Approximate Nonnegative Matrix Factorization via Alternating Minimization

Lorenzo Finesso
ISIB–CNR
Corso Stati Uniti, 4
35127 Padova – Italy
finesso@isib.cnr.it

Peter Spreij
Korteweg-de Vries Institute for Mathematics
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam – The Netherlands
spreij@science.uva.nl

February 1, 2008

Abstract

In this paper we consider the Nonnegative Matrix Factorization (NMF) problem: given an (elementwise) nonnegative matrix $V \in \mathbb{R}_+^{m\times n}$ find, for assigned k, nonnegative matrices $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{k\times n}$ such that $V = WH$. Exact, non trivial, nonnegative factorizations do not always exist, hence it is interesting to pose the approximate NMF problem. The criterion which is commonly employed is I-divergence between nonnegative matrices. The problem becomes that of finding, for assigned k, the factorization WH closest to V in I-divergence. An iterative algorithm, EM like, for the construction of the best pair (W, H) has been proposed in the literature. In this paper we interpret the algorithm as an alternating minimization procedure à la Csiszár-Tusnády and investigate some of its stability properties. NMF is widespreading as a data analysis method in applications for which the positivity constraint is relevant. There are other data analysis methods which impose some form of nonnegativity: we discuss here the connections between NMF and Archetypal Analysis. An interesting system theoretic application of NMF is to the problem of approximate realization of Hidden Markov Models.

1 Introduction

The approximate Nonnegative Matrix Factorization (NMF) of nonnegative matrices is a data analysis technique only recently introduced [6, 10]. Roughly speaking the problem is to find, for a given nonnegative matrix $V \in \mathbb{R}_+^{m \times n}$, and an assigned k, a pair of nonnegative matrices $W \in \mathbb{R}_+^{m \times k}$ and $H \in \mathbb{R}_+^{k \times n}$ such that, in an appropriate sense, $V \approx WH$. EM like algorithms for the construction of a factorization have been proposed in [6, 7]. In [10] the connection of these algorithms with the classic alternating minimization of the I-divergence [1] has been pointed out but not fully investigated. In this paper we pose the NMF problem as a minimum I-divergence problem that can be solved by alternating minimization and derive, from this point of view, the algorithm proposed in [6].

Although only recently introduced the NMF has found many applications as a data reduction procedure and has been advocated as an alternative to Principal Components Analysis (PCA) in cases where the positivity constraint is relevant (typically image analysis). The title of [10] is a clear indication of this point of view, but a complete analysis of the relations between NMF and PCA is still lacking. Other data analysis methods proposed in the literature enforce some form of positivity constraint and it is useful to investigate the connection between NMF and these methods. An interesting example is the so called Archetypal Analysis (AA) technique [2]. Assigned a matrix $X \in \mathbb{R}^{m \times n}$ and an integer k, the AA problem is to find, in the convex hull of the columns of X, a set of k vectors whose convex combinations can optimally represent X. To understand the relation between NMF and AA we choose the L_2 criterion for both problems. For any matrix A and positive definite matrix Σ define $||A||_\Sigma = (\mathrm{tr(A^T\Sigma A)})^{1/2}$. Denote $||A||_I = ||A||$. The solution of the NMF problem is then

$$(W,H) = \arg\min_{W,H} ||V - WH||$$

where the minimization is constrained to the proper set of matrices. The solution to the AA problem is given by the pair of column stochastic matrices (A, B) of respective sizes $k \times n$ and $m \times k$ such that $||X - XBA||$ is minimized (the constraint to column stochastic matrices is imposed by the convexity). Since $||X - XBA|| = ||I - BA||_{X^TX}$ the solution of the AA problem is

$$(A,B) = \arg\min_{A,B} ||I - BA||_{X^TX}.$$

AA and NMF can therefore be viewed as special cases of a more general problem which can be stated as follows. Given any matrix $P \in \mathbb{R}_+^{m \times n}$, any positive definite matrix Σ, and any integer k, find the best nonnegative factorization $P \approx Q_1Q_2$ (with $Q_1 \in \mathbb{R}_+^{m \times k}$, $Q_2 \in \mathbb{R}_+^{k \times n}$) in the L_2 sense, *i.e.*

$$(Q_1,Q_2) = \arg\min_{Q_1,Q_2} ||P - Q_1Q_2||_\Sigma.$$

Our interest in NMF stems from the system theoretic problem of approximate realization (or order reduction) of Hidden Markov Models. Partial results have already been obtained [4].

