neq l,\ B_{il'} \leq \rho B_{il}$, (b) $\sum_{i \in S_l} B_{il} \geq p_0$, and (c) $\forall i \in S_l, B_{il} \geq \gamma$. This assumption can be seen as an aggregate version of the separability condition: instead of asking for an anchor word in each topic we ask for catchwords (this assumption is essentially the catchwords assumption in (Bansal et al., 2014)). This is also a relaxation of the assumption in (Lee & Seung, 1999) who require the basis vectors to have disjoint support.

D2. Dominant Basis Vectors There is a partition of $[n]$ into $T_1, T_2, \ldots, T_k$ satisfying (a) $\forall j \in T_l, l' \neq l, C_{l,j} \geq \alpha$ and $C_{l',j} \leq \beta$, and (b) $\forall l, |T_l| \geq w_0 n$. This is similar to the dominant topic assumption in (Bansal et al., 2014) which says that each document has a dominant topic.

D3. Nearly Pure Records Assumption. For each l, there is a set P_l of at least $\varepsilon_0 n$ records in each of which the l'th basis vector has coefficient at least $1 - \varepsilon_4$. I.e., $\forall l, \exists \geq \varepsilon_0 n\ j$ with $C_{lj} \geq 1 - \varepsilon_4$.

This is similar to the previous assumption. It is true in the context of LDA (with low hyperparamter). The previous two conditions allow for the robustness of our results. In Section 5.1 we provide empirical evidence that real life corpora often have reasonable fraction, more than 2%, of pure documents. Also in the presence of heavy noise the condition is neccessary, please see Discussion at end of Section 1 in Supplementary.

D4. Noise We assume N satisfies the following quantitative version of (2):

$$\forall T \subseteq [n]\ ,\ |T| \geq \varepsilon_4 n\ ,\ \frac{1}{|T|} \left\| \sum_{j \in T} N_{\cdot,j} \right\|_1 \leq \varepsilon_4^2. \qquad (3)$$

A matrix pair B, C is defined to be part of the **Dominant NMF** family if they satisfy Assumptions **D1-3**.

4. TSVDNMF : A SVD based algorithm for NMF

In this section we describe an algorithm which yields an NMF from the SVD decomposition of matrix obtained by carefully thresholding the data-matrix A. We will also show that it provably recovers B under heavy noise if B, C comes from a Dominant NMF family.

The algorithm depends on following parameters $\alpha, \beta, \rho, \varepsilon, \varepsilon_0, \varepsilon_4, \gamma$ in $(0, 1)$ and $\nu > 1$. The constant ν is defined in the algorithm. The algorithm effectively requires only ε_0, α and ν. We will discuss the choice of parameter values after presenting the algorithm.

We assume the constants satisfy:
$\varepsilon < \varepsilon_0/20; \beta + \rho \leq (1 - 5\varepsilon)\alpha; \varepsilon_4 \leq \frac{\alpha\gamma\varepsilon}{2}, \frac{\varepsilon_0 w_0 p_0}{16k^3}, \frac{\varepsilon_0^2}{4}$.

TSVDNMF **Input:** $\mathbf{A}, k, \alpha, \varepsilon, \varepsilon_4, \beta + \rho, \varepsilon_0$ **Output:** Basis matrix $\mathbf{B}$.

1. **Thresholding**: Apply the Thresholding procedure (see below) to get $\mathbf{D}$.
2. **SVD**: Find the best rank k approximation $\mathbf{D}^{(k)}$ to $\mathbf{D}$.
3. **Identify Dominant Basis Vectors for each record**:
   (a) **Project and Cluster** Find (approximately) optimal k-means clustering of the columns of $\mathbf{D}^{(k)}$.
   (b) **Lloyd's Algorithm** Using the clustering found in Step 3(a) as the starting clustering, apply Lloyd's k-means algorithm to the columns of $\mathbf{D}$ ($\mathbf{D}$, not $\mathbf{D}^{(k)}$).
   (c) Let $R_1, R_2, \ldots, R_k$ be the $k-$partition of $[n]$ corresponding to the clustering after Lloyd's.
4. **Identify Dominant Features for each basis vector**:
   (a) For each i, l, compute $g(i, l) =$ the $(\lfloor \varepsilon_0 n/2 \rfloor)$th highest element of $\{A_{ij} : j \in R_l\}$.
   (b) $J_l = \{i : g(i, l) > \text{Max}(\gamma - 2\varepsilon_4, \text{Max}_{l' \neq l} \nu g(i, l'))\}$, where, $\nu = \frac{1-\alpha\varepsilon}{\beta+\rho+2\alpha\varepsilon}$.
5. **Find Basis Vectors** Find the $\lfloor \varepsilon_0 n/4 \rfloor$ highest $\sum_{i \in J_l} A_{ij}$ among all $j \in [n]$ and return the average of these $A_{\cdot,j}$ as our approximation $\hat{B}_{\cdot,l}$ to $B_{\cdot,l}$.

