FIRST RESULTS ON UNIQUENESS OF SPARSE NON-NEGATIVE MATRIX FACTORIZATION

Fabian J. Theis, Kurt Stadlthanner and Toshihisa Tanaka[*]

Institute of Biophysics, University of Regensburg, 93040 Regensburg, Germany
phone: +49 941 943 2924, fax: +49 941 943 2479, email: fabian@theis.name
[*] Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology (TUAT)
2-24-16, Nakacho, Koganei-shi, Tokyo 184-8588 Japan and
ABSP Laboratory, BSI, RIKEN, 2-1, Hirosawa, Wako-shi, Saitama 351-0198 Japan

ABSTRACT

Sparse non-negative matrix factorization (sNMF) allows for the decomposition of a given data set into a mixing matrix and a feature data set, which are both non-negative and fulfill certain sparsity conditions. In this paper it is shown that the employed projection step proposed by Hoyer has a unique solution, and that it indeed finds this solution. Then indeterminacies of the sNMF model are identified and first uniqueness results are presented, both theoretically and experimentally.

1. INTRODUCTION

Non-negative matrix factorization (NMF) describes a promising new technique for decomposing non-negative data sets into a product of two smaller matrices thus capturing the underlying structure [3]. In applications it turns out that additional constraints like for example sparsity enhance the recoveries; one promising variant of such a sparse NMF algorithm has recently been proposed by Hoyer [2]. It consists of the common NMF update steps, but at each step a sparsity constraint is posed. If factorization algorithms are to produce reliable results, their indeterminacies have to be known and uniqueness (except for the indeterminacies) has to be shown — so far only restricted and quite disappointing results for NMF [1] and none for sNMF are known.

In this paper we first present a novel uniqueness result showing that the projection step of sparse NMF always possesses a unique solution (except for a set of measure zero), theorems 2.2 and 2.6. We then prove that Hoyer's algorithm indeed detects these solutions, theorem 2.8. In section 3 after shortly repeating Hoyer's sNMF algorithm, we analyze its indeterminacies and show uniqueness in some restricted cases, theorem 3.3. The result is both new and astonishing, because the set of indeterminacies is much smaller than the one of NMF, namely of measure zero.

2. SPARSE PROJECTION

The sparse NMF algorithm enforces sparseness by using a projection step as follows: Given $\mathbf{x} \in \mathbb{R}^n$ and fixed $\lambda_1, \lambda_2 > 0$, find $\mathbf{s}$ such that

$$\mathbf{s} = \operatorname{argmin}_{\|s\|_1=\lambda_1, \|s\|_2=\lambda_2, s\geq 0} \|\mathbf{x}-\mathbf{s}\|_2 \qquad (1)$$

Here $\|s\|_p := \left(\sum_{i=1}^n |s_i|^p\right)^{1/p}$ denotes the p-norm; in the following we often omit the index in the case $p = 2$. Furthermore $\mathbf{s} \geq 0$ is defined as $s_i \geq 0$ for all $i = 1, \ldots, n$, so $\mathbf{s}$ is to be *non-negative*. Our goal is to show that such a projection always exists and is unique for almost all $\mathbf{x}$. This problem can be generalized by replacing the 1-norm by an arbitrary p-norm, however the (Euclidean) 2-norm has to be used as can be seen in the proof later. Other possible generalizations include projections in infinite-dimensional Hilbert spaces.

First note that the two norms are equivalent i.e. induce the same topology; indeed $\|s\|_2 \leq \|s\|_1 \leq \sqrt{n}\|s\|_2$ for all $\mathbf{s} \in \mathbb{R}^n$ as can be easily shown. So a necessary condition for any $\mathbf{s}$ to satisfy equation (1) is $\lambda_2 \leq \lambda_1 \leq \sqrt{n}\lambda_2$.

We want to solve problem (1) by projecting $\mathbf{x}$ onto

$$M := \{\mathbf{s} | \|\mathbf{s}\|_1 = \lambda_1\} \cap \{\mathbf{s} | \|\mathbf{s}\|_2 = \lambda_2\} \cap \{\mathbf{s} \geq 0\} \qquad (2)$$

In order to solve equation (1), $\mathbf{x}$ has to be projected onto a point adjacent to it in M:

Definition 2.1. *A point $\mathbf{p} \in M \subset \mathbb{R}^n$ is called* adjacent to $\mathbf{x} \in \mathbb{R}^n$ in *M, in symbols $\mathbf{p} \lhd_M \mathbf{x}$ or shorter $\mathbf{p} \lhd \mathbf{x}$, if $\|\mathbf{x}-\mathbf{p}\|_2 \leq \|\mathbf{x}-\mathbf{q}\|_2$ for all $\mathbf{q} \in M$.*

In the following we will study in which cases this is possibly, and which conditions are needed to guarantee that this projection is even unique.