2 Preliminaries and problem statement

The NMF is a long standing problem in linear algebra [5, 9]. It can be stated as follows. Given $V \in \mathbb{R}_+^{m\times n}$, and $1 \leq k \leq \min(m,n)$, find a pair of matrices $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{k\times n}$ such that $V = WH$. The smallest k for which a factorization exists is called the positive rank of V, denoted $\mathrm{prank}(V)$. This definition implies that $\mathrm{rank}(V) \leq \mathrm{prank}(V) \leq \min(m,n)$. It is well known that $\mathrm{prank}(V)$ can assume all intermediate values, depending on V. Examples for which nonnegative factorizations do not exist, and examples for which factorization is possible only for $k > \mathrm{rank}(V)$ are easily constructed [5]. The prank has been characterized only for special classes of matrices [9] and algorithms for the construction of a NMF are not known. The approximate NMF has been recently introduced in [6] independently from the exact NMF problem. The set-up is the same, but instead of exact factorization it is required that $V \approx WH$ in an appropriate sense. In [6] and in this paper the approximation is to be understood in the sense of minimum I-divergence. For two nonnegative matrices (or vectors) $M = (M_{ij})$ and $N = (N_{ij})$ of the same size the I-divergence is defined as

$$D(M||N) = \sum_{ij}(M_{ij}\log\frac{M_{ij}}{N_{ij}} - M_{ij} + N_{ij}),$$

with the conventions $0/0 = 0$, $0\log 0 = 0$ and $p/0 = \infty$ for $p > 0$. ¿From the inequality $x\log x \geq x - 1$ it follows that $D(M||N) \geq 0$ with equality iff $M = N$. The problem of approximate NMF is to find

$$\arg\min_{W,H} D(V||WH).$$

It can be shown that, if $V_{ij} > 0$, the minimum is attained. Dropping constants the problem is equivalent to finding

$$\max_{W,H} F(W,H) := \sum_{ij}(V_{ij}\log(WH)_{ij} - (WH)_{ij}).$$

Clearly the solution is not unique. In order to rule out too many trivial multiple solutions, we impose the condition that H is row stochastic, so $\sum_j H_{lj} = 1$ for all l. This is not a restriction. Indeed, excluding without loss of generality the case where H has one or more zero rows, let h be the diagonal matrix with elements $h_i = \sum_j H_{ij}$, then $WH = \tilde{W}\tilde{H}$ with $\tilde{W} = Wh$, $\tilde{H} = h^{-1}H$ and $\tilde{H}$ is by construction row stochastic. The convention that H is row stochastic still doesn't rule out non-uniqueness. Think e.g. of post-multiplying W with a permutation matrix Π and pre-multiplying H with Π^{-1}.

Although the function F is concave in each of its arguments W and H separately, it does not have this property as a function of two variables. Hence F may have several (local) maxima, that may prevent numerical algorithms for a global maximum search to converge to the global maximizer.

Let e ($e^\top$) be a column (row) vector of appropriate dimension whose elements are all equal to one. The (constrained) problem we will look at is then

$$\max_{W,H:He=e} F(W,H). \tag{1}$$

Notice that the constrained problem (1) can be rewritten as

$$\max_{W,H:He=e} F(W,H) := \sum_{ij}(V_{ij}\log(WH)_{ij} - W_{ij}).$$

To carry out the maximization numerically [6, 7] propose an iterative algorithm. Denoting by W^n and H^n the matrices at step n, the update equations are the following

$$W_{il}^{n+1} = \sum_j V_{ij}\frac{W_{il}^n H_{lj}^n}{(W^n H^n)_{ij}} \tag{2}$$

$$H_{lj}^{n+1} = \sum_i V_{ij}\frac{W_{il}^n H_{lj}^n}{(W^n H^n)_{ij}} / \sum_i\sum_j V_{ij}\frac{W_{il}^n H_{lj}^n}{(W^n H^n)_{ij}}. \tag{3}$$

There is no rationale for this algorithm although the update steps (2) and (3) are like those in the EM algorithm, known from statistics, see [3]. Likewise the convergence properties of the algorithm are unclear. In the next section we will cast the maximization problem in a different way that provides more insight in the specific form of the update equations.