4.1. Thresholding Procedure

1. Initialize $R := [d]$. /* R is the set of unpruned words.*/
2. For each i,
   (a) compute ν_i the $(1 - \frac{\varepsilon_0}{2})-$fractile of row i of A. Let $\zeta_i := \alpha\nu_i - 2\varepsilon_4$. ($\zeta_i$ is the threshold for row i of A.)
   (b) If $\zeta_i \geq 0$, set $W_i := \{j : A_{ij} \geq \zeta_i\}$. Set $D_{ij} := \sqrt{\zeta_i}$ for $j \in W_i$ and $D_{ij} := 0$ for $j \notin W_i$.
   (c) If $\zeta_i < 0$, then, set $W_i := \emptyset$; $D_{ij} := 0\, \forall j$; $R := R \setminus \{i\}$.
3. Sort the $|W_i|$ in ascending order. For convenience, renumber the i so that now $|W_i|$ are in the ascending order.
4. For $i = 1, 2, \ldots,$ in R: (If $W_i \tilde{\subseteq} W_{i'}$, we "prune" i' by zeroing out all entries not in W_i.)

- For $i' > i$ with $i' \in R$, and $|W_i| \leq |W_{i'}| - \varepsilon_0 n/8$, if $W_i \tilde{\subseteq} W_{i'}$, [3] set $D_{i',j} := 0$ for all $j \in W_{i'} \setminus W_i$; delete i' from R.

D is the $d \times n$ matrix after thresholding.

Remark The pruning step as stated above takes $O(nd^2)$ time. This can be reduced to $O^*(d^2)$ (assuming $\varepsilon_0/k \in \Omega(1)$) by the following: Instead of comparing W_i to $W_{i'}$, we just pick a uniform random sample T of $O(k \log d/\varepsilon_0)$ j 's at the outset and only compare the sets $W_i \cap T$ and $W_{i'} \cap T$. The proof that this succeeds with high probability is simple and we postpone that to the final paper.

Constants The value of ε_4 in the thresholding step is not sensitive to the recovery of the factorization if it is small enough. In our experiments we have chosen values so that $\gamma = 2\varepsilon_4$. We fixed the other constants as $\varepsilon_0 = 0.04$, $\alpha = 0.9$, $\nu = 1.15$ in all the experiments. Note that a proper tuning to find the constants may improve the result further.

Theorem 1. *Given a $d \times n$ data-matrix A and under the assumptions* **D1-4**, *the algorithm* **TSVDNMF** *finds for each l, an approximation $\hat{B}_{\cdot,l}$ satisfying*

$$\left| B_{\cdot,l} - \hat{B}_{\cdot,l} \right|_1 \leq \varepsilon_0.$$

The proof of the above theorem appears in the supplement.

Intuition for the algorithm. Thresholding is a crucial part of the algorithm. It would ideally ensure that the thresholded matrix D is a block matrix with non-zero entries precisely in $S_l \times T_l$. But with errors, this is only approximately true. Also, since there are many features which are not dominant for any basis vector, the effect of thresholding on them is more difficult to control. Our careful choice of thresholds and the pruning help in ensuring block diagonality for these features too. Then we apply SVD + k-means which clusters records according to dominant basis vector. Now taking the average of each cluster (which is what one does normally in hard clustering) does not work here, because here, basis vectors are essentially the extreme points of each cluster. We tackle this by identifying dominant features first, then identifying records which have the largest component in these features. We show that these records are near the extreme points and show that their average does the job.

