Supplement for “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

1. Heavy noise encompasses several Noise models

Recall the Heavy noise model assumed in Assumption **D4**.

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

It subsumes several noise models which include i.i.d Gaussian, Multinomial and Heavy Tailed Distribution. The focus of this section is to make these relations explicit.

To help exposition we define the following event for a random N.

$$\mathcal{E}\ :\ \exists T \subseteq [n], |T| \geq \varepsilon_4 n \text{ with } \left|\left|\sum_{j\in T} N_{\cdot,j}\right|\right|_1 \geq \varepsilon_4^2 |T|. \quad (2)$$

Our strategy will be to argue that this event occurs with extremely low probability for various noise models.

Gaussian Noise-IID Case Consider the following lemma when the entries of N are i.i.d. $\mathcal{N}(0,\sigma^2)$ random variables.

Lemma 1. *Suppose N is a $d \times n$ matrix with $n \geq d$ with i.i.d. entries drawn from $\mathcal{N}(0,\sigma^2)$, where $\sigma \leq \frac{\varepsilon_4^{5/2}}{4\sqrt{n}}$. For n large enough, $Prob(\mathcal{E}) \leq 0.01$.*

Proof. By Random Matrix Theory (see for example Theorem 5.31 in (Vershynin, 2010)) we have that the largest singular value of N is at most $4\sigma\sqrt{n}$ with probability 0.99 For contradiction, assume that there is some T, $|T| \geq \varepsilon_4 n$ with $\left|\left|\sum_{j\in T} N_{\cdot,j}\right|\right|_1 \geq \varepsilon_4^2|T|$. This implies that $\left|\left|\sum_{j\in T} N_{\cdot,j}\right|\right|_2 \geq \varepsilon_4^2|T|/\sqrt{d}$. But $\left|\left|\sum_{j\in T} N_{\cdot,j}\right|\right|_2 \leq ||N||_2\sqrt{|T|} \leq 4\sigma\sqrt{n}\sqrt{|T|}$ producing the contradiction. □

If $\sigma \leq \varepsilon_4^{5/2}/4\sqrt{d}$, (1) is satisfied. Furthermore, if $\sigma > c\varepsilon/d$, then, with high probability, a CONSTANT FRACTION of the columns j violate the condition $||N_{\cdot,j}||_1 \leq \varepsilon$ required by previous algorithms to hold for EVERY column.

Lemma 2. *Suppose $k = 1$, $n \leq c_0 d$, $||C_{\cdot,j}||_1 = 1$ for all j and N has i.i.d. entries drawn from $\mathcal{N}(0,\sigma^2)$, where, $\sigma > c_1/\sqrt{d}$ for a large constant c_1. Then, given $A = BC + N$, the Maximum Likelihood Estimator $\tilde{B}$ of B with high probability satisfies*

$$\left|\left|\tilde{B}_{\cdot,1} - B_{\cdot,1}\right|\right|_1 > \varepsilon.$$

Proof. Since $k = 1$, it is easy to see that the MLE is just the average of all columns of A. Since $\text{Var}(\sum_{j=1}^n N_{ij}) = n\sigma^2$, with high probability, for a constant fraction of i, we have $\left|\sum_{j=1}^n N_{ij}\right| > \sqrt{n}\sigma/10$ which implies that $\frac{1}{n}\left|\left|\sum_{j=1}^n N_{\cdot,j}\right|\right|_1 \geq \sigma\sqrt{n}d/10$, from which the Lemma follows. □

General Correlated Noise: While noise in different data points may be independent, noise in different coordinates of the same data point need not be. [For example, in Hyperspectral Imaging, where NMF is applied, each column of A corresponds to a pixel, noise in its intensity at different frequencies (these are the coordinates) are not necessarily independent.] We can model this more general case by having $N_{\cdot,j}$ be independent vector-valued zero-mean random variables. Suppose Σ_j is the covarience matrix of $N_{\cdot,j}$ for $j = 1, 2, \ldots, n$. [The Σ_j need not be the same.]

$$\Sigma_j = E\left(N_{\cdot,j}N_{\cdot,j}^T\right).$$

Lemma 3. *Suppose $n \geq d$. If $||\Sigma_j||_2 \leq \varepsilon_4^5/8n$, then, for n large enough,*

$$Pr\left(\mathcal{E}\right) \leq 0.01.$$

Note that for the i.i.d. spherical Gaussian case, $||\Sigma_j||_2 = \sigma^2$, so Lemma (1) is essentially a special case.

Proof. Recent results from Random Matrix Theory apply to matrices with independent vector-valued random variables as columns (so no independence among rows is assumed) (for example, Theorem 5.48 and Remark 5.49 of ((Vershynin, 2010))), and assert:

$$||N||_2 \le 4||\Sigma||_2 n$$

with high probability. Arguing as in Lemma 1 yields the current Lemma. □

Multinomial Noise For this, assume $||C_{\cdot,j}||_1 = ||A_{\cdot,j}||_1 = 1$. We 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 the example of Topic Modeling, the process above has picked m words in each document. Usually, $m << n, d$] Almost all data points can violate the condition $||N_{\cdot,j}||_1 \le \varepsilon$. However, we will also show using the bounded difference martingale inequality (McDiarmid, 1989) that our noise assumption is satisfied.

Lemma 4. *Suppose $N_{\cdot,j}$ is the average of m i.i.d. Multinomial trials, each of which picks a unit vector e_i with probabilities proportional to $(BC)_{i,j}$. Subtract the mean to make $E(N_{ij}) = 0$. For n large enough, the $Prob(\mathcal{E}) \le 0.01$*

Proof. The proof follows from Lemma 5. □

Heavy noise We now discuss a more general case of Heavy noise where each column of N can have large norm.

Lemma 5. *Suppose $||C_{\cdot,j}||_1 = ||A_{\cdot,j}||_1 = 1$ and suppose each column $N_{\cdot,j}$ of N is the average of $m \ge 8c_0^2 \log(e/\varepsilon_4)/\varepsilon_4^4$ independent zero-mean vector-valued random variables $N_{\cdot,j}^{(1)}, N_{\cdot,j}^{(2)}, \dots, N_{\cdot,j}^{(m)}$- with $||N_{\cdot,j}^{(t)}||_1 \le c_0$.* [1] *Then, (1) is satisfied whp.*

Proof. Fix attention on one $T \subseteq [n], |T| \ge \varepsilon_4 n$ for now. Let

$$X = \frac{1}{m|T|}\left|\left|\sum_{j\in T, t\in[m]} N_{\cdot,j}^{(t)}\right|\right|_1 .$$

The random variable X is a function of $m|T|$ independent random variables $\{N_{\cdot,j}^{(t)} : j \in T, t = 1, 2, \dots, m\}$. Further changing any one of these random variables causes a maximum change of $c_0/m|T|$ in X since we assumed $||N_{\cdot,j}^{(t)}||_1 \le c_0$ with probability 1. So applying the bounded difference inequality for Martingales (McDiarmid, 1989) directly, we get that for all $\lambda > 0$,

$$\Pr\left(|X - EX| \ge \lambda\right) \le 2\exp\left(-\lambda^2 m|T|/8c_0^2\right).$$

Put $\lambda = 8c_0\sqrt{\ln(e/\varepsilon_4)}/\sqrt{m}$ in the above.