3 Lifted version of the problem

In this section we lift the I-divergence minimization problem to an equivalent minimization problem where the 'matrices' (we should speak of *tensors*) have three indices. Because we insist on probabilistic interpretations we change notations as follows. $P \in \mathbb{R}_+^{m\times n}$ is a given, fixed matrix and

$\boldsymbol{\mathcal{P}} = \{\mathbf{P} \in \mathbb{R}_+^{m\times k\times n} : \sum_l \mathbf{P}(ilj) = P(ij)\}$,

$\boldsymbol{\mathcal{Q}} = \{\mathbf{Q} \in \mathbb{R}_+^{m\times k\times n} : \mathbf{Q}(ilj) = Q_-(il)Q_+(lj),\ Q_-(il),\ Q_+(lj) \geq 0,\ Q_+e = e\}$,

$\mathcal{Q} = \{Q \in \mathbb{R}_+^{m\times n} : Q(ij) = \sum_l \mathbf{Q}(ilj) \quad \text{for some } \mathbf{Q} \in \boldsymbol{\mathcal{Q}}\}$.

Notice that $\mathcal{Q}$ is the class of $m \times n$ matrices that admit exact NMF of size k. In the notation of section 2, V has become P, and W, H are now Q_-, Q_+ respectively.

The following observation (whose proof is elementary, see [8]) motivates our approach.

Lemma 3.1 *P can be factorized as $P = Q_-Q_+$ iff $\boldsymbol{\mathcal{P}} \cap \boldsymbol{\mathcal{Q}} \neq \emptyset$, so iff there exists a $\mathbf{P} \in \boldsymbol{\mathcal{P}}$ and $\mathbf{Q} \in \boldsymbol{\mathcal{Q}}$ such that $\mathbf{P} = \mathbf{Q}$.*

For a probabilistic interpretation of this lemma, and of the results below, we assume (without loss of generality) that $\mathbf{P}$ represents the joint distribution of a three dimensional random vector. Suppose that Y_- and Y_+ are finite valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whose joint distribution is given by $\mathbb{P}(Y_- = i, Y_+ = j) = P(ij)$. Then the content of the lemma is that there exists a finite valued random variable X such that Y_- and Y_+ are conditionally independent given X iff $P = Q_-Q_+$. The matrix Q_- then gives the joint distribution of Y_- and X by $Q_-(il) = \mathbb{P}(Y_- = i, X = l)$, whereas the matrix Q_+ can be interpreted as conditional distributions of Y^+ given X via $Q_+(lj) = \mathbb{P}(Y_+ = j|X = l)$. Moreover, in this case we have $\mathbb{P}(Y_- = i, X = l, Y_+ = j) = \mathbf{Q}(ilj)$. To see this we write the conditional independence relation

$$\mathbb{P}(Y_- = i, Y_+ = j|X = l) = \mathbb{P}(Y_- = i|X = l)\mathbb{P}(Y_+ = j|X = l)$$

in equivalent form as

$$\mathbb{P}(Y_- = i, X = l, Y_+ = j) = \mathbb{P}(Y_- = i, X = l)\mathbb{P}(Y_+ = j|X = l),$$

from which the above statements immediately follow.

4 Two partial minimization problems

In this section we consider the following two minimization problems. In the first one we minimize for given $\mathbf{Q} \in \boldsymbol{\mathcal{Q}}$ the I-divergence $D(\mathbf{P}||\mathbf{Q})$ over $\mathbf{P} \in \boldsymbol{\mathcal{P}}$. In the second problem we minimize for given $\mathbf{P} \in \boldsymbol{\mathcal{P}}$ the I-divergence $D(\mathbf{P}||\mathbf{Q})$ over $\mathbf{Q} \in \boldsymbol{\mathcal{Q}}$. The unique solution $\mathbf{P}^* = \mathbf{P}^*(\mathbf{Q})$ to the first problem can be computed analytically and is given by

$$\mathbf{P}^*(ilj) = \frac{\mathbf{Q}(ilj)P(ij)}{Q(ij)}, \tag{4}$$

where $Q(ij) = \sum_l \mathbf{Q}(ilj)$. A direct computation gives the useful relation

$$D(\mathbf{P}^*(\mathbf{Q})||\mathbf{Q}) = D(P||Q).$$

The interpretation in terms of random variables is that for a given probability measure $\mathbb{Q}$, random variables Y_-, X, Y_+ with law $\mathbb{Q}(Y_- = i, X = l, Y_+ = j) = \mathbf{Q}(ilj)$, the best approximating model $\mathbb{P}^*$ with marginal distribution of $Y = (Y_-, Y_+)$ described by P is given by

$$\begin{aligned}\mathbf{P}^*(ilj) &= \mathbb{P}(Y_- = i, X = l, Y_+ = j) \\ &= \mathbb{Q}(X = l|Y_- = i, Y_+ = j)P(i, j).\end{aligned}$$