2.1 Existence

Assume that $\mathbf{x}$ lies in the closure of M, but not in M. Obviously there exists no $\mathbf{p} \lhd \mathbf{x}$ as $\mathbf{x}$ 'touches' M without being an element of it. In order to avoid these exceptions, it is enough to assume that M is closed:

Theorem 2.2 (Existence)**.** *If M is closed and nonempty, then for every $\mathbf{x} \in \mathbb{R}^n$ there exists a $\mathbf{p} \in M$ with $\mathbf{p} \lhd \mathbf{x}$.*

Proof. Let $\mathbf{x} \in \mathbb{R}^n$ be fixed. Without loss of generality (by taking intersections with a large enough ball) we can assume that M is compact. Then $f : M \to \mathbb{R}, \mathbf{p} \mapsto \|\mathbf{x}-\mathbf{p}\|$ is continuous and has therefore a minimum $\mathbf{p}_0$, so $\mathbf{p}_0 \lhd \mathbf{x}$. □

2.2 Uniqueness

Definition 2.3. *Let $\mathscr{X}(M) := \{\mathbf{x} \in \mathbb{R}^n | \text{there exists more than one point adjacent to } \mathbf{x} \text{ in } M\} = \{\mathbf{x} \in \mathbb{R}^n | \#\{\mathbf{p} \in M | \mathbf{p} \lhd \mathbf{x}\} > 1\}$ denote the* exception set *of M.*

In other words, the exception set contains the set of points from which we can't uniquely project. Our goal is to show that this set vanishes or is at least very small. Figure 1 shows the exception set of two different sets.

Note that if $\mathbf{x} \in M$ then $\mathbf{x} \lhd \mathbf{x}$, and $\mathbf{x}$ is the only point with that property. So $M \cap \mathscr{X}(M) = \varnothing$. Obviously the exception set of an affine linear hyperspace is empty. Indeed, we can prove more generally:

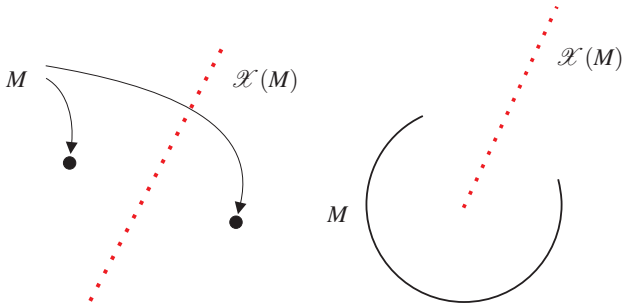
Lemma 2.4. *Let $M \subset \mathbb{R}^n$ be convex. Then $\mathscr{X}(M) = \varnothing$.*

For the proof we need the following simple lemma, which only works for the 2-norm as it uses the scalar product.

Lemma 2.5. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ such that $\|\mathbf{a}+\mathbf{b}\|_2 = \|\mathbf{a}\|_2 + \|\mathbf{b}\|_2$. Then $\mathbf{a}$ and $\mathbf{b}$ are collinear.*

Proof. By taking squares we get $\|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + 2\|\mathbf{a}\|\|\mathbf{b}\| + \|\mathbf{b}\|^2$, so

$$\|\mathbf{a}\|^2 + 2\langle \mathbf{a}, \mathbf{b}\rangle + \|\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + 2\|\mathbf{a}\|\|\mathbf{b}\| + \|\mathbf{b}\|^2$$



(a) exception set of two points (b) exception set of a sector

Figure 1: Two examples of exception sets.

if $\langle .,. \rangle$ denotes the (symmetric) scalar product. Hence $\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|\|\mathbf{b}\|$ and $\mathbf{a}$ and $\mathbf{b}$ are collinear according to the Schwarz inequality. □

Proof of lemma 2.4. Assume $\mathscr{X}(M) \neq \varnothing$. Then let $\mathbf{x} \in \mathscr{X}(M)$ and $\mathbf{p}_1 \neq \mathbf{p}_2 \in \mathscr{M}$ such that $\mathbf{p}_i \lhd \mathbf{x}$. By assumption $\mathbf{q} := \frac{1}{2}(\mathbf{p}_1 + \mathbf{p}_2) \in M$. But

$$\|\mathbf{x} - \mathbf{p}_1\| \leq \|\mathbf{x} - \mathbf{q}\| \leq \frac{1}{2}\|\mathbf{x} - \mathbf{p}_1\| + \frac{1}{2}\|\mathbf{x} - \mathbf{p}_2\| = \|\mathbf{x} - \mathbf{p}_1\|$$

because both $\mathbf{p}_i$ are adjacent to $\mathbf{x}$. Therefore $\|\mathbf{x} - \mathbf{q}\| = \|\frac{1}{2}(\mathbf{x} - \mathbf{p}_1)\| + \|\frac{1}{2}(\mathbf{x} - \mathbf{p}_2)\|$ and application of lemma 2.5 shows that $\mathbf{x} - \mathbf{p}_1 = \alpha(\mathbf{x} - \mathbf{p}_2)$. Taking norms (and using the fact that $\mathbf{q} \neq \mathbf{x}$) shows that $\alpha = 1$ and hence $\mathbf{p}_1 = \mathbf{p}_2$, which is a contradiction. □

In a similar manner, it is easy to show for example that the exception set of the sphere consists only of its center, or to calculate the exception sets of the sets M from figure 1. Another property of the exception set is that it behaves nicely under non-degenerate affine linear transformation.

Hence in general, we cannot expect $\mathscr{X}(M)$ to vanish altogether. However we can show that in practical applications we can easily neglect it:

Theorem 2.6 (Uniqueness)**.** $\mathrm{vol}(\mathscr{X}(M)) = 0$.