So far we have bounded the probability that a single T violates (1). We wish to take the union bound over all such T. For this, note that the number of T of cardinality t is

$$\binom{n}{t} \le e^{t\ln(e/\varepsilon_4)},$$

noting that $t \ge \varepsilon_4 n$. Plugging this in, we see that the probability that (1) does not hold for a particular t is at most $1/n^2$. Again taking the union bound over all t, we get the Lemma. □

On Assumption D3 To ensure identifiability in presence of noise we need few *Pure Documents* to be present in the corpus, assumption **D3**. Consider the case when $k = 2$ and say there are less than $\varepsilon_4 n/2$ nearly pure records for $l = 1$. These can all be corrupted when noise is adversarial. Then, it is easy to see that $\tilde{B}$ with $\tilde{B}_{\cdot,1}$ = a convex combination of $B_{\cdot,1}$ and $B_{\cdot,2}$ which is not near $B_{\cdot,1}$ can satisfy $\|B_{\cdot 1} - \tilde{B}_{\cdot 1}\|_1 \le \epsilon$. This shows the necessity of **D3** for Identifiability under heavy noise.

2. TSVDNMF : A SVD based Algorithm for NMF

In this section we prove the correctness of **TSVDNMF** .

2.1. Notation

Let [2] $\alpha, \beta, \rho, \varepsilon, \varepsilon_0, \varepsilon_4$ be non-negative reals in $(0, 1)$ satisfying:

$$\varepsilon < \varepsilon_0/20. \tag{3}$$

$$\beta + \rho \le \text{Min}\left((1 - 5\varepsilon)\alpha, 0.6 - 4\alpha\varepsilon\right). \tag{4}$$

$$\varepsilon_4 \le \alpha\gamma\varepsilon/2\ ,\ \frac{\varepsilon_0 w_0 p_0}{16k^3}\ ,\ \varepsilon_0^2/4.. \tag{5}$$

Algorithm **Input:** Un-normalized matrix **A**; **Output:** Basis matrix **B**.

1. **Thresholding**: Apply the Thresholding procedure (see below) to get **D** and k.
2. **SVD**: Find the best rank k approximation $\mathbf{D}^{(k)}$ to **D**.

[1] Note that this allows the noise to be as large (in l_1 norm) as data.

[2] The algorithm uses the actual values of γ, ε_4 only for a technical reason in the proof. The terms $\gamma - 2\varepsilon_4$ and $-2\varepsilon_4$ can be dropped for small size problems.

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 $B_{\cdot,l}$.

2.2. 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 $j \in W_{i'} \setminus W_i$; set $W_{i'}$ to W_i and delete i' from R.

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

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

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

Before we state the main proof we establish a few lemmas which will help in deriving the proof.

2.3. Dominant Features and Primary Basis Vectors

Wlg Assume $|B_{\cdot,j}|_1 = 1$ and $|A_{\cdot,j}|_1 \leq 1$ for all j. These imply $|C_{\cdot,j}|_1 \leq 1$.

Dominant Basis Vector Assumption There are $T_1, T_2, \ldots, T_k \subseteq [n]$ satisfying

$$\forall j \in T_l, l' \neq l, \qquad C_{l,j} \geq \alpha \text{ and } C_{l'j} \leq \beta \tag{6}$$
$$\forall l, \qquad |T_l| = w_l n. \tag{7}$$

We assume $w_l \geq w_0$ for all l.

Nearly Pure Records Assumption For each l, there are at least $\varepsilon_0 n$ records in each of which the l th basis vector has coefficient at least $1 - \varepsilon_4$. I.e.,

$$\forall l, \exists \geq \varepsilon_0 n \; j \text{ with } C_{lj} \geq 1 - \varepsilon_4. \tag{8}$$

Dominant Features Assumption: There are k disjoint sets of features - $S_1, S_2, \ldots, S_k$ such that

$$\forall i \in S_l, \; \forall l' \neq l, \; B_{il'} \leq \rho B_{il} \tag{9}$$
$$\sum_{i \in S_l} B_{il} \geq p_0 \tag{10}$$
$$\forall i \in S_l, B_{il} \geq \gamma. \tag{11}$$

Noise:

$$\forall T \subseteq [n] \text{ with } |T| \geq \varepsilon_4 n \; , \; \frac{1}{|T|} \left| \sum_{j \in T} (A - BC)_{\cdot,j} \right|_1 \leq \varepsilon_4^2. \tag{12}$$

Lemma 6. *For $i \in [d]$, let $l(i) = \arg\max_l B_{il}$. Consider the W_i at the end of step 6. If $W_i \neq \emptyset$, then,*

$$|T_{l(i)} \setminus W_i| \leq 2\varepsilon_4 n.$$

In addition, if $i \in S_l$, then

$$|W_i \setminus T_l| \leq 2\varepsilon_4 n.$$

Proof. Fix attention on one i with $|W_i| \neq 0$ and let $l = l(i)$. W_i initially has at least $\varepsilon_0 n/2$ elements. If it ever gets pruned by another W, it is easy to see that it will not become empty and remains non-empty.

For any j, $(BC)_{ij} = \sum_{l'} B_{il'} C_{l'j} \leq B_{il}$, since, $B_{il'} \leq B_{il}$ and $\sum_{l'} C_{l'j} \leq 1$. Let

$$H_i = \{j : |A_{ij} - (BC)_{ij}| \geq \varepsilon_4\}. \tag{13}$$

We claim that $|H_i| \leq 2\varepsilon_4 n$. To see this, let $H_i^+ = \{j : A_{ij} \geq (BC)_{ij} + \varepsilon_4\}$. $\sum_{j \in H_i^+} (A - BC)_{ij} \geq \varepsilon_4 |H_i^+|$.

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

So, (12) implies $|H_i^+| \le \varepsilon_4 n$. Similarly for $H_i^- = \{j : (A-BC)_{ij} \le -\varepsilon_4\}$.

For $j \notin H_i$, $A_{ij} \le B_{il} + \varepsilon_4$ and so we have with ν_i as defined in the thresholding procedure, $\nu_i \le B_{il} + \varepsilon_4$ (using (3, 5) from which it follows that

$$\zeta_i \le \alpha B_{il} - (2-\alpha)\varepsilon_4. \tag{14}$$

If $j \in T_l$, then $C_{lj} \ge \alpha$ implies $(BC)_{ij} \ge B_{il}C_{lj} \ge \alpha B_{il}$. So for $j \in T_l \setminus H_i$, $A_{ij} \ge \alpha B_{il} - \varepsilon_4 \ge \zeta_i$ (by (14)) proving the first assertion of the theorem.

To prove the second assertion, now suppose $i \in S_l$. By Assumption (8), there are $\varepsilon_0 n$ nearly pure records and since $\varepsilon_4 \le \varepsilon_0/16$ (3, 5), we have at least $\varepsilon_0 n/2$ of these nearly pure records must not be in H_i and so have $|A_{ij} - (BC)_{ij}| \le \varepsilon_4$. So, we have $\nu_i \ge B_{il}(1-\varepsilon_4) - \varepsilon_4 \ge B_{il} - 2\varepsilon_4$ (since for pure records, $(BC)_{ij} \ge (1-\varepsilon_4)B_{il}$) which implies (using $\varepsilon_4 \le \alpha\gamma/8$ from (5))

$$\zeta_i \ge \alpha B_{il} - 4\varepsilon_4 > 0, \tag{15}$$

since by (11), $B_{il} \ge \gamma > 4\varepsilon_4/\alpha$.