Discussion: The Dominant NMF model closely parallels the work of (Bansal et al., 2014) on Topic Modeling, in particular Dominant Features are inspired by Catchwords while Dominant Basis Vectors are inspired by Dominant topics. **TSVDNMF** is also similar to **TSVD** proposed in (Bansal et al., 2014). However there are significant differences with (Bansal et al., 2014). There are at least three important new aspects in this paper. Firstly, the main purpose of this paper is to handle NMF under heavy noise. We introduce a heavy noise model which subsumes large variety of noise models. The second is the thresholding step: in (Bansal et al., 2014) an assumption called 'no-local-min" was crucially used in this step. For NMF, that assumption does not hold and thresholding instead uses a new step **Step 4** in **TSVDNMF** ; its properties are proved in Lemmas 6 and 7 of the Supplementary and this forms a crucial part now of the proof of the Main Theorem. The third important new aspect is the proof of Identifiability which is not considered at all in (Bansal et al., 2014). Further, the randomness of C was crucial throughout the proofs in (Bansal et al., 2014), but in the NMF setting, there is no stochastic model of C. So, the proofs of correctness of Steps 4 and 5 of **TSVDNMF** are now completely different

[3] If $W, W' \subseteq [n]$, we write $W \tilde{\subseteq} W'$ to denote: $|W \setminus W'| \leq \varepsilon_0 n/4$

4.2. Identifiability of Dominant NMF

In their influential paper (Donoho & Stodden, 2003), Donoho and Stodden consider the question of uniqueness of nonnegative matrix factorization for exact NMF. There has been considerable work on understanding uniqueness conditions since then; we refer to Huang et al. (Huang et al., 2014) and Gillis (Gillis, 2014) and references therein for an up-to-date review of the literature. Some of these conditions are necessary and sufficient; unfortunately they do not seem to be easy to check or use. These conditions are often geometric in nature and not directly related to the application at hand. Donoho et al. gave a set of necessary conditions called Separable Factorial Articulation Family including separability. Laurberg et al. (Laurberg et al., 2008) gave further such conditions and also studied the approximate NMF case.

By identifiability of approximate NMF, we mean the following: Given k and a $d \times n$ matrix A, a $d \times k$ matrix B is (approximately) *identifiable* from A if there exists a C with $||A - BC||_1 \leq \varepsilon$ and for any B', C' ($d \times k$ and $k \times n$ respectively) with $||A - B'C'||_1 \leq \varepsilon$, we have $||B - B'||_1 \leq \varepsilon'$, where, $\varepsilon' \to 0$ as $\varepsilon \to 0$ and we may need to permute the columns of B' appropriately.

Under the above conditions, we have the following theorem.

Theorem 2. *Suppose that B, C are $d \times k$ and $k \times n$ matrices with non-negative entries where each column of B sums to 1 and they satisfy Assumptions* **D1** *and* **D3**. *Suppose that $B' \in \mathbb{R}_+^{d \times k}$ and $C' \in \mathbb{R}_+^{k \times n}$ satisfy*

$$\| \sum_{j \in R_\ell} (BC)_{\cdot,j} - \sum_{j \in R_\ell} (B'C')_{\cdot,j} \|_1 \leq \delta \sum_{j \in P_\ell} \|C_{\cdot,j}\|_1 \forall \ell. \tag{4}$$

Let $\delta' := 2\epsilon_4 + 6\delta$. Also assume that the following conditions are satisfied: (a) $(p_0 - \delta')^2 > 4k(p_1 + \delta')$, and

(b) $2\delta' < p_0 - p_1$. *Then* B *is close to* B' *up to permutations and scalings of the columns. More precisely, there exist a permutation* $\pi : [k] \to [k]$ *and constants* $\alpha_j \in \mathbb{R}$ *for* $j \in [k]$ *such that*

$$\|B_{.,j} - \alpha_j B'_{.,\pi(j)}\|_1 \leq 2\delta' + \frac{4k(p_1 + \delta')}{p_0 - \delta'}.$$

Proof. See Theorem 3 in Supplement. □

We remark that the conditions used for the identifiability result are not exactly the same as those for our algorithmic results. This is essential and is explained in Supplement after the proof of Theorem 3.