This means that the Lebesgue measure of the exception set is zero i.e. that it does not contain any open ball. In other words, if $\mathbf{x}$ is drawn from a continuous probability distribution on $\mathbb{R}^n$, then $\mathbf{x} \in \mathscr{X}(M)$ with probability 0. We simplify the proof by introducing the following lemma:

Lemma 2.7. *Let $\mathbf{x} \in \mathscr{X}(M)$ with $\mathbf{p} \lhd \mathbf{x}, \mathbf{p}' \lhd \mathbf{x}$ and $\mathbf{p} \neq \mathbf{p}'$. Assume $\mathbf{y}$ lies on the line between $\mathbf{x}$ and $\mathbf{p}$. Then $\mathbf{y} \notin \mathscr{X}(M)$.*

Proof. So $\mathbf{y} = \alpha\mathbf{x} + (1-\alpha)\mathbf{p}$ with $\alpha \in (0,1)$. Note that then also $\mathbf{p} \lhd \mathbf{y}$ — otherwise we would have another $\mathbf{q} \lhd \mathbf{y}$ with $\|\mathbf{q} - \mathbf{y}\| < \|\mathbf{p} - \mathbf{y}\|$. But then $\|\mathbf{q} - \mathbf{x}\| \leq \|\mathbf{q} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}\| < \|\mathbf{p} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}\| = \|\mathbf{p} - \mathbf{x}\|$, which contradicts the assumption.

Now assume that $\mathbf{y} \in \mathscr{X}(M)$. Then there exists $\mathbf{p}'' \lhd \mathbf{y}$ with $\mathbf{p}'' \neq \mathbf{p}$. But $\|\mathbf{p}'' - \mathbf{x}\| \leq \|\mathbf{p}'' - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}\| = \|\mathbf{p} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}\| = \|\mathbf{p} - \mathbf{x}\|$. Then $\mathbf{p} \lhd \mathbf{x}$ induces $\|\mathbf{p}'' - \mathbf{x}\| = \|\mathbf{p} - \mathbf{x}\|$. So

$$\|\mathbf{p}'' - \mathbf{x}\| = \|\mathbf{p}'' - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}\|.$$

Application of lemma 2.5 then yields $\mathbf{p}'' - \mathbf{y} = \alpha(\mathbf{y} - \mathbf{x})$, and hence $\mathbf{p}'' - \mathbf{y} = \beta(\mathbf{p} - \mathbf{y})$. Taking norms (and using $\mathbf{p} \lhd \mathbf{x}$) shows that $\beta = 1$ and hence $\mathbf{p} = \mathbf{p}''$, which is a contradiction. □

Proof of theorem 2.6. Assume there exists an open set $U \subset \mathscr{X}(M)$, and let $\mathbf{x} \in U$. Then choose $\mathbf{p} \neq \mathbf{p}' \in M$ with $\mathbf{p} \lhd \mathbf{x}, \mathbf{p}' \lhd \mathbf{x}$. But

$$\{\alpha\mathbf{x} + (1 - \alpha\mathbf{p}) | \alpha \in (0,1)\} \cap U \neq \varnothing,$$

which contradicts lemma 2.7. □

2.3 Algorithm

From here on, let M be defined by equation (2). In [2], Hoyer proposes algorithm 1 to project a given vector $\mathbf{x}$ onto $\mathbf{p} \in M$ such that $\mathbf{p} \lhd \mathbf{x}$ (we added a slight simplification by not setting *all* negative values of $\mathbf{s}$ to zero but only a single one in each step). The algorithm iteratively detects $\mathbf{p}$ by first satisfying the 1-norm condition (line 1) and then the 2-norm condition (line 3). The algorithm terminates if the constructed vector is already positive; otherwise a negative coordinate is selected, set to zero (line 4) and the search is continued in $\mathbb{R}^{n-1}$.

Algorithm 1: Sparse projection

Input: vector $\mathbf{x} \in \mathbb{R}^n$, norm conditions λ_1 and λ_2
Output: closest non-negative $\mathbf{s}$ with $\|\mathbf{s}\|_i = \lambda_i$

1 Set $\mathbf{r} \leftarrow \mathbf{x} + (\|\mathbf{x}\|_1 - \lambda_1/n)\mathbf{e}$ with $\mathbf{e} = (1, \ldots, 1)^\top \in \mathbb{R}^n$.
2 Set $\mathbf{m} \leftarrow (\lambda_1/n)\mathbf{e}$.
3 Set $\mathbf{s} \leftarrow \mathbf{m} + \alpha(\mathbf{r} - \mathbf{m})$ with $\alpha > 0$ such that $\|\mathbf{s}\|_2 = \lambda_2$.
if *exists j with $s_j < 0$* **then**
4 Fix $s_j \leftarrow 0$.
5 Remove j-th coordinate of $\mathbf{x}$.
6 Decrease dimension $n \leftarrow n - 1$.
7 **goto** 1.
end

The projection algorithm terminates after maximally $n-1$ iterations. However it is not obvious that it indeed detects $\mathbf{p}$. In the following we will prove this given that $\mathbf{x} \notin \mathscr{X}(M)$ — of course we have to exclude non-uniqueness points. The idea of the proof is to show that in each step the new estimate has $\mathbf{p}$ as closest point in M.

