A rank estimation criterion using an NMF algorithm under an inner dimension condition

Nam H. Lee I-Jeng Wang Runze Tang
Michael Rosen Carey E. Priebe

June 25, 2014

Abstract
We introduce a rank selection criterion for non-negative factorization algorithms, for the cases where the rank of the matrix coincides with the inner dimension of the matrix. The criteria is motivated by noting that provided that a unique factorization exists, the factorization is a solution to a fixed point iteration formula that can be obtained by rewriting non-negative factorization together with singular value decomposition. We characterize the asymptotic error rate for our fixed point formula when the non-negative matrix is observed with noise generated according to the so-called random dot product model for graphs.

1 Introduction

Non-negative factorization of a matrix whose columns are sub-probability vectors We consider an \( n \times p \) non-negative matrix \( X \) such that the rank of \( X \) is \( r \) and

\[
X = WH,
\]

(1)

where \( W, H \geq 0 \), \( 1^T W = 1^T \in \mathbb{R}^r \), and \( 1^T H \leq 1^T \in \mathbb{R}^p \). Note also that \( 1^T X \leq 1^T \). We call the integer \( r \) the inner dimension of the decomposition \( WH \). In general, a non-negative factorizable matrix can be transformed to a form stated in (1) by way of the so-called “pull-back” map (c.f. [1]). Given such \( X \), generally speaking, finding the pair \((W, H)\) is known to be an NP-hard problem (c.f. [2] and [3]), and even validating the uniqueness of the factorization remains a challenging task (c.f. [4] and [5]). As such, one is often interested in an algorithm for obtaining such a factorization approximately by numerically solving the following optimization problem:

\[
(\hat{W}, \hat{H}) := \arg \min_{W \geq 0, H \geq 0, 1^T W = 1^T, 1^T H \leq 1^T} D(X|WH),
\]

(2)

where \( D \) is some measure of discrepancy between \( X \) and the product \( WH \). For example, one can iterate until \( \|X - \hat{W}\hat{H}\|_F \) is sufficiently small. For data
encountered in practice, the error made by estimates $\hat{W}$ and $\hat{H}$ can not be known because true $W$ and $H$ are unknown. For more detailed discussion on non-negative factorization beyond a brief summary of the topic in this section we refer the reader to [1].

We now introduce a rank estimation criterion expressed in terms of deviation of $\hat{W}$ and $\hat{H}$ from their respective fixed points together with $\|X - \hat{W}\hat{H}\|_F$. Let

\[
\varepsilon(\hat{W}, \hat{H}) := \|X - \hat{W}\hat{H}\|_F, \quad \varepsilon(W) := \|F(W, \hat{H})\|_F, \quad \text{and} \quad \varepsilon(H) := \|G(\hat{W}, \hat{H})\|_F,
\]

where $X = UV^\top$ is a singular value decomposition of $X$, and

\[
F(W, H) := W - XH^TW^T(U \Sigma^{-2} U^\top)W, \quad \quad G(W, H) := H^T - X^TWH(V \Sigma^{-2} V^\top)H^T.
\]

**Definition 1.1.** Let $W$ and $H$ be non-negative matrices such that $WH$ is an $n \times p$ non-negative matrix. Then, the fixed-point information criteria (FIC) is

\[
\text{FIC}(W, H) := \varepsilon(W, H) + \varepsilon(W) + \varepsilon(H).
\]

We denote by $\hat{r}$ the inner dimension used to obtain an approximate non-negative factorization $(\hat{W}, \hat{H})$. In general, the bigger the value of $\hat{r}$ is, the smaller $\varepsilon(\hat{W}, \hat{H})$ is. On the other hand, it can be argued that for $\hat{r} > r$, the $\hat{r}$th singular value approximates zero, and as such, $\varepsilon(\hat{W})$ and $\varepsilon(\hat{H})$ are expected to be a large number if not infinity. It motivates our formulation behind our FIC formula. In Lemma 2.1 and 2.2, we motivate functions in (3) and (4).

Our analysis to come will depend on the following assumption (Condition 1) which reduces our problem to estimating the inner dimension of the non-negative matrix. Under Condition 1, if the matrix is observed when there is no noise, then one can see that for $\hat{r} > r$, $\varepsilon(W) = \varepsilon(H) = \infty$ and $\varepsilon(W, H) = 0$.

Hence, for such simple cases, while the exact analysis of $\hat{r} \leq r$ with respect to its FIC value remains elusive, our FIC will always suggest $\hat{r} \leq r$.

**Condition 1.** A non-negative matrix $X$ is a rank $r$ matrix, and there exists a unique non-negative factorization $W_H$ with inner dimension $r$.