5. Experiments

In this section we do a comprehensive empirical evaluation of **TSVDNMF** on synthetic and real datasets. We follow (Gillis & Luce, 2014) for experiments on synthetic datasets and (Mizutani, 2014) on real-world datasets. Our main objectives are three fold. More specifically

- **Checking of Dominant NMF assumptions:** In section 5.1 we show that Dominant Basis assumption and Pure record assumption (as described in section 3) are satisfied by the existing state of the art NMF algorithms on often used real world datasets in literature (Kumar et al., 2013; Mizutani, 2014).
- **Quality of Factorization recovered by TSVDNMF under heavy noise:** Following (Gillis & Luce, 2014) we generate matrices B, C and add noise N to generate the observed matrix A. We compare the factorizations of state of the art provable algorithms with **TSVDNMF** . We experiment with various noise models and report results in Section 5.2.
- **Performance of TSVDNMF on real world datasets:** In section 5.3, we compare **TSVDNMF** on the task of clustering on various datasets used in (Kumar et al., 2013; Mizutani, 2014).

Experimental setup: To find B in separable algorithms and C in **TSVDNMF** , we used non-negative least squares method as proposed in (Kumar et al., 2013). All experiments are performed using Matlab on a system with 3.5 GHz processor and 8GB RAM.

5.1. Checking Dominant NMF assumptions on real data

To check the validity of the dominant basis vector assumption and the pure records assumption (as defined in **D2** and **D3** of section 3), we run the existing state-of-the-art NMF algorithm PW-SPA (Gillis & Ma, 2014) on the following text and face datasets, and analyze the coefficient matrix C.

Table 1. Checking validity of Dominant Basis Assumption and Pure Records Assumption: % records satisfying dominant basis vector assumption with $\alpha = 0.4$ and $\beta = 0.3$ and % records satisfying pure record assumption with $\epsilon_4 = 0.05$

Datasets	**% Records: Dominant basis vectors**	**% Pure records**
Reuters10	60.97	10.11
Reuters48	35.08	2.28
20NG	55.73	7.02
TDT2	64.56	7.70
Yale	78.78	14.54

.

Dataset: Reuters [4], 20 Newsgroups [5], TDT-2 [6], Yale face dataset[7].

In the literature (Kumar et al., 2013; Mizutani, 2014), Reuters has been truncated to have only largest 10 or largest 48 classes. We consider both the versions and refer them as Reuters10 and Reuters48 respectively. The Yale dataset contains 15 subjects with 11 images per subject (total 165 faces), one per different facial expression or configuration. We chose the vocabulary size as 4096 while creating a bag of words representation of the dataset from the SIFT features. We further preprocessed the data by removing standard stop-words and words with less than 5 occurrences in the corpus. A brief statistics of the datasets after preprocessing is (1) Reuters10 ($n = 7285, d = 12418, k = 10$), (2) Reuters48 ($n = 8258, d = 13647, k = 48$), (3) 20 Newsgroups ($n = 18846, d = 24287, k = 20$), (4) TDT-2 ($n = 9394, d = 20687, k = 30$). (5) Yale ($n = 165, d = 4096, k = 15$). We used the tf-idf representation (Manning et al., 2008) to construct the data matrix A. We do not normalize the data in any form.

Observation: The percentage of records satisfying the dominant basis assumption with $\alpha = 0.4$ and $\beta = 0.3$, and the pure record assumption with $\epsilon_4 = 0.05$ are reported in Table 1 for different datasets. We see significant amount of records satisfy the two assumptions and conclude that the existing NMF method PW-SPA recovers the coefficient matrix satisfying the proposed assumptions whence the assumptions are empirically realistic.

[4] archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[5] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

[6] http://www.itl.nist.gov/iad/mig/tests/tdt/1998

[7] http://vision.ucsd.edu/content/yale-face-database

Table 2. Synthetic experiment under heavy noise: Data is generated using either separability or dominant assumption. Noise added is either Gaussian or Multinomial (a special case of heavy-tailed noise). Higher value signifies better performance.

	Separable data + Gaussian Noise			**Dominant data + Gaussian Noise**			**Separable data + Multinomial Noise**			**Dominant data + Multinomial Noise**		
$\tilde{\beta}$ or m	$\tilde{\beta} = 0.5$	1	2	$\tilde{\beta} = 0.5$	1	2	$m = 10$	60	100	$m = 10$	60	100
TSVDNMF	0.759	**0.659**	**0.402**	**0.757**	**0.478**	**0.114**	**0.094**	**0.587**	**0.654**	**0.017**	**0.51**	**0.605**
Gillis-LP	0.725	0.577	0.27	0.554	0.334	0.006	0.056	0.488	0.603	-0.017	0.364	0.443
PW-SPA	0.761	0.568	0.249	0.619	0.424	0.054	0.054	0.529	0.619	-0.014	0.423	0.497
SPA	0.764	0.566	0.246	0.618	0.437	0.074	0.056	0.527	0.616	-0.012	0.416	0.493
ER-SPA	**0.765**	0.57	0.252	0.619	0.433	0.065	0.056	0.528	0.618	-0.009	0.419	0.493

5.2. Quality of Factorization recovered by TSVDNMF under heavy noise

We perform synthetic experiments to compare the performance of **TSVDNMF** with NMF algorithms which have *provable* error guarantees.