Equivalently, we can say that $\mathbb{P}^*$ is such that the marginal distribution of Y under $\mathbb{P}^*$ is given by P and the conditional distribution of X given Y under $\mathbb{P}^*$ is equal to the conditional distribution under $\mathbb{Q}$. Below we will see that this is not a coincidence.

The solution $\mathbf{Q}^* = \mathbf{Q}^*(\mathbf{P})$ to the second problem is given by

$$Q^*_-(il) = \sum_j \mathbf{P}(ilj) \tag{5}$$

$$Q^*_+(lj) = \frac{\sum_i \mathbf{P}(ilj)}{\sum_{ij} \mathbf{P}(ilj)}. \tag{6}$$

The interpretation in probabilistic terms is that for a given distribution $\mathbb{P}$ of (Y_-, X, Y_+), the best model $\mathbb{Q}^*$ that makes Y_- and Y_+ conditionally independent given X is such that

$$\mathbb{Q}^*(Y_- = i, X = l) = \mathbb{P}(Y_- = i, X = l)$$

and

$$\mathbb{Q}^*(Y_+ = j|X = l, Y_- = i) = \mathbb{Q}^*(Y_+ = j|X = l) = \mathbb{P}(Y_+ = j|X = l).$$

We see that the optimal solution $\mathbb{Q}^*$ is such that the marginal distributions of (X, Y_-) under $\mathbb{P}$ and $\mathbb{Q}^*$ coincide and that the same happens for the conditional distributions of Y_+ given X. Again, this is not a coincidence, as we will explain below. First we will state for the two partial minimization problems above the following two Pythagorean rules.

Lemma 4.1 *For fixed* $\mathbf{P}$ *and* $\mathbf{Q}^* = \mathbf{Q}^*(\mathbf{P})$ *it holds that for any* $\mathbf{Q} \in \boldsymbol{\mathcal{Q}}$

$$D(\mathbf{P}||\mathbf{Q}) = D(\mathbf{P}||\mathbf{Q}^*) + D(\mathbf{Q}^*||\mathbf{Q}), \tag{7}$$

whereas for fixed $\mathbf{Q}$ *and* $\mathbf{P}^* = \mathbf{P}^*(\mathbf{Q})$ *it holds that for any* $\mathbf{P} \in \boldsymbol{\mathcal{P}}$

$$D(\mathbf{P}||\mathbf{Q}) = D(\mathbf{P}||\mathbf{P}^*) + D(\mathbf{P}^*||\mathbf{Q}), \tag{8}$$

and

$$D(\mathbf{P}^*||\mathbf{Q}) = D(P||Q), \tag{9}$$

where Q *is given by* $Q(ij) = \sum_l \mathbf{Q}(ilj)$.

Proof. To prove the first relation we first introduce some notation. Let $\mathbf{P}(il\cdot) = \sum_j \mathbf{P}(ilj)$, $\mathbf{P}(\cdot lj) = \sum_i \mathbf{P}(ilj)$ and $\mathbf{P}(j|l) = \mathbf{P}(\cdot lj)/\sum_j \mathbf{P}(\cdot lj)$. For $\mathbf{Q}$ we use similar notation and so we have $\mathbf{Q}(il\cdot) = Q_-(il)$ and $\mathbf{Q}(j|l) = Q_+(lj)/\sum_j Q_+(lj)$ and $Q^*_-(il) = \mathbf{P}(il\cdot)$ and $Q^*_+(lj) = \mathbf{P}(j|l)$. Consider

$$\begin{aligned} D(\mathbf{P}||\mathbf{Q}) - D(\mathbf{P}||\mathbf{Q}^*) &= \sum_{ilj} \mathbf{P}(ilj) \log \frac{\mathbf{P}(il\cdot)}{Q_-(ij)} + \log \frac{\mathbf{P}(j|l)}{Q_+(lj)} \\ &= \sum_{il} \mathbf{P}(il\cdot) \log \frac{\mathbf{P}(il\cdot)}{Q_-(ij)} + \sum_{lj} \mathbf{P}(\cdot lj) \log \frac{\mathbf{P}(j|l)}{Q_+(lj)}. \end{aligned}$$