**Related works** The choice for the inner dimension is often left to the practitioner’s discretion with little guiding principles for an informed choice. Previous related works of the model selection aspect of using NMF include [6], [7], [8]. In [6], so-called core consistency metric is introduced for a PARAFAC tensor model. In [7], [8], the core consistency is used to estimate the model dimension. It has been shown in [9] that there is a close relationship between non-negative matrix factorization and $K$-means clustering. Since the rank of a matrix is the number of non-zero singular values, a reasonable and popular approach, especially for noisy data, is to find the “effective” number of non-zero singular values. For example, in [10], an automatic “elbow” finding algorithm has been proposed, and their theory is developed based on the theory of multivariate normal random vectors.
2 Fixed Point Error Criterion

Note that whenever the inverses of $HH^\top$ and $WW^\top$ exist, we have that

$$W = U \left( \Sigma V^\top H (HH^\top)^{-1} \right),$$

(6)

$$H = \left( (W^\top W)^{-1} W^\top U \Sigma \right) V^\top,$$

(7)

where $X = U \Sigma V^\top$ is a singular value decomposition. Note that the indexes of the extreme points of the rows of $U$ match the indexes of the extreme points of the rows of $W$, and similarly, the indexes of the extreme points of the rows of $V$ match the indexes of the extreme points of the columns of $H$. This interplay between singular value decomposition and non-negative factorization has been exploited, for example, in [11] and [4] as we do in this paper as well.

**Lemma 2.1.** Under Condition 1,

$$W = X H^\top (U \Sigma^{-2} U^\top) W,$$

(8)

$$H^\top = X^\top W H (V \Sigma^{-2} V^\top) H^\top.$$  

(9)

**Proof.** Since the rank of $W$ must be $r$, $U^\top W$ and $W^\top U$ are full rank square matrices, whence their inverse exist. Next, we observe that

$$\Sigma^{-1} U^\top W (HH^\top) W^\top U \Sigma^{-1} = \Sigma^{-1} U^\top (U \Sigma V^\top) V \Sigma U^\top U \Sigma^{-1} = \Sigma^{-1} I (\Sigma \Sigma) I \Sigma^{-1} = I,$$

and by rearranging non-singular matrices, it follows that $HH^\top = (U^\top W)^{-1} \Sigma^2 (U^\top U)^{-1}$. Hence, we have $(HH^\top)^{-1} = (U^\top W)^{\top} \Sigma^{-2} U^\top W$. Next, we note that $WHH^\top = U \Sigma V^\top H^\top$, and therefore,

$$W = X H^\top (HH^\top)^{-1} = X H^\top (U^\top W)^{\top} \Sigma^{-2} (U^\top W) = X H^\top W (U \Sigma^{-2} U^\top) W.$$

By a similar argument, we can see that $H^\top = X^\top W H (V \Sigma^{-2} V^\top) H^\top$. □

**Lemma 2.2.** Under Condition 1, any pair $(W,H)$ of $n \times r$ and $r \times p$ non-negative matrices that satisfies (8) and (9) replacing the pair $(W,H)$ is such that their product yields $X$.

**Proof.** Consider the problem of finding a pair $(W,H)$ such that

$$W = X (Z)^{\top} (U \Sigma^{-2} U^\top) W,$$

(10)

$$H^\top = X^\top (Z) (V \Sigma^{-2} V^\top) H^\top,$$

(11)

where $Z := WH$. First, we write $Z = U' \Sigma (V')^\top$. Then, note the following equations:

$$Z = WH = (X H^\top W^\top U \Sigma^{-2} U^\top W) H,$$

(12)

$$X^\top ZZ^\top = X^\top X Z^\top U \Sigma^{-2} U^\top ZZ^\top,$$

(13)
where (13) is obtained from (12) by left and right multiplying $X^\top$ and $Z^\top$ to (12). Then, simplifying (13) with singular value decomposition, we have 
\[ \hat{V} \Sigma \hat{U}^\top = \Sigma_2 \Sigma^{-1} \Sigma V^\top \Sigma \Sigma^{-1} \Sigma U \hat{U}^\top \]  
and subsequently, 
\[ \Sigma = \Sigma_2 \Sigma^{-1} (Z)^\top \Sigma \Sigma^{-1} \Sigma U \hat{U}^\top. \]
Therefore, we have $I = \Sigma V^\top Z^{-1} \Sigma^{-2} \Sigma^\top U \hat{U}^\top$, or equivalently, 
\[ \Sigma = \Sigma_2 \Sigma^{-1} (Z)^\top \Sigma \Sigma^{-1} \Sigma U \hat{U}^\top. \]
Hence, 
\[ \Sigma V^\top \Sigma = \Sigma V^\top (Z)^\top \Sigma \Sigma^{-1} \Sigma U \hat{U}^\top \]  
and hence, 
\[ \Sigma = (\Sigma V^\top (Z)^\top \Sigma \Sigma^{-1} \Sigma U \hat{U}^\top) \Sigma V^\top (Z)^\top \Sigma \Sigma^{-1} \Sigma U \hat{U}^\top. \]
Subsequently, we have 
\[ \Sigma_2 = U \Sigma U^\top. \]
This implies $X = \Sigma V^\top = \Sigma U^\top Z$. On other hand, since both $\Sigma$ and $Z$ have rank $r$, it follows, from $Z = \Sigma U^\top Z + (I - \Sigma U^\top)Z$, that the rank of $(I - \Sigma U^\top)Z$ must be zero, or equivalently, $Z = \Sigma U^\top Z = \Sigma$ as desired. 
\[ \square \]
Combining Lemma 2.1 and 2.2 we have an if-and-only-if statement, and we list Theorem 2.1 for future reference:

**Theorem 2.1.** Under Condition 1, $X = WH$ is a non-negative factorization with inner dimension $r$ if and only if $(W,H)$ is a solution to the fixed point problem specified by (3) and (4), and the product $WH$ is of rank $r$.

Our next result in Theorem 2.2 gives an upper bound on the fixed point error $\|e(W)\|_F$ when $W$ may be different from $\hat{W}$, which will be used in the proof of Theorem 3.1. Note that we do not explicitly state the error bound associated with $\hat{H}$, but by symmetry, i.e., by transposing the matrix $X$, one can obtain a similar bound.

**Theorem 2.2.** There exist $\gamma_1$, $\gamma_2$, $\gamma_3$ and $\gamma_4$, depending only on $X$, such that for any rank $r$ matrix $X = USV^\top = WH$ with inner dimension $r$ decomposition,
\begin{align*}
\|e(W)\|_F &\leq \left(\|I - \Sigma U^\top\|_F + \gamma_1\|\Sigma\|_F\right)\|\Sigma U^\top\|_F + \gamma_2\|\Sigma\|_F\|\Sigma V - V\|_F \\
&\quad + \gamma_3\|\Sigma - \Sigma\|_F + \gamma_4\|\Sigma\|_F\|\Sigma U^\top - U\|_F\|\Sigma V - V\|_F,
\end{align*}
(14) (15)

**Proof.** Let $M := X(X)^\top (\Sigma_2 \Sigma^{-2} \Sigma^\top)$, and note that using the singular value decomposition,
\[ E := W - MW = U - MU, \]
where the fact that $\hat{W}$, $\hat{H}$, $W$ and $H$ are of rank $r$ is used for the identity. Now,
\[ M = \Sigma_2 \Sigma^{-2} \Sigma^\top V \Sigma U^\top U \Sigma^{-2} \Sigma^\top = \Sigma_2 \Sigma^{-2} \Sigma^\top (V - V) + I \Sigma (U^\top (U - U) + I) \Sigma^{-2} \Sigma^\top. \]
Then, we let $E = U - MU$, and note that
\begin{align*}
E &= U - U \Sigma \Sigma^{-2} \Sigma^\top U - U \Sigma (V^\top (V - V)) \Sigma (U^\top (U - U)) \Sigma^{-2} \Sigma^\top \\
&\quad - U \Sigma (V^\top (V - V)) \Sigma^{-2} \Sigma^\top U - U \Sigma \Sigma^{-2} \Sigma^\top (U - U) \Sigma^{-2} \Sigma^\top.
\end{align*}
(16)
Also, we have
\[ U - U\Sigma\Sigma^{-1}U^T U = U - U + UU^TU - U\Sigma\Sigma^{-1}U^TU \]
\[ = (U - U) + UU^T(U - U) + U(I - \Sigma\Sigma^{-1})U^TU \]
\[ = (I - UU^T)(U - U) + U(I - \Sigma\Sigma^{-1})U^TU. \quad (17) \]

Hence, using the triangular inequality,
\[
\|E\|_F \leq \|I - UU^T\|_F \|U - U\|_F + \|U\|_F \|I - \Sigma\Sigma^{-1}\|_F \|U^TU\|_F
+ \|X\|_F \|V - V\|_F \|\Sigma\|_F \|U\|_F \|U^T - U^T\|_F \|\Sigma^{-2}\|_F \|U^T\|_F
+ \|X\|_F \|V - V\|_F \|\Sigma\|_F \|\Sigma^{-2}\|_F \|U^T\|_F
\]
\[
+ \|\Sigma\|_F \|Y\|_F \|\Sigma\|_F \|\Sigma^{-2}\|_F \|U^T\|_F \|\Sigma^{-2}\|_F \|U^T\|_F. \]

Note that \(\|U\|_F = \|U\|_F = \|V\|_F = \|V\|_F = \sqrt{r}\) since \(U, V, U\) and \(V\) are unitary matrices. Then, let \(\gamma_1 := \|U\|_F \|\Sigma\|_F \|\Sigma^{-2}\|_F \|U\|_F, \gamma_2 := \|X\|_F \|\Sigma^{-2}\|_F \|U^T\|_F, \gamma_3 := \|\Sigma\|_F \|Y\|_F \|\Sigma^{-1}\|_F \|Y\|_F, \gamma_4 := \|X\|_F \|\Sigma^{-2}\|_F \|U^T\|_F \|U^T\|_F. \)

\[ \begin{aligned}
\end{aligned} \]

3 Application to Random Dot Product Graphs

In this section, we consider our fixed point error \(\varepsilon(W)\) with respect to a case where instead of observing \(X\) exactly, a noisy version of \(X\) is observed. To this end, we consider a random dot product graph model, which was originally introduced in [12].