Theorem 2.8 (Sparse projection)**.** *Given $\mathbf{x} \geq 0$ such that $\mathbf{x} \notin \mathscr{X}(M)$. Let $\mathbf{p} \in M$ with $\mathbf{p} \lhd \mathbf{x}$. Furthermore assume that $\mathbf{r}$ and $\mathbf{s}$ are constructed by lines 1 and 3 of algorithm 1. Then:*

(i) $\sum r_i = \lambda_1$, $\mathbf{p} \lhd \mathbf{r}$ *and* $\mathbf{r} \notin \mathscr{X}(M)$.
(ii) $\sum s_i = \lambda_1$, $\|\mathbf{s}\|_2 = \lambda_2$ *and* $\mathbf{p} \lhd \mathbf{s}$ *and* $\mathbf{s} \notin \mathscr{X}(M)$.
(iii) If $s_j < 0$ then $p_j = 0$.
(iv) Define $\mathbf{u} := \mathbf{s}$ but set $u_j = 0$. Then $\mathbf{p} \lhd \mathbf{u}$ and $\mathbf{u} \notin \mathscr{X}(M)$.

This theorem shows that if $\mathbf{s} \geq 0$ then already $\mathbf{s} \in M$ and $\mathbf{p} \lhd \mathbf{s}$ *(ii)* so $\mathbf{s} = \mathbf{p}$. If $s_j < 0$ then it is enough to set $s_j := 0$ (because $p_j = 0$ *(iii)*) and continue the search in one dimension lower *(iv)*.

Proof. Let $H := \{\mathbf{x} \in \mathbb{R}^n | \sum x_i = \lambda_1\}$ denote the affine hyperplane given by the 1-norm. Note that $M \subset H$.

(i) By construction $\mathbf{r} \in H$. Furthermore $\mathbf{e} \perp H$, so $\mathbf{r}$ is the orthogonal projection of $\mathbf{x}$ onto H. Let $\mathbf{q} \in M$ be arbitrary. We then get $\|\mathbf{q} - \mathbf{x}\|^2 = \|\mathbf{q} - \mathbf{r}\|^2 + \|\mathbf{r} - \mathbf{x}\|^2$. By definition $\|\mathbf{p} - \mathbf{x}\| \leq \|\mathbf{q} - \mathbf{x}\|$, so $\|\mathbf{p} - \mathbf{r}\|^2 = \|\mathbf{p} - \mathbf{x}\|^2 - \|\mathbf{r} - \mathbf{x}\|^2 \leq \|\mathbf{q} - \mathbf{x}\|^2 - \|\mathbf{r} - \mathbf{x}\|^2 = \|\mathbf{q} - \mathbf{r}\|^2$ and therefore $\mathbf{p} \lhd \mathbf{r}$. Furthermore $\mathbf{r} \notin \mathscr{X}(M)$ because if $\mathbf{q} \in \mathbb{R}^n$ with $\mathbf{q} \lhd \mathbf{r}$, then $\|\mathbf{q} - \mathbf{r}\| = \|\mathbf{p} - \mathbf{r}\|$. Then by the above also $\|\mathbf{q} - \mathbf{x}\| = \|\mathbf{p} - \mathbf{x}\|$, hence $\mathbf{q} = \mathbf{p}$ (because $\mathbf{x} \notin \mathscr{X}(M)$).

(ii) First note that $\mathbf{s}$ is a linear combination of $\mathbf{m}$ and $\mathbf{r}$, and both lie in H so also $\mathbf{s} \in H$ i.e. $\sum s_i = \lambda_1$. Furthermore by construction $\|\mathbf{s}\| = \lambda_2$. Now let $\mathbf{q} \in M$. For $\mathbf{p} \lhd \mathbf{s}$ to hold, we have to show that $\|\mathbf{p} - \mathbf{s}\| \leq \|\mathbf{q} - \mathbf{s}\|$. This follows (see *(i)*) if we can show

$$\|\mathbf{q} - \mathbf{r}\|^2 = \|\mathbf{s} - \mathbf{r}\|^2 + \frac{1}{\alpha_0}\|\mathbf{q} - \mathbf{s}\|^2. \tag{3}$$

We can prove this equation as follows: By definition $\lambda_2^2 = \|\mathbf{q} - \mathbf{m}\|^2 = \|\mathbf{q} - \mathbf{s}\|^2 + \|\mathbf{s} - \mathbf{m}\|^2 + 2\langle \mathbf{q} - \mathbf{s}, \mathbf{s} - \mathbf{m} \rangle$, hence $\|\mathbf{q} - \mathbf{s}\|^2 = -2\langle \mathbf{q} - \mathbf{s}, \mathbf{s} - \mathbf{m} \rangle = -2\frac{\alpha_0}{\alpha_0 - 1}\langle \mathbf{q} - \mathbf{s}, \mathbf{s} - \mathbf{r} \rangle$, where we have used $\mathbf{s} - \mathbf{m} = \alpha_0(\mathbf{r} - \mathbf{m})$ i.e. $\mathbf{m} = \frac{\mathbf{s} - \alpha_0\mathbf{r}}{1 - \alpha_0}$ so $\mathbf{s} - \mathbf{m} = \frac{\alpha_0}{\alpha_0 - 1}(\mathbf{s} - \mathbf{r})$.