Baselines After comprehensive experimentation (Gillis & Luce, 2014) recommends **Gillis-LP** for their robustness. We also consider **SPA** which has been recently shown to be provably robust, and recently proposed variants of SPA, namely **ER-SPA** and **PW-SPA**. In summary we compare **TSVDNMF** with the following algorithms. (1) **SPA**: the successive projection algorithm of (Araújo & et al., 2001). (2) **ER-SPA**: ellipsoid rounding of (Mizutani, 2014) applied to SPA. (3) **PW-SPA**: SPA preconditioned with the SVD based approach proposed in (Gillis & Ma, 2014). (4) **Gillis-LP**: Solves equation (8) of (Gillis & Luce, 2014).[8] [9]

Datasets: Noisy data matrix $A_{d\times n}$ is set to be $\Pi(BC+N)$ where matrices $B_{d\times k}$ and $C_{k\times n}$ are generated to satisfy either Separability or Dominant assumption and the noise matrix $N_{d\times n}$ is generated under two different noise models, Gaussian noise or Multinomial noise. $\Pi_{d\times d}$ is a permutation matrix generated uniformly at random. We chose $d = 100, n = 100, k = 10$.

Separable data: Following (Gillis & Luce, 2014), each entry of C is generated uniformly at random from the interval [0,1]. The matrix B is taken as $[I \;\; \tilde{B}]^T$ where columns of $\tilde{B}$ are generated from Dirichlet distribution with parameters drawn from a uniform distribution over interval [0,1].

Dominant data: Each column of C is generated from a symmetric Dirichlet distribution with hyper-parameter $\frac{1}{2k}$. Columns of B are also generated from Dirichlet by randomly selecting c features and putting weight $\propto \eta_0$ on these features. Concretely, let c be the number of dominant features for each basis vector and η_0 be the sum of the weights of the dominant features in each basis vector. Assume $\eta = (\frac{\eta_0}{1-\eta_0}) \cdot (\frac{d-c}{c})$. For the jth basis vector, we set $\widetilde{\alpha}_j = \mathbf{1}_{d\times 1}$ and multiply the elements of $\widetilde{\alpha}_j$ indexed from $(c(j-1)+1)$ to cj by η. jth column of B, is then generated from a Dirichlet with hyper-parameter $\widetilde{\alpha}_j$. We chose $c = 3, \eta_0 = 0.1$.

Gaussian Noise: Noise matrix N is generated with each entry from $\mathcal{N}(0,1)$ and jth column $N_{.j}$ is multiplied by $\frac{\tilde{\beta}}{\sqrt{d}} \times \|(BC)_{.,j}\|_2$ (to make the noise comparable to the true matrix BC) where $\tilde{\beta} \in \{0.5, 1, 2\}$ is the noise level.

Multinomial Noise: The matrices B and C are normalized such that each of their column sums to one. $N_{.j}$, jth column of noise matrix N, is generated by picking m samples (a sample is a 1-of-d coding) from $[1 \cdots d]$ with prob$(i) = (BC)_{ij}$ and taking an average of the samples to find $\tilde{N}_{.j}$. Then $N_{.j}$ is set as $\tilde{N}_{.j} - (BC)_{.j}$. Lower m implies high noise and vice versa, $m \in \{10, 60, 100\}$.

Performance measures: Let $A = BC + N$ be the noisy data matrix with factorization BC and $\widehat{B}\widehat{C}$ be the output of an NMF algorithm. We compare the performance of various algorithms by the ℓ_1-*residual*, $1 - \|BC - \widehat{B}\widehat{C}\|_s / \|BC\|_s$ where $\|\cdot\|_s$ is the sum of absolute values of the entries of the input, suggested in (Gillis & Luce, 2014). Higher value of the measure signifies better performance.