For $j \notin T_l$,

$$\begin{aligned}(BC)_{i,j} &= B_{il}C_{lj} + \sum_{l' \ne l} B_{il'}C_{l'j} \\ &\le B_{il}C_{lj} + \rho B_{il}(1 - C_{lj}) \\ &\le B_{il}(\beta + \rho) \\ &\le B_{il}\alpha(1-5\varepsilon),\end{aligned}$$

since $C_{lj} \le \beta$ by (6) and by (3), $\beta + \rho \le (1-5\varepsilon)\alpha$.

Now consider $j \notin (T_l \cup H_i)$.

$$A_{ij} \le B_{il}\alpha(1-5\varepsilon) + \varepsilon_4 < \alpha B_{il} - 4\varepsilon_4 \le \zeta_i,$$

where, for the last inequality, we have used (15) and for the previous step, (3) and (11). This implies the second assertion of the Lemma.

□

Lemma 7 (No Threshold Splits any T_l). *At the end of the thresholding step of the Algorithm, for every i, (i need not be in any S_l), and every l, we have that either there are at most $4\varepsilon_4 n$ $j \in T_l$ with $D_{ij} > 0$ or at most $4\varepsilon_4 n$ $j \in T_l$ with $D_{ij} = 0$.*

Proof. From the previous Lemma, for $i \in S_l$, we see that

$$|T_l \setminus W_i| \ , \ |W_i \setminus T_l| \le 2\varepsilon_4 n.$$

We claim that for $i \in S_l$, there is no i' such that $|W_{i'}| \le |W_i| - \varepsilon_0 n/8$ and $W_{i'} \tilde{\subseteq} W_i$ (and hence, the pruning step 5 does not modify the assignment $D_{ij} = \sqrt{\zeta_i}$ made in step 4 for all $j \in W_i$). Clearly this suffices to prove the Lemma for $i \in S_l$. Now for the claim: suppose for some i', $W_{i'} \tilde{\subseteq} W_i$. By Lemma (6), $W_{i'}$ must contain almost all of some $T_{l'}$. If $l' = l$, then, $|W_{i'}| \ge |T_l| - 2\varepsilon_4 n \ge |W_i| - 3\varepsilon_4 n$ contradicting the hypothesis of the Claim (by (3, 5). If on the other hand, $l' \ne l$, then, we must have that W_i also contains all but $2\varepsilon_4 n$ elements of $T_{l'}$ contradicting $|W_i \setminus T_l| \le 3\varepsilon_4 n$. This proves the claim and the Lemma for $i \in S_l$.

Now for $i \notin \cup_l S_l$. If $|W_i| \le |T_{l(i)}| + 2\varepsilon_4 n$, then, since $W_{l(i)}$ already contains all but $2\varepsilon_4 n$ of $T_{l(i)}$. it must contain at most $4\varepsilon_4 n$ elements of any $T_l, l \ne l(i)$, and so cannot split any T_l. So assume $|W_i| > |T_{l(i)}| + 2\varepsilon_4 n$. Take $i' \in S_{l(i)}$. $|W_{i'} \setminus T_{l(i)}| \le 2\varepsilon_4 n$ implies that $|W_{i'}| \le |T_{l(i)}| + 2\varepsilon_4 n$. So $W_{i'}$ comes before W_i Also, $|W_i| \ge |W_{i'}| + (\varepsilon_0 n/8)$. So i' prunes W_i if it is not pruned already. We have proved that W_i is pruned from some $i' \in \cup_l S_l$ and it is easy to see that after pruning, it cannot split any T_l.

□

Let μ be a $d \times n$ matrix whose columns are given by

$$\forall j \in T_l \ , \ \mu_{.,j} = \frac{1}{|T_l|}\sum_{j \in T_l} D_{.,j}.$$

μ 's columns corresponding to all $j \in T_l$ are the same.

Lemma 8.

$$\forall l, \forall j \in T_l, \forall i, \quad \mu_{ij} \le \frac{4\varepsilon_4}{w_l}\sqrt{\zeta_i} \textit{ OR } \mu_{ij} \ge \sqrt{\zeta_i}(1 - \frac{4\varepsilon_4}{w_l})$$

$$\frac{1}{|T_l|}\sum_{j \in T_l}(D_{ij} - \mu_{ij})^2 \le \frac{8\varepsilon_4}{w_l}\zeta_i, \textit{ where, } |T_l| = w_l n.$$

Thus,

$$||D - \mu||_F^2 \le 16\alpha k^2 \varepsilon_4 n.$$

Proof. Fix attention on one i. Let

$$a_l = \left|\{j \in T_l : D_{ij} = \sqrt{\zeta_i}\}\right|.$$

From Lemma (7), we know that a_l is either at least $(w_l - 4\varepsilon_4)n$ or at most $4\varepsilon_4 n$ do. Consider the first case. Then, clearly,

$$\mu \ge \frac{(w_l - 4\varepsilon_4)n}{w_l n}\sqrt{\zeta_i} = \left(1 - \frac{4\varepsilon_4}{w_l}\right)\sqrt{\zeta_i}.$$

Also, in this case,

$$\frac{1}{|T_l|}\sum_{j \in T_l}(D_{ij} - \mu_{ij})^2 \le \frac{\zeta_i}{w_l n}\left(a_l \frac{16\varepsilon_4^2}{w_l^2} + (w_l n - a_l)\right).$$

Now, from (3, 5) and the fact that $w_0 \ge \varepsilon_0$, we know that $4\varepsilon_4 < w_0 \le w_l$, so in the above expression, the maximum value is when $a_l = (w_l - 4\varepsilon_4)n$. From this, the first two assertions of the Lemma follow for this case. The other case is symmetrically handled.

To bound $||D-\mu||_F^2$, it is first clear that $||D-\mu||_F^2 \le \sum_l \frac{\varepsilon_3}{w_l} w_l n \sum_i \zeta_i$. it remains to bound $\sum_i \zeta_i$. Let $I = \{i : \zeta_i > 0\}$. For $i \in I$, $\alpha\nu_i \ge 2\varepsilon_4$ (by the way ζ_i is determined in the algorithm.) Now,

$$\nu_i \le B_{l(i),i} + \varepsilon_4 \le 2B_{l(i),i},$$

for $i \in I$, since, $|H_i| \le 2\varepsilon_4 n < \varepsilon_0 n/3$. Thus, $\zeta_i \le \nu_i \le 2B_{l(i),i}$. This implies that

$$\sum_i \zeta_i \le 2\sum_i B_{l(i),i} \le 2\sum_{i,l} B_{il} \le 2k, \qquad (16)$$

since, B has column sums equal to 1. The last assertion of the Lemma follows. □

Lemma 9. *For $j \in T_l$ and $j' \in T_{l'}$ with $l' \ne l$,*

$$|\mu_{\cdot,j} - \mu_{\cdot,j'}|^2 \ge c\alpha p_0.$$

Proof. From Lemma (6), it follows that for $i \in S_l$, at most $\varepsilon_4 n$ $j \in T_l$ have $D_{ij} = 0$ and at most $\varepsilon_4 n$ $j' \notin T_l$ have $D_{ij} = \sqrt{\zeta_i}$. This implies that $|\mu_{i,j} - \mu_{i,j'}| \ge \frac{1}{4}\sqrt{\zeta_i}$. Squaring and adding over all $i \in S_l$, we get

$$|\mu_{i,j} - \mu_{i,j'}|^2 \ge \frac{1}{16}\sum_{i \in S_l} \zeta_i.$$

Now we show that $\sum_{i \in S_l} \zeta_i$ is high enough to prove the Lemma. We showed in the proof of Lemma (6) that H_i as defined there satisfies $|H_i| \le \varepsilon_4 n$. For the $\varepsilon_0 n$ pure records of l, we have $(BC)_{ij} \ge (1-\varepsilon_4)B_{l,i}$. So at least $\varepsilon_0 n/2$ of these do not belong to H_i and have $A_{ij} \ge (BC)_{ij} - \varepsilon_4 \ge B_{il}(1-\varepsilon_4) - \varepsilon_4 \ge B_{il}/2$, (since for $i \in S_l, B_{il} \ge \gamma \ge 4\varepsilon_4$). This implies that $\zeta_i \ge B_{il}\alpha/6$. So, we have

$$\sum_{i \in S_l} \zeta_i \ge \alpha \sum_{i \in S_l} B_{li}/6 \ge \alpha p_o/6, \qquad (17)$$

by (10). □

Lemma 10.