Using the above, we can now calculate

$$\begin{aligned}\|\mathbf{q}-\mathbf{r}\|^2 &= \|\mathbf{q}-\mathbf{s}\|^2+\|\mathbf{s}-\mathbf{r}\|^2+2\langle\mathbf{q}-\mathbf{s},\mathbf{s}-\mathbf{r}\rangle \\ &= \|\mathbf{q}-\mathbf{s}\|^2+\|\mathbf{s}-\mathbf{r}\|^2+\frac{1-\alpha_0}{\alpha_0}\|\mathbf{q}-\mathbf{s}\|^2 \\ &= \|\mathbf{s}-\mathbf{r}\|^2+\frac{1}{\alpha_0}\|\mathbf{q}-\mathbf{s}\|^2.\end{aligned}$$

Similarly, from formula 3, we get $\mathbf{s}\notin\mathscr{X}(M)$, because if there exists $\mathbf{q}\in\mathbb{R}^n$ with $\|\mathbf{q}-\mathbf{s}\|=\|\mathbf{p}-\mathbf{s}\|$, then also $\|\mathbf{q}-\mathbf{r}\|=\|\mathbf{p}-\mathbf{r}\|$ hence $\mathbf{q}=\mathbf{p}$.

(iii) Assume $s_j<0$. First note that $\mathbf{m}$ does not lie on the line $\beta\mathbf{s}+(1-\beta)\mathbf{p}$ (in other words $\mathbf{m}\neq(\mathbf{p}+\mathbf{s})/2$), because otherwise due to symmetry there would be at least two points in M closest to $\mathbf{s}$, but $\mathbf{s}\notin\mathscr{X}(M)$. Now assume the claim is wrong, then $p_j>0$ (because $\mathbf{p}\geq 0$). Define $\mathbf{g}:[0,1]\to H$ by $\mathbf{g}(\beta):=\mathbf{m}+\alpha_\beta(\beta\mathbf{s}+(1-\beta)\mathbf{p}-\mathbf{m})$, where $\alpha_\beta>0$ has been chosen such that $\|\mathbf{g}(\beta)\|=\lambda_2$. The curve $\mathbf{g}$ describes the shortest arc in $H\cap\{\|\mathbf{q}\|=\lambda_2\}$ connecting $\mathbf{p}$ to $\mathbf{s}$. We notice that $p_j>0, r_j<0$ and $\mathbf{g}$ is continuous. Hence determine the (unique) β_0 such that $\mathbf{q}:=\mathbf{g}(\beta_0)$ has the property $q_j=0$. By construction $\mathbf{q}\in M$, but $\mathbf{q}$ lies closer to $\mathbf{s}$ than $\mathbf{p}$ (because $\|\mathbf{g}(\beta-\mathbf{r})\|^2=2\langle\mathbf{g}(\beta)-\mathbf{m},\mathbf{m}-\mathbf{r}\rangle+2\lambda_2^2$ is decreasing with increasing β). But this is a contradiction to $\mathbf{p}\lhd\mathbf{s}$.

(iv) The vector $\mathbf{u}$ is defined by $u_i=s_i$ if $i\neq j$ and $u_j=0$ i.e. $\mathbf{u}$ is the orthogonal projection of $\mathbf{s}$ onto the coordinate hyperplane given by $x_j=0$. So we calculate $\|\mathbf{p}-\mathbf{s}\|^2=\|\mathbf{p}-\mathbf{u}\|^2+\|\mathbf{u}-\mathbf{s}\|^2$ and the claim follows directly as in *(i)*. □

3. MATRIX FACTORIZATION

Matrix factorization models have already been used successfully in many applications when it comes to find suitable data representations. Basically, a given $m\times T$ data matrix $\mathbf{X}$ is factorized into a $m\times n$ matrix $\mathbf{W}$ and a $n\times T$ matrix $\mathbf{H}$

$$\mathbf{X}=\mathbf{WH}, \tag{4}$$

where $m\leq n$.

3.1 Sparse non-negative matrix factorization

In contrast to other matrix factorization models such as principal or independent component analysis, *non-negative matrix factorization (NMF)* strictly requires both matrices $\mathbf{W}$ and $\mathbf{H}$ to have non-negative entries, which means that the data can be described using only additive components. Such a constraint has many physical realizations and applications, for instance in object decomposition [3].

Although NMF has recently gained popularity due to its simplicity and power in various applications, its solutions frequently fail to exhibit the desired sparse object decomposition. Therefore, Hoyer [2] proposed a modification of the NMF model to include sparseness: he minimizes the deviation of (4) under the constraint of fixed sparseness of both $\mathbf{W}$ and $\mathbf{H}$. Here, using a ratio of 1- and 2-norms of $\mathbf{x}\in\mathbb{R}^n\setminus\{0\}$, the sparseness is measured by $\sigma(\mathbf{x}):=(\sqrt{n}-\|\mathbf{x}\|_1/\|\mathbf{x}\|_2)/(\sqrt{n}-1)$. So $\sigma(\mathbf{x})=1$ (maximal) if $\mathbf{x}$ contains $n-1$ zeros, and it reaches zero if the absolute value of all coefficients of $\mathbf{x}$ coincide.