Observation: For each setting the performance is averaged over 10 independent datasets and reported in Table 2. In the experiment on separable data with high Gaussian noise, Gills-LP and ER-SPA outperform SPA which is consistent with (Gillis & Luce, 2014) and (Mizutani, 2014). **TSVDNMF** outperforms the baselines by 27.4% (on an average) in all the settings except in separable data with gaussian noise ($\tilde{\beta} = 0.5$) where ER-SPA performs better, but with a very small (0.8%) improvement over **TSVDNMF** . Another interesting observation is in the multinomial noise setting with $m = 10$ (high noise) where all the separable based algorithms gives negative result which is worse than making the matrix $C = 0_{k\times n}$ with zero performance. However in this setting **TSVDNMF** performs better with very small accuracy (0.017). In section 4.1 of supplement material we report results of **Leave one out** experiment to demonstrate the robustness of **TSVDNMF** .

[8] For **ER-methods** and **Gillis-LP**, as suggested by their authors, we set $\eta = 0.99, \theta = 5$ and $\rho = 1$ respectively.

[9] Codes for **SPA, PW-SPA, Gillis-LP** are obtained from their first author's website

Table 3. Clustering : NMI and Accuracy achieved by different NMF algorithms on five datasets.

	Reuters10		**Reuters48**		**20 Newsgroups**		**TDT-2**		**Yale face**	
	NMI	AC	NMI	AC	NMI	AC	NMI	AC	NMI	AC
TSVDNMF	**0.54**	**0.618**	**0.456**	**0.482**	**0.436**	**0.414**	**0.708**	**0.635**	**0.784**	**0.77**
SPA	0.211	0.314	0.297	0.195	0.271	0.258	0.478	0.41	0.523	0.467
ER-SPA	0.316	0.348	0.359	0.241	0.282	0.26	0.534	0.454	0.74	0.667
PW-SPA	0.369	0.382	0.363	0.24	0.191	0.186	0.603	0.515	0.658	0.655

5.3. Clustering performance of TSVDNMF on real world datasets

In this section we compare **TSVDNMF** with the recent provable NMF methods on clustering experiment with real text and face datasets. For clustering, first we find the coefficient matrix C and then for each datapoint, we assign a cluster label based on the maximum element in the corresponding column of C.

Baselines: We consider the recent provable NMF methods SPA, ER-SPA and PW-SPA. [10] We also compare with clustering by Kmeans, XRAY (Kumar et al., 2013) due to their popularity and TSVD as it shares the same framework with **TSVDNMF** , however as our focus is on provable NMF methods, we put these results in section 4.2 of supplement.

Dataset: Following (Mizutani, 2014; Kumar et al., 2013), we use four text datasets Reuter10, Reuters48, 20 Newsgroups and TDT2. Additionally we consider the Yale face dataset to compare **TSVDNMF** on facial images. All these datasets are described in section 5.1.

Performance measures: Following (Mizutani, 2014), we used two metrics, accuracy (AC) and the normalized mutual information (NMI) to evaluate the clustering performance. Let $\mathcal{T} = \{\mathcal{T}_1, \cdots \mathcal{T}_k\}$ be a partition of the dataset (of n data points) according to the true class labels and $\mathcal{S} = \{\mathcal{S}_1, \cdots \mathcal{S}_k\}$ be a partition of the dataset according to the cluster labels returned by an algorithm. To find AC, we align $\mathcal{S}$ and $\mathcal{T}$ to maximize the total number of common data points and compute AC $= \frac{1}{n} \sum_{j=1}^{k} |\mathcal{S}_j \cap \mathcal{T}_j|$. NMI is defined as NMI$(\mathcal{S}, \mathcal{T}) = 2 \cdot I(\mathcal{S}, \mathcal{T}) / [H(\mathcal{S}) + H(\mathcal{T})]$ where I and H denote the mutual information and entropy respectively. See (Manning et al., 2008) for details.

Observation: Table 3 shows the Accuracy and NMI of the clustering by **TSVDNMF** and the baselines. For all the datasets, **TSVDNMF** outperforms the Baselines comprehensively. On average, over the provable algorithms, **TSVDNMF** gives atleast $\mathbf{33\%}$ improvement in NMI on 4 Text datasets and $\mathbf{6.6\%}$ improvement in NMI on the Yale face recognition dataset. This real data experiment empirically justifies robustness of **TSVDNMF** over the other NMF algorithms. From runtime perspective, we observed in Reuters10, **TSVDNMF** took 8 secs as opposed to ER-SPA's 3.2 secs and PW-SPA's 1.2 secs, which is reasonable given the improvement in performance.