$$\sigma_{k+1}^2(D) \le 2\varepsilon_4 nk^2$$
$$\sigma_k(D)^2 \ge w_0 p_0 n/4.$$

Proof. Let U be the span of the incidence vectors of $T_1, T_2, \ldots, T_k$. Let $\hat{D}_{i,\cdot}$ be the component of $D_{i,\cdot}$ orthogonal to U. Lemma (7) implies that $||\hat{D}_{i,\cdot}||_2^2 \le \varepsilon_4 nk\zeta_i$. So, by (16),

$$||\hat{D}||_F^2 \le \varepsilon_4 nk \sum_{i=1}^{d} \zeta_i \le 2\varepsilon_4 nk^2,$$

which implies by the Min-Max Theorem the first part of the Lemma.

For the other part, we have again by min-max:

$$\begin{aligned}
\sigma_k(D)^2 &\ge \text{Min}_{l=1}^{k} \frac{1}{|T_l|} \sum_{i=1}^{d} \left(\sum_{j \in T_l} D_{ij} \right)^2 \\
&\ge \text{Min}_l \frac{1}{|T_l|} \sum_{i \in S_l} \left(\sum_{j \in T_l} D_{ij} \right)^2 \\
&\ge \text{Min}_l \frac{|T_l|}{4} \sum_{i \in S_l} \zeta_i \ge c\alpha |T_l| p_0,
\end{aligned}$$

by (17). So the second part of the Lemma follows. □

Using a results from K-means clustering due to (Kumar & Kannan, 2010) which states that our algorithm correctly classifies most records by dominant basis vectors, the number of records misclassified by the algorithm being at most

$$\frac{k||D-\mu||_2^2}{\min_{l \ne l'} |\mu_l - \mu_{l'}|^2 w_0},$$

which by Lemmas (9) and (8) is at most $O(k^3\varepsilon_4 n/(w_0 p_0))$ which is at most $\varepsilon_0 n/4$ from (5) .

Lemma 11. *At the end of the $k-$means clustering step, the algorithm correctly identifies the dominant basis vector in all but at most $\varepsilon_0 n/4$ records.*

2.4. Identifying Dominant Features

We now show that the step of identifying dominant features for each basis vector works correctly. This will be proved in two lemmas which are roughly converses of each other, asserting that $J_l \approx S_l$ in essence.

Lemma 12. *For all l, $S_l \subseteq J_l$.*

Proof. Suppose $i \in S_l$. By Lemma (11), we know $|R_l \triangle T_l| \le \varepsilon n$. There are $\varepsilon_0 n$ pure records for l, at least $3\varepsilon_0 n/4$ are in R_l and at least $\varepsilon_0 n/2$ of them are not in H_i. Thus, $g(i,l) \ge (1-\varepsilon_4)B_{il} - \varepsilon_4 \ge B_{il}(1-\alpha\varepsilon)$ (since for $i \in S_l$, $B_{il} \ge \gamma$). For $j \notin T_l \cup H_i$, we have $A_{ij} \le B_{il}(\beta+\rho)+\varepsilon_4 \le B_{il}(\beta+\rho+\alpha\varepsilon)$. For $l' \ne l$, $R_{l'}$ has at most $\varepsilon_0 n/4 + |H_i|$ records j with $A_{ij} \ge B_{il}(\beta+\rho+2\alpha\varepsilon)$ and this implies that $g(i,l') \le B_{il}(\beta+\rho+2\alpha\varepsilon)$. So, $g(i,l)/g(i,l') \ge \nu$. Also, $g(i,l) \ge \gamma - 2\varepsilon_4$. So $i \in J_l$.

□

2.5. Finding Basis Vectors

Lemma 13. *For $i \in J_l$, and $l' \ne l$, we have:*

$$B_{i,l} \ge \mathit{Max}\left(\frac{\gamma}{2}\,,\ \frac{1-2\varepsilon\alpha}{\beta+\rho+5\alpha\varepsilon} B_{i,l'} \right).$$

Proof. Let $i \in J_l$. Let $l(i) = \arg\max_{l'} B_{il'}$. $g(i,l) \leq B_{i,l(i)} + \varepsilon_4$, since $|H_i| \leq \varepsilon_4 n$. Also, $g(i,l) \geq \gamma - 2\varepsilon_4$ by the definition of J_l. So, we have $B_{i,l(i)} \geq \gamma - 3\varepsilon_4$ which implies that

$$\varepsilon_4 \leq B_{il(i)}\alpha\varepsilon. \tag{18}$$

Now, for any l', we have $g(i,l') \geq (1-\varepsilon_4)B_{il'} - \varepsilon_4$ by the pure records for l' which are not in H_i. If $l' \neq l$, we also have $g(i,l) \geq (1-\alpha\varepsilon)g(i,l')/(\beta+\rho+2\alpha\varepsilon)$, so we get

$$\begin{aligned}\frac{B_{i,l(i)} + \varepsilon_4}{(1-\varepsilon_4)B_{i.l'} - \varepsilon_4} &\geq \frac{1-\alpha\varepsilon}{\beta+\rho+2\alpha\varepsilon}\\ B_{i,l(i)}(\beta+\rho+2\alpha\varepsilon) + 3\varepsilon_4 &\geq (1-\alpha\varepsilon)(1-\varepsilon_4)B_{il'},\end{aligned} \tag{19}$$

from which the Lemma follows if $l(i) = l$. But it is easy to see that $l(i) = l$ and this proves the Lemma.

□

Lemma 14. *The number of j among the $\varepsilon_0 n/4$ highest $\sum_{i\in J_l} A_{ij}$ in the Find Basis Vectors Step of the algorithm for which $C_{lj} \leq 1 - 5\alpha\varepsilon$ is at most $\varepsilon_4 n$.*

Proof. Let $i \in J_l$. Let $Q_l = \{j : C_{lj} \geq 1 - \varepsilon_4\}$. $|Q_l| \geq \varepsilon_0 n$ by Nearly Pure Records Assumption.

$$\begin{aligned}&\forall j \in Q_l, (BC)_{ij} \geq B_{il}(1-\varepsilon_4) \implies\\ &\sum_{i\in J_l}(BC)_{ij} \geq (1-\varepsilon_4)\sum_{i\in J_l} B_{il}.\end{aligned} \tag{20}$$

Let

$$H = \{j : \sum_{i\in J_l}(BC - A)_{ij} \geq \varepsilon_4\}.$$

$$\left|\sum_{j\in H}(A - BC)_{\cdot,j}\right|_1 \geq \sum_{j\in H}\sum_{i\in J_l}(BC-A)_{ij} \geq \varepsilon_4|H|.$$

So, (12) implies $|H| \leq \varepsilon_4 n$. For $j \in Q_l \setminus H$,

$$\sum_{i\in J_l} A_{ij} \geq (1-\varepsilon_4)\sum_{i\in J_l} B_{il} - \varepsilon_4 \geq (1-\alpha\varepsilon)\sum_{i\in J_l} B_{il},$$

using $B_{il} \geq \gamma/2$ from Lemma (13).