**Definition 3.1.** Let \(E\) be a subset of \(\mathbb{R}^r\) such that, for all \(\omega, \omega' \in E\), \(0 \leq \langle \omega, \omega' \rangle \leq 1\). For any given \(n \geq 1\), let \(Y\) be a \(n \times d\) matrix whose rows \(\{Y_i\}_{i=1}^n\) are elements of \(E\). The adjacency matrix \(A\) of a random dot product graph (RDPG) with latent positions \(Y\) is a random \(n \times n\) symmetric non-negative matrix such that its entries take values in \(\{0, 1\}\) and
\[
P[A(Y_i)_{i=1}^n] = \prod_{i<j} (Y_iY_j^T)^{A_{ij}}(1 - Y_iY_j^T)^{1 - A_{ij}}.
\]

Given an \(n \times d\) matrix of latent positions \(Y\), the random dot product model generates a symmetric (adjacency) matrix \(A\) whose edges \(\{A_{ij}\}_{i<j}\) are independent Bernoulli random variables with parameters \(\{P_{ij}\}_{i<j}\) where \(P = YY^T\). Random dot product graphs are a specific example of latent position graphs [13].

Our analysis in this section is an asymptotic one in which the number \(n\) of rows is taken to infinity. As such, to indicate this explicitly, we may write, for example, \(\varepsilon_n(W)\) instead of \(\varepsilon(W)\). But, for the most parts, to simplify our notation, we suppress our notation’s dependence on \(n\).

Let \(A\) be an \(n \times n\) random matrix such that \(P = E[A|Y] = YY^T\), where the rows \(\{Y_i\}_{i=1}^n\) of \(Y\) form a sequence of independent and identically distributed
random probability vectors, i.e., $Y \mathbf{1} = \mathbf{1}$, and conditioning on $Y$, each $A_{ij}$ is an independent Bernoulli random variable. Let $\hat{Y}$ be the adjacency spectral embedding of $A$ at rank $r$. That is, given that $A = \hat{U} \hat{\Sigma} \hat{V}^\top$ is the singular value decomposition with its singular values in the decreasing order, $\hat{Y} = \hat{U}_+ \hat{\Sigma}_+^{1/2}$, where $\hat{U}_+$ is the first $r$ columns of $\hat{U}$ and $\hat{\Sigma}_+$ is the upper $r \times r$ diagonal sub-matrix of $\hat{\Sigma}$. The matrix $\hat{Y}$ is known as the adjacency spectral embedding, which is shown to have good properties. For more detailed analysis of $\hat{Y}$, we refer the interested reader to [14] and to the references therein.

The following condition is a key assumption in [14] to which we make references for several inequalities used in the proof of Theorem 3.1.

**Condition 2.** The distribution of $Y_i$ does not change with $n$, and the $r \times r$ second moment matrix $\Delta := \mathbb{E}[Y_i^\top Y_i]$ has distinct and strictly positive eigenvalues.

Our discussion on estimation of the rank for non-negative factorization connects to the random dot product model by a simple observation that even if $n \times r$ matrix $Y$ is not non-negative, if $P = YY^\top$ is a non-negative matrix with rank $r$, then there exists an $n \times r$ non-negative matrix $W$ such that $P = WW^\top$ (c.f. [4]). Our next condition is a stronger version of this observation.

**Condition 3.** Suppose that for each $n$, there exists an orthogonal matrix $Q$ such that
\[
\frac{1}{n} \hat{Y} Q = \hat{W} := \arg \min_{W \in \mathbb{R}^{n \times r}_+} \sum_{i<j} \left| \frac{1}{n} A_{ij} - e_i^T W W^\top e_j \right|^2.
\] (18)

**Condition 4.** $P = U \Sigma U^\top$ is the eigenvalue decomposition, where $\Sigma$ is a non-singular $r \times r$ diagonal matrix whose diagonal values are in decreasing order. There exists a constant $\xi_0 < \infty$ such that almost surely, for all $n$,
\[
\tau(P) := \frac{\Sigma_{11}}{\Sigma_{rr}} \leq \xi_0.
\] (19)

Our next result characterizes the asymptotic error rate for our fixed point formula:

**Theorem 3.1.** Let $\{a_n\}$ be a sequence such that $\lim_{n \to \infty} \sqrt{\log(n)}/a_n = 0$. Under Condition [4], [3] and [4] almost surely,
\[
\lim_{n \to \infty} \varepsilon_n(\hat{W})/a_n = 0,
\] (20)

where $\varepsilon_n(\cdot)$ denotes the fixed point error for $\bar{X} := P/n = YY^\top/n$ with $\bar{W} := (1/\sqrt{n})Y$ and $\bar{H} := (1/\sqrt{n})Y^\top$.