[10] We don't consider Gillis-LP (Gillis & Luce, 2014) due to its high run time (as stated by their authors).

5.4. A Summary of the Empirical results

First we empirically justified the dominant basis vector and the pure record assumptions. Then we extensively compared our proposed approach with the state-of-the-art provably robust NMF algorithms on synthetic datasets under different noise models and on real text and face datasets for the task of clustering. We observed that when the data satisfies the separability assumption (with high noise) or the dominant assumption, **TSVDNMF** is more robust than the baselines on both the Gaussian and multinomial noise models. In the clustering experiment **TSVDNMF** outperforms all the other methods by a significant margin.

We also conducted leave-one-out experiments on synthetic data under separability and dominant assumptions with both Gaussian and multinomial noise models. We see **TSVDNMF** to be more robust and an improvement of 12.7%(on average) over the baselines. For details see section 4.1 of supplement. We also include clustering results with more baselines in section 4.2 of supplement where **TSVDNMF** outperforms over the extended baselines on all the datasets except Reuters48 where it is comparable to the best performing TSVD.

6. Conclusion

Existing NMF models are based on separability assumption and are extremely sensitive to noise. This paper proposes an alternate model which does not require separability, is realistic, is resilient to noise in the data, and gives robustly unique NMF. Inferring the correct factorization is also efficiently done by the proposed TSVDNMF algorithm. This is the first work towards proposing a provable NMF model for heavy noise setting along with a substantially improved noise-tolerant and efficient algorithm and a proof of uniqueness of factorization. Among the possible directions for future work: Further improve the running time. SVD is a bottleneck here, but perhaps for suitable models one can do much better. Also, weaker conditions for robust uniqueness would be of interest.

Acknowledgements

The authors graciously acknowledge an Unrestricted gift from Microsoft.

References

Araújo, Mário César Ugulino and et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

Arora, Sanjeev, Ge, Rong, Kannan, Ravindran, and Moitra, Ankur. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 145–162. ACM, 2012.

Bansal, Trapit, Bhattacharyya, Chiranjib, and Kannan, Ravindran. A provable svd-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems 27*, pp. 1997–2005, 2014.

Donoho, David and Stodden, Victoria. When does nonnegative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, pp. 1141–1148, 2003.

Gillis, N. and Vavasis, S.A. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(4):698–714, April 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.226.

Gillis, Nicolas. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12, 2014.

Gillis, Nicolas and Luce, Robert. Robust near-separable nonnegative matrix factorization using linear optimization. *The Journal of Machine Learning Research*, 15(1): 1249–1280, 2014.

Gillis, Nicolas and Ma, Wing-Kin. Enhancing pure-pixel identification performance via preconditioning. *arXiv preprint arXiv:1406.5286*, 2014.

Huang, Kejun, Sidiropoulos, Nicholas D, and Swami, Ananthram. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *Signal Processing, IEEE Transactions on*, 62(1):211–224, 2014.

Kumar, Abhishek, Sindhwani, Vikas, and Kambadur, Prabhanjan. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013.

Laurberg, Hans, Christensen, Mads Græsbøll, Plumbley, Mark D, Hansen, Lars Kai, and Jensen, Søren Holdt. Theorems on positive data: On the uniqueness of nmf. *Computational intelligence and neuroscience*, 2008, 2008.

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Manning, Christopher D, Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

McDiarmid, C. On the method of bounded differences. *Surveys in Combinatorics*, 141 (August):148–188, 1989.

Mizutani, Tomohiko. Ellipsoidal rounding for nonnegative matrix factorization under noisy separability. *The Journal of Machine Learning Research*, 15(1):1011–1039, 2014.

Oh, Hyun Jun, Lee, Kyoung Mu, and Lee, Sang Uk. Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. *Image and Vision Computing*, 26(11):1515–1523, 2008.

Recht, Ben, Re, Christopher, Tropp, Joel, and Bittorf, Victor. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pp. 1214–1222, 2012.

Rong, Ge and Zou, James. Intersecting faces: Non-negative matrix factorization with new guarantees. In *ICML*, pp. 2295–2303, 2015.