This implies that the $1 - \varepsilon_0/2$ fractile of $\{\sum_{i\in J_l} A_{ij}, j = 1,2,\ldots,n\}$ is at least $\sum_{i\in J_l} B_{il}(1-\alpha\varepsilon)$.

On the other hand, define $\hat{H} = \{j : \sum_{i\in J_l}(A - BC)_{ij} \geq \varepsilon_4\}$. Then, $|\hat{H}| \leq \varepsilon_4$ by (12). if for $j \notin H_i$, we have $C_{lj} \leq 1 - 5\alpha\varepsilon$, then,

$$\begin{aligned}(BC)_{ij} &= B_{il}C_{lj} + \sum_{l'\neq l} B_{il'}C_{l'j}\\ &\leq B_{il}(1-5\alpha\varepsilon) + B_{il}\frac{\beta+\rho+\alpha\varepsilon}{1-3\alpha\varepsilon}(5\alpha\varepsilon)\\ &\leq B_{il}(1-\alpha\varepsilon)\end{aligned}$$

$$\sum_{i\in J_l} A_{ij} \leq \sum_{i\in J_l} B_{il}(1-\alpha\varepsilon/2),$$

for all $j \notin \hat{H}$.

□

We now present the proof of Theorem 1.

Proof. (**of Theorem 1**) From the last Lemma, we see that the approximation $\hat{B}_{\cdot,l}$ the algorithm finds to $B_{\cdot,l}$ is the average of a set $T \subseteq [n]$ with $|T| = \varepsilon_0 n/2$ of $A_{\cdot,j}$'s , at most $\varepsilon_4 n$ of which have $C_{lj} \leq 1 - 5\alpha\varepsilon$. We have

$$\begin{aligned}\text{For } j \text{ with } \quad & C_{lj} \geq 1 - 5\alpha\varepsilon,\\ B_{\cdot,l} + u \geq\ & (BC)_{\cdot,j} \geq (1-5\alpha\varepsilon)B_{\cdot,l}\\ \text{For all } j, \quad & |(BC)_{\cdot,j} - B_{\cdot,l}|_1 \leq 1,\end{aligned}$$

for some u with $|u|_1 \leq 5\alpha\varepsilon$. Adding over $j \in T$, we get

$$\frac{1}{|T|}\left|\left|\sum_{j\in T}((BC)_{\cdot,j} - B_{\cdot,l})\right|\right|_1 \leq 5\alpha\varepsilon + \frac{\varepsilon_4 n}{|T|} \leq \varepsilon_0/4,$$

using (5). By (12), we also have

$$\frac{1}{|T|}\left|\left|\sum_{j\in T}(A - (BC))_{\cdot,j}\right|\right|_1 \leq \frac{\varepsilon_0}{2}.$$

Adding the last two inequalities, we get the theorem:

$$\frac{1}{|T|}\left|\left|\sum_{j\in T}(A_{\cdot,j} - B_{\cdot,l})\right|\right|_1 \leq \varepsilon_0.$$

□

3. Uniqueness of nonnegative matrix factorization

Given a non-negative factorization $A = BC$ (with $||B_{\cdot,l}||_1 = 1$ for all l, without loss of generality) we can generate many others of the form $A = B'C'$ where $B' = BP$ and $C' = P^{-1}C$ and P is a $k \times k$ matrix obtained by applying a permutation matrix to a diagonal matrix with nonnegative entries. This just amounts to scaling and permuting of the columns of B and correspondingly of the rows of C. If these account for all factorizations of A then we say that A has unique NMF. For some applications of NMF, it is desirable to have the property that the NMF is unique: It gives us confidence that NMF has found the "right" structure in the data and not some spurious explanation. E.g., in clustering applications it tells us that the clusters are unique, in topic modeling it tells us that the "right" topics have been found. In general, NMF need not be unique and this raises the question of which matrices A have unique NMF.

In their influential paper Donoho and Stodden (Donoho & Stodden, 2003) consider the question of uniqueness of nonnegative matrix factorization. They give conditions on the input matrix for uniqueness to hold. There has been considerable work on understanding uniqueness conditions since then; we refer to Huang et al. (Huang et al., 2014) and references therein for an up-to-date review of the literature. Some of these conditions are necessary and sufficient however they do not seem to be easy to check and use; they are often geometric in nature and not directly related to the application at hand. Donoho et al. gave a necessary condition they called Separable Factorial Articulation Family and applied this to an image library of special type of images, and showed that NMF gives a unique decomposition of the images into parts. One of the conditions used for defining Separable Factorial Articulation Family is the separability of the factorization. This condition has been influential in later work as it leads to efficient algorithms. Laurberg et al. (Laurberg et al., 2008) gave further such conditions and also studied the case when there is noise in the data; see also Huang et al(Huang et al., 2014). However, these conditions do not seem to have wide applicability.

Uniqueness can also be achieved by modifying the optimization formulation of the NMF problem in some way; e.g., the factors can be required to be sparse, a regularizer term can be added to the objective function, or a determinant associated with the factorization can be required to have minimum volume. Considerable work exists on this approach; we refer to Gillis (Gillis, 2014) and Huang et al. (Huang et al., 2014) for more details and references.

In the present paper we give new conditions under which NMF is unique without any change to the problem. Our conditions are arguably natural and easy to interpret and verify (if a factorization is given). Our condition is robust in the sense that it can be adapted to the case when we allow approximate factorizations. This is important because in practice we can only expect to find approximate NMFs.

3.1. A uniqueness theorem for exact NMF

Two conditions on NMF, *separability* and *pure records* have been studied in the literature (we already mentioned the former). We will recall these. We will prove that these conditions together are sufficient for uniquness of NMF.

An NMF $A = BC$ is said to be *separable* if for each $\ell \in [k]$, there is an $i \in [d]$ such that $B_{i,\ell}$ is the unique non-zero entry in row i of B. An NMF $A = BC$ is said to have the *pure records* property if for each $\ell \in [k]$, there is a j such that $C_{\ell,j} = 1$ (and so $C_{\ell',j} = 0$ for all $\ell' \neq \ell$). In the context of topic modeling, this is the same as assuming that for each topic, there is a document purely on that topic.

Theorem 2. *If an NMF $A = BC$, where $\mathrm{rank}(A) = k$, has both the separability and the pure records properties, then, the NMF is unique. I.e., if for $d \times k$ and $k \times n$ (resp.) matrices B', C' with non-negative entries, we have $A = B'C'$, then there is a diagonal matrix D with positive diagonal entries and a permutation matrix Π and $P = D\Pi$, such that $B' = BP$ and $C' = P^{-1}C$.*

Notice that in the second factorization $A = B'C'$, we did not make any assumptions apart from non-negativity.