**Proof.** Note that $1^T A e_j \leq n$ for all $j = 1, \ldots, n$. First, by Proposition 4.5 and Theorem 4.6 in [14], for each $\varepsilon > 0$, for all sufficiently large values of $n$, with probability $1 - \varepsilon$,
\[
\|\Sigma - \hat{\Sigma}_+\|_F \leq C\delta^{-2} r \log(n/\varepsilon),
\] (21)
\[
\|U - \hat{U}_+\|_F \leq 4\delta^{-2} \sqrt{2r \log(n/\varepsilon)/n},
\] (22)
where $C$ is the constant in Remark 3.7 of [14] which depends only on the (common) distribution of $Y_i$ and $\delta_r$ denotes the smallest eigenvalue of the second moment matrix $\Delta = E[Y_i^T Y_i]$. Now, since $\bar{X} = P/n$ and $P = USU^T$, $\bar{X} = U(\Sigma/n)U^T$ is an eigenvalue decomposition of $\bar{X}$. Now, we write $U = U$ and $\Sigma = \Sigma/n$. Note that using/adapting the equations (16) and (17) from the proof of Theorem 2.2.

\[ \|\xi_n(\hat{W})\|_F \leq (\sqrt{n} + r)\|\hat{U}_+ - U\|_F + r^{3/2}(\sqrt{n} + r\|\hat{\Sigma}_+\|_F\|\Sigma^{-1}\|_F) + r^2\|\Sigma\|_F\|\hat{\Sigma}_+\|_F\|\Sigma^{-2}\|_F\|\hat{U}_+ - U\|_F + r^{3/2}\|\Sigma\|_F\|\hat{\Sigma}_+\|_F\|\Sigma^{-2}\|_F, \]

where we implicitly used the fact that $\Sigma = \Sigma/n$ to . for example, simplify $\|\hat{\Sigma}_+\|_F\|\Sigma^{-1}\|_F$. Appearing Proposition 4.5 of [14] once again, we also have that

\begin{align*}
\|\hat{\Sigma}_+\|_F\|\Sigma^{-1}\|_F & \leq r(\|\hat{\Sigma}_+\|_F/\|\Sigma\|)\|\Sigma\|\|\Sigma^{-1}\|_F \leq r(2/\delta_r)\xi_0, \quad (23) \\
\|\Sigma\|_F\|\hat{\Sigma}_+\|_F\|\Sigma^{-2}\|_F & \leq r^{3/2}\|\Sigma\|_F\|\Sigma^{-2}\|_F\|\Sigma^{-2}\|_F\|\Sigma^{-1}\|_F \leq r^{3/2}\xi_0^2/\delta_r, \quad (24)
\end{align*}

where $\|\cdot\|$ denotes the spectral norm, i.e., the largest singular value of the matrix. Therefore, we have the following inequality from which our claim follows:

\begin{align*}
\|\xi_n(\hat{W})\|_F & \leq (\sqrt{n} + r)\|\hat{U}_+ - U\|_F + r^{3/2}(\sqrt{n} + r(2/\delta_r)\xi_0) + r^2r^{3/2}\xi_0^2/\delta_r\|\hat{U}_+ - U\|_F + r^{3/2}r^{3/2}\xi_0^2/\delta_r, \quad (25) \\
& + r^2r^{3/2}\xi_0^2/\delta_r\|\hat{U}_+ - U\|_F + r^{3/2}r^{3/2}\xi_0^2/\delta_r, \quad (26)
\end{align*}

\[\square\]

\section{4 Numerical Experiments}

\subsection{4.1 Computing environment}

For non-negative factorization during our numerical experiments, we have used \texttt{nmf} function from the \texttt{NMF} package from R 3.0.1 (64-bit) and \texttt{nmf}, \texttt{FastConicalHull} and \texttt{FastSepNMF} from MATLAB R2013b 8.2.0.701 (64-bit) under Mac OS X 10.9 on an Intel Core i5 @ 1.3 GHz machine with 4 GB RAM.