Formally, *sparse NMF (sNMF)* [2] can be defined as the task of finding

$$\mathbf{X}=\mathbf{WH}\qquad\text{subject to}\qquad\begin{cases}\mathbf{X},\mathbf{W},\mathbf{H}\geq 0\\ \sigma(\mathbf{W}_{*i})=\sigma_{\mathbf{W}}\\ \sigma(\mathbf{H}_{i*})=\sigma_{\mathbf{H}}\end{cases} \tag{5}$$

Here $\sigma_{\mathbf{W}},\sigma_{\mathbf{H}}\in[0,1]$ denote fixed constants describing the sparseness of the columns of $\mathbf{W}$ respectively the rows of $\mathbf{H}$. Usually, the linear model in NMF is assumed to hold only approximately, hence the above formulation of sNMF represents the limit case of perfect factorization. sNMF is summarized by algorithm 2, which uses algorithm 1 separately in each column respectively row for the sparse projection.

Algorithm 2: Sparse non-negative matrix factorization

Input: observation data matrix $\mathbf{X}$
Output: decomposition $\mathbf{WH}$ of $\mathbf{X}$ fulfilling given sparseness constraints σ_H and σ_W

Initialize $\mathbf{W}$ and $\mathbf{H}$ to random non-negative matrices.
Project the rows of $\mathbf{H}$ and the columns of $\mathbf{W}$ such that they meet the sparseness constraints $\sigma_{\mathbf{H}}$ and $\sigma_{\mathbf{W}}$ respectively.
repeat
  Set $\mathbf{H}\leftarrow\mathbf{H}-\mu_{\mathbf{H}}\mathbf{W}^\top(\mathbf{WH}-\mathbf{X})$.
  Project the rows of $\mathbf{H}$ such that they meet the sparseness constraint $\sigma_{\mathbf{H}}$.
  Set $\mathbf{W}\leftarrow\mathbf{W}-\mu_{\mathbf{W}}(\mathbf{WH}-\mathbf{X})\mathbf{H}^\top$.
  Project the rows of $\mathbf{W}$ such that they meet the sparseness constraint $\sigma_{\mathbf{W}}$.
until *convergence*;

3.2 Indeterminacies

Obvious indeterminacies of model 5 are permutation and positive scaling of the columns of $\mathbf{W}$ (and correspondingly of the rows of $\mathbf{H}$), because if $\mathbf{P}$ denotes a permutation matrix and $\mathbf{L}$ a positive scaling matrix, then $\mathbf{X}=\mathbf{WH}=(\mathbf{WP}^{-1}\mathbf{L}^{-1})(\mathbf{LPH})$ and the conditions of positivity and sparseness are invariant under scaling by a positive number. Another maybe not as obvious indeterminacy comes into play due to the sparseness assumption.

Definition 3.1. *The $n\times T$-matrix $\mathbf{H}$ is said to be* degenerate *if there exist $\mathbf{v}\in\mathbb{R}^n$, $\mathbf{v}>0$ and $\lambda_t\geq 0$ such that $\mathbf{H}_{*t}=\lambda_t\mathbf{v}$ for all t.*

Note that in this case all rows $\mathbf{h}_i^\top:=\mathbf{H}_{i*}$ of $\mathbf{H}$ have the same sparseness $\sigma(\mathbf{h}_i)=(\sqrt{n}-\|\lambda\|_1/\|\lambda\|_2)/(\sqrt{n}-1)$ independent of $\mathbf{v}$, where $\lambda:=(\lambda_1,\ldots,\lambda_T)^\top$. Furthermore, if $\mathbf{W}$ is any matrix with positive entries, then $\mathbf{Wv}>0$ and $\mathbf{WH}_{*t}=\lambda_t(\mathbf{Wv})$, so the signals $\mathbf{H}$ and its transformations $\mathbf{WH}$ have rows of equal sparseness. Hence if the sources are degenerate we get an indeterminacy of sNMF: Let $\mathbf{W},\tilde{\mathbf{W}}$ be non-negative such that $\tilde{\mathbf{W}}^{-1}\mathbf{Wv}>0$ (for example $\mathbf{W}>0$ arbitrary and $\tilde{\mathbf{W}}:=\mathbf{I}$), and let $\mathbf{H}$ be degenerate. Then $\tilde{\mathbf{H}}:=\tilde{\mathbf{W}}^{-1}\mathbf{WH}$ is of the same sparseness as $\mathbf{H}$ and $\mathbf{WH}=\tilde{\mathbf{W}}\tilde{\mathbf{H}}'$, but the mixing matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ do not coincide up to permutation and scaling.

3.3 Uniqueness

In this section we will discuss the uniqueness of sNMF solutions i.e. we will formulate conditions under which the set of solutions is satisfactorily small. We will see that in the perfect factorization case, it is enough to put the sparseness condition either onto $\mathbf{W}$ or $\mathbf{H}$ — we chose $\mathbf{H}$ in the following to match the picture of sources with a given sparseness.