Proof. We will first bring A in a convenient form by permuting its columns and rows. It's clear that $\mathsf{CH}(A) \subseteq \mathsf{CH}(B)$. The pure records property of factorization $A = BC$ gives that in fact $\mathsf{CH}(A) = \mathsf{CH}(B)$ and the columns of B occur as columns of A. Permute the columns of A, if needed, so that the first k columns correspond to the columns of B. Now, by the separability property of the factorization, and permuting the rows of A if needed, we can arrange that the top left $k \times k$ submatrix D of A is a diagonal matrix with positive entries.

Now consider any other NMF $A = B'C'$. As before, we have $\mathsf{CH}(A) \subseteq \mathsf{CH}(B')$. $A|_{[k]}$ is the matrix consisting of the first k rows of A, and similarly for $B'|_{[k]}$. It is now clear that the only way $\mathsf{CH}(D) \subseteq \mathsf{CH}(B'|_{[k]})$ can hold is that $B'|_{[k]}$ is itself a diagonal matrix with positive entries, possibly after applying a permutation—proving the separability of NMF $B'C'$. We now permute the columns of B' so that $B'|_{[k]}$ is a diagonal matrix. Now, it is clear that the first k columns of C' must also form a diagonal matrix with positive entries—proving the pure records property of NMF $B'C'$. □

3.2. A uniqueness theorem for approximate NMF

Suppose that $A \in \mathbb{R}_+^{d\times n}$ has factorization of the form

$$A = BC,$$

where $B \in \mathbb{R}_+^{d\times k}$ and $C \in \mathbb{R}_+^{k\times n}$ and $k = \mathrm{rank}(A)$. In addition, B and C satisfy the following properties:

- **Normalization Assumption.** Each column of B sums to 1.
- **Dominant Features Assumption.** There are constants $p_1 < p_0 \in (0,1)$ and pairwise disjoint sets $S_1, \ldots, S_k \subset [d]$ such that for $\ell \neq \ell' \in [k]$ we have

$$\sum_{i\in S_\ell} B_{i,\ell} \geq p_0, \tag{21}$$

$$\sum_{i\in S_\ell} B_{i,\ell'} \leq p_1. \tag{22}$$

- **Pure Records Assumption.** There are pairwise disjoint sets $R_1, R_2, \ldots, R_k \subseteq [n]$ and a constant $\epsilon \in (0,1)$ such that for each $\ell \in [k]$ we have
$$\sum_{j \in R_\ell} C_{\ell,j} \geq (1-\epsilon) \sum_{j \in R_\ell} \|C_{\cdot,j}\|_1. \tag{23}$$

Under the above conditions, we have the following theorem.

Theorem 3. *Suppose that B, C are $d \times k$ and $k \times n$ matrices with non-negative entries satisfying the Normalization, Dominant Features and Pure Records Assumptions above. Suppose that $B' \in \mathbb{R}_+^{d\times k}$ and $C' \in \mathbb{R}_+^{k \times n}$ are such that $B'C'$ is close to BC in the following aggregate sense: There is a constant $\delta \in (0,1)$ such that for all $\ell \in [k]$ we have*
$$\|\sum_{j \in R_\ell} (BC)_{\cdot,j} - \sum_{j\in R_\ell} (B'C')_{\cdot,j}\|_1 \leq \delta \sum_{j \in R_l} \|C_{\cdot,j}\|_1. \tag{24}$$

Let $\delta' := 2\epsilon + 6\delta$. Also assume that the following conditions are satisfied for constants $p_0 > p_1$:
$$(p_0 - \delta')^2 > 4k(p_1 + \delta'), \tag{25}$$
$$2\delta' < p_0 - p_1. \tag{26}$$

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

Note that the second (approximate) factorization $B'C'$ has no conditions on the factors apart from nonnegativity. The proof of the above theorem is built upon the proof for the exact case.

Proof. Without loss of generality we may assume that each column of B' also sums to 1. We aggregate the columns in C and C' by simply replacing the set of columns in R_ℓ by their sum for each ℓ (thus replacing the set of $|R_\ell|$ columns by a single column). Now assume that for $l = 1, 2, \ldots k$, column l of C (respectively of C') is the sum of all columns in R_l of the original C (respectively C'). The Pure Records condition on C can now be restated as: For $\ell \in [k]$
$$C_{\ell,\ell} \geq (1-\epsilon)\|C_{\cdot,\ell}\|_1. \tag{27}$$

Condition (24) becomes: for $\ell \in [k]$
$$\|(BC)_{\cdot,\ell} - (B'C')_{\cdot,\ell}\|_1 \leq \delta \|C_{\cdot,\ell}\|_1. \tag{28}$$

We first outline the proof for which we need one piece of notation: We rescale columns of C and C' to get matrices $\bar{C}$ and $\bar{C}'$, respectively, as follows: $\bar{C}_{\cdot,\ell} := C_{\cdot,\ell}/\|C'_{\cdot,\ell}\|_1$ and $\bar{C}'_{\cdot,\ell} := C'_{\cdot,\ell}/\|C'_{\cdot,\ell}\|_1$ (note that in both cases we are dividing by $\|C'_{\cdot,\ell}\|_1$). For the outline, let's say two matrices P, Q are approximately equal, denoted $P \approx Q$ if the l_1 distance between each column of P and the corresonding column of Q is small (to be quantified later). The proof has the following steps:

1. We first prove
$$B \approx B\bar{C}.$$
This plus the hypothesis (24) will imply that $B \approx B'\bar{C}'$ which says that the convex hull of the columns of B, denoted CH(B) is contained in the convex hull of the columns of B':
$$\text{CH}(B) \subseteq \text{CH}(B').$$
2. We prove that for any $d \times k$ matrix P with non-negative entries and all column sums 1, with CH(B) $\subseteq$CH(P), we must have $P \approx B$ after possibly permuting columns of P.

Note first that for $\ell = 1, 2, \ldots, k$,
$$\begin{aligned}\|(BC)_{\cdot,\ell} - (B'C')_{\cdot,\ell}\|_1 &= \|BC_{\cdot,\ell} - B'C'_{\cdot,\ell}\|_1 \\ &\geq \left|\|BC_{\cdot,\ell}\|_1 - \|B'C'_{\cdot,\ell}\|_1\right| \\ &= \left|\|C_{\cdot,\ell}\|_1 - \|C'_{\cdot,\ell}\|_1\right|,\end{aligned}$$

where we have used the Normalization Assumption to get the last equation. Hence by (28) we have
$$k\left|\|C_{\cdot,\ell}\|_1 - \|C'_{\cdot,\ell}\|_1\right| \leq \delta\|C_{\cdot,\ell}\|_1. \tag{29}$$

We assume, w.l.o.g., that the columns of B' sum to 1. From (29), we get
$$\|C_{\cdot,\ell}\|_1 \leq (1+2\delta)\|C'_{\cdot,\ell}\|_1, \tag{30}$$

and so
$$\left|\|\bar{C}_{\cdot,\ell}\|_1 - 1\right| \leq \frac{\delta}{k}\frac{\|C_{\cdot,\ell}\|_1}{\|C'_{\cdot,\ell}\|_1} \leq 2\delta. \tag{31}$$