For \textit{Near Separable Cases} experiments, we use \texttt{FastConicalHull} and \texttt{FastSepNMF} which are implementations of the so-called “XRAY” algorithm in [11] respectively, and the linear programming algorithm in [15], and their MATLAB source codes can be found on the web\footnote{http://bit.ly/1eDKNem}. For \textit{Not Separable Cases} and \textit{Sensor Network Data} experiments, we use the \texttt{pe-nmf} option in \texttt{nmf} function that solves numerically the following problem:

\[ (\hat{W}, \hat{H}) := \arg\min_{W \geq 0, H \geq 0} \frac{1}{2} \|\bar{X} - WH\|_F^2 + \alpha \sum_{k_1 \neq k_2} e_{k_1}^T W e_{k_2} + \beta \sum_{k,t} H_{k,t}, \quad (27) \]
Using FastConicalHull together with FIC, 100 Monte Carlo experiments are performed.

| (a) $\epsilon = 0.3$ | (b) $\epsilon = 0.2$ | (c) $\epsilon = 0.1$ |
|----------------------|----------------------|----------------------|
| $\hat{r}$ | Count | $\hat{r}$ | Count | $\hat{r}$ | Count |
| 8- | 21 | 8- | 5 | 8- | 1 |
| 8 | 4 | 8 | 4 | 8 | 0 |
| 9 | 20 | 9 | 14 | 9 | 3 |
| 10 | 58 | 10 | 77 | 10 | 96 |
| 10+ | 3 | 10+ | 0 | 10+ | 0 |

where $1^\top W = 1^\top$ and $1^\top H = 1^\top$, and $\alpha, \beta \geq 0$ are chosen appropriately. More specifically, the particular numerical algorithm solving the aforementioned problem for our experiment is known as pattern-expression non-negative factorization (c.f. [16]). Unless said otherwise, we set $\alpha = 0$ and $\beta = 1$.

### 4.2 Near Separable Cases

**A case where the rank and inner dimension coincide but errorfully observed** Using the example in [17], we conduct a Monte Carlo simulation comparing the so-called “XRAY” algorithm in [11] and the algorithm in [15]. In particular, for each Monte Carlo simulation, we randomly generate a $50 \times 100$ matrix $\bar{M}$. To do this, we sample the entries of $50 \times 10$ non-negative matrix $W$ uniformly from $[0, 1]$ independently, and subsequently, we normalize the column so that each column is a probability vector, i.e., $W \leftarrow W \text{diag}(1^\top W)^{-1}$. Then, we generate a $10 \times 100$ matrix $\bar{H}$ by sampling each column from a Dirichlet distribution whose parameter is sampled uniformly from $[0, 1]^{10}$. Then, $\bar{M} = \bar{W} \bar{H}$. Then, the observed matrix $X$ is obtained from $\bar{M}$ by adding normally distributed errors:

\[ X = \bar{W}\bar{H} + N \text{diag}(1^\top |N|)^{-1} \varepsilon, \quad (28) \]

where $N_{ij}$ is a normal random variable with mean zero and standard deviation 1, the matrix $|N| = (|N_{ij}| : i, j)$, and $\varepsilon > 0$. If $X_{ij} < 0$ after adding noise, then we set the value of $X_{ij}$ to zero to enforce the non-negativity. The results are reported in Table [1] and Table [2] respectively. The results are comparable. We also mention that using the usual $L_2$ minimization results in a poor performance; our experiment results in $\hat{r} = 1$ for all 100 Monte Carlo experiments for all $\varepsilon = 0.1, 0.2, 0.3$.

**Swimmer Data Sets** The swimmer data set is a often-tested data set for bench-marking NMF algorithms (c.f. [18] and [17]). In our present notation, each column of $220 \times 256$ data matrix $X$ is a vectorization of a binary image, and each row corresponds to a particular pixel. Each image is a binary images (20-by-11 pixels) of a body with four limbs which can be each in four different
Table 2: Using FastSepNMF together with FIC, 100 Monte Carlo experiments are performed.

|       | (a) $\varepsilon = 0.3$ | (b) $\varepsilon = 0.2$ | (c) $\varepsilon = 0.1$ |
|-------|--------------------------|--------------------------|--------------------------|
| $\hat{r}$ | Count | $\hat{r}$ | Count | $\hat{r}$ | Count |
| 8-   | 15   | 8-   | 1   | 8-   | 1   |
| 8    | 3    | 8    | 0   | 8    | 1   |
| 9    | 23   | 9    | 13  | 9    | 1   |
| 10   | 52   | 10   | 86  | 10   | 97  |
| 10+  | 0    | 10+  | 0   | 10+  | 0   |

positions. It is known that the matrix $X$ is 16-separable while the rank of $X$ is 13. Such, Condition 1 is not met, and application of our FIC criteria using FastConicalHull and FastSepNMF yields the estimated $\hat{r}$ as 13 while using nnmf yields $\hat{r} = 1$.