Assume that two solutions $(\mathbf{W},\mathbf{H})$ and $(\tilde{\mathbf{W}},\tilde{\mathbf{H}})$ of the sNMF model (2) are given with $\mathbf{W}$ and $\tilde{\mathbf{W}}$ of full rank; then

$$\mathbf{WH}=\tilde{\mathbf{W}}\tilde{\mathbf{H}}, \tag{6}$$

and $\sigma(\mathbf{H})=\sigma(\tilde{\mathbf{H}})$. As before let $\mathbf{h}_i=\mathbf{H}_{i*}^\top$ respectively $\tilde{\mathbf{h}}_i=\tilde{\mathbf{H}}_{i*}^\top$ denote the rows of the source matrices. In order to avoid the scaling indeterminacy, we can set the source scales to a given value, so we may assume

$$\|\mathbf{h}_i\|_2=\|\tilde{\mathbf{h}}_i\|_2=1 \tag{7}$$

for all i. Hence, the sparseness of the rows is already fully determined by their 1-norms, and

$$\|\mathbf{h}_i\|_1=\|\tilde{\mathbf{h}}_i\|_1. \tag{8}$$

We can then show the following lemma (even without positive mixing matrices).

Lemma 3.2. *Let* $\mathbf{W}, \tilde{\mathbf{W}} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{H}, \tilde{\mathbf{H}} \in \mathbb{R}^{n\times T}$, $\mathbf{H}, \tilde{\mathbf{H}} \geq 0$, *such that equations (6–8) hold. Then for all* $i \in \{1,\dots,m\}$

(i) $\sum_j w_{ij} = \sum_j \tilde{w}_{ij}$

(ii) $\sum_{j<k} w_{ij} w_{ik} (1 - \langle \mathbf{h}_j, \mathbf{h}_k \rangle) = \sum_{j<k} \tilde{w}_{ij} \tilde{w}_{ik} (1 - \langle \tilde{\mathbf{h}}_j, \tilde{\mathbf{h}}_k \rangle)$

Proof. (i) Let $\mathbf{e} := (1,\dots,1)^\top \in \mathbb{R}^T$, and $\tau := \|\mathbf{h}_1\|_1$ the constant 1-norm of the rows. Then $\mathbf{h}_i^\top \mathbf{e} = \tau$, so application of $\mathbf{e}$ to equation (6) guarantees $\mathbf{W}(\tau,\dots,\tau)^\top = \tilde{\mathbf{W}}(\tau,\dots,\tau)^\top$ and hence *(i)*.

(ii) For readability let $\alpha_j := w_{ij}, \beta_j := \tilde{w}_{ij}$ and $\mu_{jk} := \langle \mathbf{h}_j, \mathbf{h}_k \rangle$, $\nu_{jk} = \langle \tilde{\mathbf{h}}_j, \tilde{\mathbf{h}}_k \rangle$. By equation (7) we get

$$\sum_j \alpha_j^2 + 2 \sum_{j<k} \alpha_j \alpha_k \mu_{jk} = \sum_j \beta_j^2 + 2 \sum_{j<k} \beta_j \beta_k \nu_{jk}. \tag{9}$$

Using *(i)*, the left hand side of this equation can be rewritten as $\sum_{j<n} \alpha_j^2 + (\sum_j \beta_j - \sum_{j<n} \alpha_j)^2 + 2\sum_{j<k} \alpha_j \alpha_k \mu_{jk}$, which, after some algebraic manipulations, can be seen to equal $2\sum_{j<n} \alpha_j^2 + 2\sum_{j<k<n} \alpha_j \alpha_k - 2\sum_{j<n,k} \alpha_j \beta_k + 2\sum_{j<k} \alpha_j \alpha_k \mu_{jk} + \sum_j \beta_j^2 + 2\sum_{j<k} \beta_j \beta_k$. Plugging this into equation (9) yields

$$\sum_{j=1}^n \alpha_j \left(\alpha_j + \sum_{k=j+1}^n \alpha_k - \sum_k \beta_k \right) + \sum_{j<k} \alpha_j \alpha_k \mu_{jk} = \sum_{j<k} \beta_j \beta_k (\nu_{jk} - 1).$$

Together with (8) the first sum on the left hand side can be shortened to read $\sum_{j=1}^n \alpha_j (-\sum_{k=1}^{j-1} \alpha_k - \alpha_k) = -\sum_{j<k} \alpha_j \alpha_k$, and plugging this into the above equation finishes the proof. □

This lemma can now be used to prove uniqueness of sNMF in some special cases — note that in more general settings some additional indeterminacies (specific to $n > 3$) will come into play; however to our present knowledge they are thin i.e. of measure zero and hence of no practical importance.

Theorem 3.3 (Uniqueness of sNMF)**.** *In addition to the assumptions from lemma 3.2, assume that* $\mathbf{H}$ *is non degenerate and that either*

(i) $\tilde{\mathbf{W}} = \mathbf{I}$ *and* $\mathbf{W} \geq 0$,

(ii) $\tilde{\mathbf{W}}^{-1}\mathbf{W} \geq 0$, *or*

(iii) $n = 2$.

Then $\mathbf{W} = \tilde{\mathbf{W}}\mathbf{P}$ *with a permutation matrix* $\mathbf{P}$.