From (27), for $\ell \leq k$, $\bar{C}_{\ell,\ell} \geq (1-\varepsilon)\|\bar{C}_{\cdot,\ell}\|_1 \geq (1-\varepsilon)(1-2\delta/k)$ and so, $\sum_{\ell':\ell'\neq\ell} \bar{C}_{\ell',\ell} \leq \varepsilon\|\bar{C}_{\cdot,\ell}\|_1 \leq \varepsilon(1+2\delta)$. Also, $\bar{C}_{\ell,\ell} \leq \|\bar{C}_{\cdot,\ell}\|_1 \leq 1 + 2\delta/k$. Using this, we show that the first k columns of $B\bar{C}$ are close to the columns of B:
$$\begin{aligned}\|B_{\cdot,\ell} - B\bar{C}_{\cdot,\ell}\|_1 &= \|(1-\bar{C}_{\ell,\ell})B_{\cdot,\ell} + \sum_{\ell':\ell'\neq\ell} \bar{C}_{\ell',\ell}B_{\cdot,\ell'}\|_1 \\ &\leq \left|1-\bar{C}_{\ell,\ell}\right| + \sum_{\ell':\ell'\neq\ell}\left|\bar{C}_{\ell',\ell}\right| \\ &\leq (\epsilon + 2\delta) + \epsilon(1+2\delta) \\ &\leq 2\epsilon + 4\delta.\end{aligned}$$

Now it follows that the first k columns of $B'\bar{C}'$ are close to the columns of B:

$$\begin{aligned}\|B_{\cdot,\ell} - B'\bar{C'}_{\cdot,\ell}\|_1 &\le \|B_{\cdot,\ell} - B\bar{C}_{\cdot,\ell}\|_1 + \|B'\bar{C'}_{\cdot,\ell} - B\bar{C}_{\cdot,\ell}\|_1 \qquad (32)\\ &\le 2\epsilon + 4\delta/k + 2\delta \qquad (33)\\ &< 2\epsilon + 6\delta =: \delta' \text{ (say).}\end{aligned}$$

This completes the first step of the outline. To give some intuition, it is useful to interpret (32) in words. Since the column sums of $\bar{C}'$ are all 1, the columns of $B'\bar{C}'$ are convex combinations of the columns of B' and so we have proved that for each column of B, there is a vector in the convex hull of columns of B' at l_1-distance at most δ'.

The second step is proved in the following lemma:

Lemma 15. *Suppose B satisfies the hypothesis and B' is any $d \times k$ matrix with non-negative entries and all column sums equal to 1 such that the convex hull of the columns of B' contains a point at l_1-distance at most δ' from each column of B. Then, for $\ell \in [k]$ after permuting the columns of B', we have*

$$\|B_{\cdot,\ell} - B'_{\cdot,\ell}\|_1 \le \delta' + \frac{4k(p_1+\delta')}{(p_0-\delta')}.$$

Proof. We now have to introduce some more notation: Let

$$S_{k+1} = [d] \setminus \cup_{\ell=1}^{k} S_\ell.$$

We define two $(k+1) \times k$ matrices V, U obtained by summing the rows in each S_ℓ respectively of B' and $B'\bar{C}$:

$$V_{\ell,\cdot} := \sum_{i\in S_\ell} B'_{i,\cdot} \;\; ; \;\; U_{\ell,\cdot} := \sum_{i\in S_\ell} (B'\bar{C})_{i,\cdot} \quad \text{for } \ell = 1, 2, \ldots, k+1.$$

Then, from (32) and the Dominant Features Assumption (21), we have for $\ell \in [k]$,

$$U_{l,l} \ge p_0 - \delta' =: q_0 \;\; ; \;\; U_{l',l} \le p_1 + \delta' =: q_1 \quad \text{for } \ell \ne \ell' \in [k]. \tag{34}$$

Note that $q_0 > 0$ by our hypothesis (26). Clearly, we have

$$U_{\cdot,l} = \bar{C}'_{1,\ell}V_{\cdot,1} + \bar{C}'_{2,\ell}V_{\cdot,2} + \ldots + \bar{C}'_{k,\ell}V_{\cdot,k}. \tag{35}$$

Our goal would be to show that there is a permutation $\pi : [k] \to [k]$ such that for each $U_{\cdot,\ell}$ there is a corresponding $V_{\cdot,\pi(\ell)}$ that is close.

Let $\tau \in (\epsilon, q_0)$ be a parameter whose value will be chosen later. We say that a $V_{\cdot,\ell}$ is *bad* if it has at least two coordinates each with value at least τ. We will show that for sufficiently large τ, no $V_{\cdot,\ell}$ is bad. To this end, first note that if $V_{\cdot,\ell}$ is bad then $\bar{C}'_{\ell,\ell'} \le q_1/\tau$ for all ℓ'. Thus bad $V_{\cdot,\ell}$'s contribute at most kq_1/τ to $U_{\cdot,1}$ via decomposition (35). We will choose τ to satisfy $q_0 - kq_1/\tau > \tau$. Hence there must exist an $\ell \in [k]$ with $V_{1,\ell} \ge U_{1,1} - kq_1/\tau$ and $V_{\cdot,\ell}$ is not bad. We associate this $V_{\cdot,\ell}$ to $U_{\cdot,1}$. We swap columns 1 and ℓ of B' (and so of V) as well as rows 1 and ℓ of C' (and also of $\bar{C}'$). After the swap, we have (using (34)),

$$V_{1,1} \ge U_{1,1} - \frac{kq_1}{\tau} \ge \frac{q_0}{2},$$

now with the choice $\tau = \frac{2kq_1}{q_0}$ (note that this satisfies our requirement $q_0 - kq_1/\tau > \tau$ above by our hypothesis (26)). For all $\ell \ge 2$, from (35), we get

$$U_{1,\ell} \ge V_{1,1}\bar{C}'_{1,\ell} \quad \Longrightarrow \quad \bar{C}'_{1,\ell} \le \frac{q_1}{V_{11}} \le \frac{2q_1}{q_0}.$$

Thus,

$$\bar{C}'_{1,1} \ge 1 - \frac{2kq_1}{q_0}.$$

After repeating the argument and permuting the columns of B' and rows of $\bar{C}'$ suitably, we get

$$\bar{C}'_{\ell,\ell} \ge 1 - \frac{2kq_1}{q_0} \;\; ; \;\; \bar{C}'_{\ell',\ell} \le \frac{2q_1}{q_0} \text{ for } \ell' \ne \ell \in [k].$$

Hence, we get

$$\begin{aligned}\|B_{\cdot,\ell} - B'_{\cdot,\ell}\|_1 &\le \|B_{\cdot,\ell} - B'\bar{C}'_{\cdot,\ell}\|_1 + \|B'\bar{C}'_{\cdot,\ell} - B'_{\cdot,\ell}\|_1\\ &\le \|B_{\cdot,\ell} - B'\bar{C}'_{\cdot,\ell}\|_1 + \|B'_{\cdot,\ell}(1 - \bar{C}'_{\ell,\ell})\|\\ &+ \sum_{\ell':\ell'\ne\ell} \|B'_{\cdot,\ell'}\bar{C}'_{\ell',\ell}\|_1\\ &\le \delta' + \frac{2kq_1}{q_0} + \frac{2kq_1}{q_0} = \delta' + \frac{4kq_1}{q_0}.\end{aligned}$$

□

□

Let us remark that Identifiability does not need the Dominant Features Assumption (D2). But we do need a condition we did not need for the computation. Indeed, take the simple case when $k = 2 = d$ and

$$B = \begin{pmatrix} p_0 & p_1 \\ p_1 & p_0 \end{pmatrix},$$

where $1 > p_0 > p_1 > 01$. The two columns: $(p_0, p_1)^T, (p_1, p_0)^T$ are points on the line $\{(x_1, x_2) : x_1 + x_2 = p_0 + p_1\}$. We could take two other points $(1 + \varepsilon)(p_0, p_1)^T - \varepsilon(p_1, p_0)^T$ and $(1 + \varepsilon)(p_1, p_0)^T - \varepsilon(p_0, p_1)^T$ as the columns of $\tilde{B}$ as long as $(1 + \varepsilon)p_1 \ge \varepsilon p_0$ and the columns of B are convex combinations, so there is some $\tilde{C}$ satisfying $\tilde{B}\tilde{C} = BC$, upsetting identifiability. Note that if we required $\tilde{B}, \tilde{C}$ also to be Dominant NMF, then by the Pure Records property for $\tilde{B}, \tilde{C}$, we cannot have ε too big and so we would have some approximate identifiability. But we do not want to assume that $\tilde{B}, \tilde{C}$ satisfy Dominant NMF. So, instead, to ensure ε is not too big in this example, we need $p_1 << p_0$.