4.3 Not Separable & Not unique Cases

In this experiment, using the parameters for the example considered in [4] and [5], we generate Monte Carlo samples of the matrix $X$ whose expected value is parametrized by $\kappa \in [0,1]$ and $\lambda \in (0, \infty)$. To this end, we let $X$ be a matrix of independent Poisson random variables such that

$$E[X] = WH \Lambda,$$

where $\Lambda = \lambda I$,

$$W = \frac{1}{2(1+\kappa)} \begin{pmatrix} \kappa & 1 & 0 \\ 1 & \kappa & 0 \\ 1 & 0 & \kappa \\ \kappa & 0 & 1 \\ 0 & \kappa & 1 \end{pmatrix} \quad \text{and} \quad H = \frac{1}{1+\kappa} \begin{pmatrix} \kappa & 1 & 1 & 0 & 0 \\ 1 & \kappa & 0 & 0 & 1 \\ 0 & 0 & \kappa & 1 & 1 \end{pmatrix}.$$  (30)

It can be shown that $WH$ is uniquely factorizable if and only if $\kappa \in [0,0.5)$ (c.f. [4]). In particular, by changing $\kappa$ from 0 to 1, we can gradually transition from a uniquely factorizable model to one that can be decomposed into more than two distinct but equally viable solutions.

We focus on correctly estimating $r = 3$ from $X \text{diag}(1^T X)^{-1}$. In Table 3 and Table 4 we consider the problem of estimating $r$ while fixing $\kappa = 0.1$ and $\kappa = 0.5$. We perform our experiments while varying the event intensity $\lambda$ from 10 to 10000. For given $\lambda$, i.e, each row, each entry counts the number of times out of 100 Monte Carlo experiments that resulted in choosing $\hat{r}$ as the inner dimension based on the baseline method and our FIC criteria. For each $\gamma$, we conduct 100 Monte Carlo simulation experiments for performance of choosing $r$ achieving the minimum FIC value. For comparison, we use as the
Table 3: The baseline procedure “dimSelect ◦ svd” is used for choosing \( \hat{r} \) for each of 100 Monte Carlo simulation experiments. The true rank \( r \) is 3. For \( \kappa = 0.1 \), the baseline procedure performs well for all values of \( \lambda \), but for \( \kappa = 0.5 \), the baseline procedure performs poorly.

| \( \lambda \) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 10 | 0 | 3 | 96 | 1 | 0 | 0 |
| 100 | 0 | 0 | 100 | 0 | 0 | 0 |
| 1000 | 0 | 0 | 100 | 0 | 0 | 0 |
| 10000 | 0 | 0 | 100 | 0 | 0 | 0 |

(a) \( \kappa = 0.1 \)  

| \( \lambda \) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 10 | 48 | 39 | 12 | 1 | 0 | 0 |
| 100 | 100 | 0 | 0 | 0 | 0 | 0 |
| 1000 | 100 | 0 | 0 | 0 | 0 | 0 |
| 10000 | 100 | 0 | 0 | 0 | 0 | 0 |

(b) \( \kappa = 0.5 \)

Table 4: FIC is used for choosing \( \hat{r} \) for each of 100 Monte Carlo simulation experiments for each \( \gamma \). The true rank \( r \) is 3. Performance of the procedure is behind that of dimSelect ◦ pca for \( \kappa = 0.1 \) but better for \( \kappa = 0.5 \), demonstrating robustness.

| \( \lambda \) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 10 | 0 | 3 | 10 | 13 | 22 | 52 |
| 100 | 0 | 0 | 66 | 16 | 11 | 7 |
| 1000 | 0 | 0 | 100 | 0 | 0 | 0 |
| 10000 | 0 | 0 | 100 | 0 | 0 | 0 |

(a) \( \kappa = 0.1 \)  

| \( \lambda \) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 10 | 0 | 19 | 18 | 28 | 29 | 6 |
| 100 | 0 | 14 | 38 | 28 | 19 | 1 |
| 1000 | 0 | 16 | 75 | 9 | 0 | 0 |
| 10000 | 0 | 17 | 83 | 0 | 0 | 0 |

(b) \( \kappa = 0.5 \)

baseline method the procedure of “dimSelect ◦ svd” for selecting the underlying dimension (or rather estimating the rank of the matrix), where the baseline method first computes the singular values of \( X \) and then, use the “elbow” finding method in [10] to estimate the rank. The result is reported in Table 3 and we note that the performance of our method is more robust than the baseline method, with respect to non-uniqueness of the underlying factorization.

4.4 Sensor Network Data

We next consider data collected over six hours and thirty minutes from a group of 19 actors working in an office. Each actor is equipped with one or more Bluetooth device(s) that detect other Bluetooth devices when in proximity. The overall idea behind collecting such data is similar to the one in [19]. There are 22 sensors in total, and they are associated with office staffs (doctors and nurses) as well as stationary objects such as a desk and a room. Three individuals were associated with two sensors and the rest are given a single sensor. The original
data is given in the following format: \( \mathcal{D}_T = \{(s, i, j) : i, j, s \in [0, T]\} \), where 
1 \leq i, j \leq 22 denotes the collection of all 22 sensors and 
\((s, i, j)\) denotes the event that sensor \(i\) detected sensor \(j\) at time \(s\). The interval \([0, T]\) is divided into 
13 equal length subintervals, and each interval represents a 30-minute window. The data matrix whose rank is estimated is an \(n \times p\) non-negative matrix \(X = X \text{diag}(1^\top X)^{-1}\), where \(X_{\ell,t}\) is the number of times that sensor \(i\) detected sensor \(j\) during \(t\)th interval and \(\ell = 22 \times (i - 1) + j\). Specifically, \(n = 22 \times 21 = 462\) and \(p = 13\).