Proof. (i) $\tilde{\mathbf{W}} = \mathbf{I}$, so the right hand side of lemma 3.2, *(ii)* is zero. Hence $\sum_j w_{ij} = 1$ and $\sum_{j<k} w_{ij} w_{ik} (1 - \langle \mathbf{h}_j, \mathbf{h}_k \rangle) = 0$ for all i. But $\mathbf{W} \geq 0$ and due to Schwarz's inequality, $\langle \mathbf{h}_j, \mathbf{h}_k \rangle \leq \|\mathbf{h}_j\|_2 \|\mathbf{h}_k\|_2 = 1$, so all factors in the right sum are positive. Therefore each term of the sum vanishes. But $\mathbf{H}$ is non degenerate, so $\langle \mathbf{h}_j, \mathbf{h}_k \rangle \neq 1$, because equality in the Schwarz inequality only holds if the two vectors are parallel. Hence at most one $w_{ij} \neq 0$, so $= 1$ for fixed i. But $\mathbf{W}$ is of full rank, so it $\mathbf{W}$ equals the unit matrix up to permutation.

(ii) Multiplication of (6) by $\tilde{\mathbf{W}}^{-1}$ and application of *(i)* show the claim.

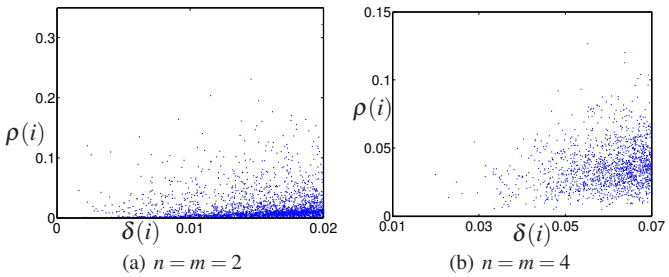
(iii) Without loss of generality we may assume $m = 2$. From (6) we get $\mathbf{VH} = \tilde{\mathbf{H}}$ with $\mathbf{V} := \tilde{\mathbf{W}}^{-1}\mathbf{W}$, and we have to show that $\mathbf{V}$ is a permutation. Similar to *(i)*, application of lemma 3.2, *(ii)* yields $v_{i1} v_{i2} (1 - \langle \mathbf{h}_1, \mathbf{h}_2 \rangle) = 0$ for $i = 1, 2$. But $\mathbf{H}$ is non degenerate and the claim follows. □

4. SIMULATIONS

In order to experimentally confirm the uniqueness result of theorem 3.3 and to show its validness in more general situations, we perform two simple simulations. Our goal is to find indeterminacies i.e. solutions to equation (6).

A fixed number of runs of the following construction is performed. In iteration i, non-negative matrices $\mathbf{W}, \mathbf{W}', \mathbf{H}$ are generated by drawing coefficients uniformly from $[0,1]$. For the second source matrix we simply set $\mathbf{H}' := \mathbf{W}'^{-1}\mathbf{WH}$ and accept the result if and only if $\mathbf{H}' \geq 0$. We then calculate the joint sparseness vector $\zeta := (\sigma(\mathbf{h}_1),\dots,\sigma(\mathbf{h}'_n))^\top$ containing the sparseness values of the rows of both $\mathbf{H}$ and $\mathbf{H}'$. The maximal difference $\delta(i) := \max_j |\zeta_j - \bar{\zeta}|$ between the sparseness of any row and the mean sparseness measures the deviation from the mixture model (6). For each iteration i, we compare this deviation with a measure $\rho(i)$ of the non-degeneracy of $\mathbf{H}$ (and hence $\mathbf{H}'$):

$$\rho(i) := \frac{2}{n(n-1)} \sum_{j<k} \left(1 - \frac{\langle \mathbf{h}_j, \mathbf{h}_k \rangle}{\|\mathbf{h}_j\|_2 \|\mathbf{h}_k\|_2} \right)^2$$



(a) $n = m = 2$ (b) $n = m = 4$

Figure 2: Simulation results in the case of two and four dimensions.

Figure 2 presents the simulation results in two cases, for $T = 10$ samples: For $n = m = 2$ (a), 2393 points out of 10^5 iterations were found with $\delta < 0.02$. In the case of $n = m = 4$ (b), 7211 solutions out of $2 \cdot 10^4$ iterations were identified with $\delta < 0.07$. In both graphs we can see that the better the mixing model is fulfilled (lower δ) the closer the two different solutions are to degeneracy (lower ρ). This confirms the claimed uniqueness of sNMF (theorem 3.3) and generalizations to higher dimensions, because when sampling for different models fulfilling the sNMF conditions, we only found degenerate solutions.

5. CONCLUSION

We have shown that the sparseness constraints in sparse NMF are almost everywhere uniquely fulfilled. Furthermore, we have been able to prove that Hoyer's projection algorithm indeed finds the closest points of fixed sparseness. Finally we have analyzed uniqueness of the sNMF model, identified a non-uniqueness condition and proved that given non-degenerate sources, uniqueness holds, at least in the case of two dimensions or some other restrictions. We have confirmed these findings and possible extensions by simulations. In later work, in addition to fully proving uniqueness, we are working on existence results and on generalizations to overcomplete situations (which we are already able to confirm experimentally).

Acknowledgements

F.T. would like to thank Patrick Hoyer for helpful discussions and Marco Hien and Peter Gruber for their ideas on the projection uniqueness and the sNMF indeterminacies respectively. Partial financial support by the DFG (GRK 638) and the BMBF (project 'ModKog') is gratefully acknowledged.

REFERENCES

[1] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Proc. NIPS 2003*. MIT Press, 2004.

[2] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[3] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 40:788–791, 1999.