4. Additional Empirical Results on Synthetic and Real data:

In main paper we discuss synthetic experiments and show the robustness of **TSVDNMF** over the baselines. In this section, we extend the experiments by performing leave-one-out experiment on synthetic data to compare robustness of **TSVDNMF** with recent provable methods.

We also report the clustering results on some more baselines.

4.1. Leave-one-out Experiments under heavy Noise

The goal of this experiment is to compare the behavior of the NMF algorithms in extreme noise case when one of the data point is fully corrupted.

Baselines: We compare with Gillis-LP, SPA, ER-SPA and PW-SPA as described section 5.2 in main paper.

Datasets: The noisy data matrix A is generated under separability or dominant assumption with Gaussian or Multinomial noise, in the same way as described in section 5.2 in main paper. Additionally, at each iteration we drop ith data point (column of A) to form the observed matrix A_i and compute $\hat{B}_i$ and $\hat{C}_i$ to evaluate the performance by different algorithms and report the minimum over n iterations.

Performance Measure: We chose the same performance measure "ℓ_1-*residual*" as defined in section 5.2 of main paper, which is a variant of measure used by (Gillis & Luce, 2014). Let $A = BC+N$ be noisy data matrix with true factorization BC, A_i is the observed data matrix after removing ith column from A and C_i is the true co-efficient matrix after removing ith column from C. Let $\widehat{B_i}\widehat{C_i}$ be the output of an NMF algorithm with input A_i. The measure"ℓ_1-*residual*" $= 1 - \frac{\|BC_i - \widehat{B_i}\widehat{C_i}\|_s}{\|BC_i\|_s}$ 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, we generate a data matrix and report the worst performance over n leave-one-out iterations in Table 1. We observe **TSVDNMF** outperforms the baselines in all the settings except in separable data with Gaussian noise with noise level $\beta = 0.5$ where PW-SPA performs better and dominant data with Gaussian noise ($\beta = 1$) where SPA and ER-SPA performs better. On an average improvement of **TSVDNMF** is 12.7% over the baselines. Another interesting observation is in the dominant data with high Gaussian noise ($\beta = 2$) or multinomial noise ($m = 10$) 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 (0.011 and 0.003 respectively).

4.2. Comparison of TSVDNMF in task of clustering documents and faces:

In this section we perform clustering on four real text datasets and one face dataset to show the applicability of the proposed method on real-world tasks. For clustering by an NMF algorithm, we find the factorization BC by running the algorithm with input data matrix A and 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. [4] 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** (even if TSVD is not a provable NMF algorithm). Additionally we also compare with the heuristic LR-ER-SPA (A low rank variant of ER-SPA) as the recent work (Mizutani, 2014) claimed it to be the best for clustering.

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 of main paper. For completeness we reiterate the datasets as follows.

We use Reuters [5], 20 Newsgroups [6], TDT-2 [7], Yale face dataset[8]. 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

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

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

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

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

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

Table 1. Synthetic leave-one-out 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**		
β **or** m	$\beta = 0.5$	1	2	$\beta = 0.5$	1	2	$m = 10$	60	100	$m = 10$	60	100
Gillis-LP	0.724	0.603	0.308	0.449	0.198	-0.089	0.03	0.469	0.583	-0.03	0.332	0.39
PW-SPA	**0.776**	0.618	0.305	0.552	0.36	-0.046	0.044	0.516	0.595	-0.036	0.438	0.49
SPA	0.762	0.603	0.301	0.56	0.374	-0.024	0.045	0.532	0.591	-0.032	0.41	0.463
ER-SPA	0.767	0.62	0.313	0.558	**0.378**	-0.033	0.037	0.525	0.591	-0.032	0.418	0.474
TSVDNMF	0.756	**0.667**	**0.364**	**0.684**	0.367	**0.011**	**0.053**	**0.582**	**0.669**	**0.003**	**0.489**	**0.567**

Table 2. 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**
TSVD	0.501	0.58	**0.486**	**0.542**	0.402	0.345	0.685	0.535	0.746	0.763
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
LR-ER-SPA	0.496	0.581	0.43	0.313	0.364	0.314	0.633	0.575	0.734	0.745
XRAY	0.26	0.37	0.299	0.238	0.172	0.173	0.611	0.591	0.615	0.606
K-means	0.468	0.434	0.346	0.292	0.361	0.401	0.673	0.58	0.714	0.666
PW-SPA	0.369	0.382	0.363	0.24	0.191	0.186	0.603	0.515	0.658	0.655

from the SIFT features(Lowe, 1999). We further preprocessed the data by removing standard stop-words and by removing 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.

Performance measures: (same as in the main paper) Following (Mizutani, 2014), we used two metrics, accuracy (AC) and the normalized mutual information (NMI) to evaluate the clustering performance on real datasets. Let $\mathcal{T} = \{\mathcal{T}_1, \cdots \mathcal{T}_k\}$ be a partition of the dataset (of n data points) according to the class labels provided with the dataset 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 solve a bipartite matching problem to align $\mathcal{S} = \{\mathcal{S}_1, \cdots \mathcal{S}_k\}$ and $\mathcal{T} = \{\mathcal{T}_1, \cdots \mathcal{T}_k\}$ such that total number of common data points $|\mathcal{S}_i \cap \mathcal{T}_j|$ is maximized and compute AC $= \frac{1}{n} \sum_{j=1}^{k} |\mathcal{S}_j \cap \mathcal{T}_j|$. NMI is defined as NMI$(\mathcal{S}, \mathcal{T}) = \frac{I(\mathcal{S},\mathcal{T})}{[H(\mathcal{S})+H(\mathcal{T})]/2}$ where I and H denote the mutual information and entropy of $\mathcal{T}$ and $\mathcal{S}$ respectively. See (Manning et al., 2008) for details.

Observation: Table 2 shows the Accuracy and NMI of the clustering by **TSVDNMF** and the baselines. For all the datasets, **TSVDNMF** outperforms the provable baselines comprehensively. On average, over the provable algorithms, **TSVDNMF** gives atleast **33%** improvement in NMI on 4 text datasets and **6.6%** improvement in NMI on the Yale-face recognition dataset. **TSVDNMF** also outperforms all the methods on all the datasets except TSVD on Reuters48 where the performance is comparable. This may be due to the presence of large number of small sized classes (containing few data points) in Reuters48. 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.

References

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, 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.

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.

Kumar, Amit and Kannan, Ravindran. Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA'*, pp. 299–308, 2010.

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.

Lowe, David G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pp. 1150–1157. Ieee, 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.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. 2010. URL http://arxiv.org/abs/1011.3027.