As displayed in Table 5, the best choice in terms of FIC for rank \(r\) was 2. The values of \(\varepsilon(\hat{X})\) decreases monotonically while the graphs of the values of 
\(\varepsilon(\hat{W})\) and \(\varepsilon(\hat{H})\) exhibit convexity in \(\hat{r}\). In Figure 1, the basis elements from fitting the rank \(r = 2\) is illustrated in an “adjacency” matrix-like layout. An inspection of \(\hat{H}\) for \(\hat{r} = 2\) yields that the estimated model is nearly separable, where the first pattern in Table 5 dominates during the first few hours in the morning, and the second pattern in Table 5 dominates during the rest of the day.
Table 5: Selecting $\hat{r}$ for sensor network data. FIC suggests that $\hat{r} = 2$ and this coincides with the suggestion made by using “dimSelect $\circ$ svd”.

|     | 1   | 2   | 3   | 4   | 5   | 6   |
|-----|-----|-----|-----|-----|-----|-----|
| FIC | 0.84| **0.60** | 2.01 | 2.04 | 2.43 | 3.12 |
| $\varepsilon(\hat{W})$ | 0.11 | 0.03 | 0.06 | 0.13 | 0.14 | 0.22 |
| $\varepsilon(\hat{H})$ | **0.27** | 0.37 | 1.79 | 1.80 | 2.19 | 2.87 |
| $\varepsilon(\hat{X})$ | 0.46 | 0.19 | 0.15 | 0.10 | 0.09 | **0.02** |

References

[1] N. Gillis. The why and how of nonnegative matrix factorization. In J.A.K. Suykens, M. Signoretto, and A. Argyriou, editors, Regularization, Optimization, Kernels, and Support Vector Machines. Chapman & Hall/CRC, 2014.

[2] S. Vavasis. On the Complexity of Nonnegative Matrix Factorization. SIAM Journal on Optimization, 20(3):1364–1377, 2010.

[3] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In Proceedings of the 44th symposium on Theory of Computing, pages 145–162. ACM, 2012.

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

[5] Hans Laurberg, Mads Fræsboøll Christensen, Mark D Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on positive data: On the uniqueness of NMF. Computational Intelligence and Neuroscience, 2008, 2008.

[6] Rasmus Bro and Henk A. L. Kiers. A new efficient method for determining the number of components in parafac models. Journal of Chemometrics, 17:274–286, 2003.

[7] A. Barrat, C. Cattuto, A. E. Tozzi, P. Vanhems, and N. Voirin. Measuring contact patterns with wearable sensors: methods, data characteristics and applications to data-driven simulations of infectious diseases. Clinical Microbiology and Infection, 20(1):10–16, 2014.

[8] Gauvin L, Panisson A, and Cattuto C. Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach. PLoS ONE, 9(1), 2014.

[9] C. Ding and X. He. On the equivalence of nonnegative matrix factorization and spectral clustering. In Fifth SIAM Conference on Data Mining, pages 600–610. 2005.
[10] M Zhu and A. Ghodshi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51:918–930.

[11] Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 231–239. JMLR Workshop and Conference Proceedings, 2013.

[12] S. Young and E. Scheinerman. Random dot product graph models for social networks. In *Proceedings of the 5th international conference on algorithms and models for the web-graph*, pages 138–149, 2007.

[13] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[14] A. Athreya, V. Lyzinski, D.J. Marchette, C.E. Priebe, D.L. Sussman, and M. Tang. A limit theorem for scaled eigenvectors of random dot product graphs. [http://arxiv.org/abs/1305.7388](http://arxiv.org/abs/1305.7388), 2013.

[15] N. Gillis and S.A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(4):698–714, April 2014.

[16] Junying Zhang, Le Wei, Xuerong Feng, Zhen Ma, and Yue Wang. Pattern Expression Nonnegative Matrix Factorization: Algorithm and Applications to Blind Source Separation. *Computational Intelligence and Neuroscience*, 2008.

[17] N. Gillis and R. Luce. Robust near-separable nonnegative matrix factorization using linear optimization. *Journal of Machine Learning Research*, pages 1249–1280, 2014.

[18] David Donoho and Victoria Stodden. When does non-negative matrix factorization give correct decomposition into parts? page 2004. MIT Press, 2003.

[19] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-Francois Pinton, Nagham Khanafer, Corinne Regis, Byeul-a Kim, Brigitte Comte, and Nicolas Voiron. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, 8(9):e73970, 09 2013.