On the other hand we have

$$
\begin{aligned}
D(\mathbf{Q}^*||\mathbf{Q}) &= \sum_{ilj} \mathbf{P}(il\cdot)\mathbf{P}(j|l)(\log \frac{\mathbf{P}(il\cdot)}{Q_-(il)} + \log \frac{\mathbf{P}(j|l)}{Q_+(lj)} \\
&= \sum_{il} \mathbf{P}(il\cdot) \log \frac{\mathbf{P}(il\cdot)}{Q_-(il)} + \sum_{lj} \mathbf{P}(\cdot lj) \log \frac{\mathbf{P}(j|l)}{Q_+(lj)}.
\end{aligned}
$$

The first assertion follows. The second Pythagorean rule follows from

$$
\begin{aligned}
&D(\mathbf{P}||\mathbf{P}^*) + D(\mathbf{P}^*||\mathbf{Q}) \\
= &\sum_{ilj} \mathbf{P}(ilj) \log \frac{\mathbf{P}(ilj)Q(ij)}{\mathbf{Q}(ilj)\mathbf{P}(ij)} + \sum_{ilj} \mathbf{Q}(ilj) \frac{P(ij)}{Q(ij)} \log \frac{P(ij)}{Q(ij)} \\
= &\sum_{ilj} \mathbf{P}(ilj) \log \frac{\mathbf{P}(ilj)}{\mathbf{Q}(ilj)} + \sum_{ilj} \mathbf{P}(ilj) \log \frac{Q(ij)}{P(ij)} \\
&+ \sum_{ij} Q(ij) \frac{P(ij)}{Q(ij)} \log \frac{P(ij)}{Q(ij)} \\
= &D(\mathbf{P}||\mathbf{Q}).
\end{aligned}
$$

□

For a probabilistic interpretation of the $\mathbf{P}^*$ and $\mathbf{Q}^*$ above as well as the Pythagorean rules we use a general result on the I-divergence between two joint laws of a random vector (U, V). We denote the law of this vector under probability measures $\mathbb{P}$ and $\mathbb{Q}$ by $P^{U,V}$ and $Q^{U,V}$. The conditional distributions of U given V are summarized by the matrices $P^{U|V}$ and $Q^{U|V}$, with the obvious convention $P^{U|V}(ij) = \mathbb{P}(U = i|V = j)$ and likewise for $Q^{U|V}$.

Lemma 4.2 *It holds that*

$$
D(P^{U,V}||Q^{U,V}) = \mathbb{E}_{\mathbb{P}} D(P^{U|V}||Q^{U|V}) + D(P^V||Q^V), \tag{10}
$$

where

$$
D(P^{U|V}||Q^{U|V}) = \sum_i P(U = i|V) \log \frac{P(U = i|V)}{Q(U = i|V)}.
$$

Proof. This follows from elementary manipulations. □

The above relation can be refined as follows. Suppose that V is bivariate, $V = (V_1, V_2)$ say and that U and V_2 are conditionally independent given V_1 under $\mathbb{Q}$, so the conditional distribution of U given V is the same as the conditional distribution of U given V_1 under $\mathbb{Q}$. Then the first term on the right hand side of equation (10) can be decomposed as

$$
\mathbb{E}_{\mathbb{P}} D(P^{U|V}||Q^{U|V}) = \mathbb{E}_{\mathbb{P}} D(P^{U|V}||P^{U|V_1}) + \mathbb{E}_{\mathbb{P}} D(P^{U|V_1}||Q^{U|V_1}). \tag{11}
$$

We apply this lemma to the first partial minimization problem above by an appropriate choice of U and V. Since $D(\mathbf{P}^*||\mathbf{Q}) = D(P||Q) = D(P^Y||Q^Y)$,

where P^Y is given by P, we see that for $U = X$, $V = Y = (Y_-, Y_+)$ the decomposition (8) can alternatively be written as $\mathbb{E}_{\mathbb{P}} D(P^{X|Y}||Q^{X|Y}) + D(P||Q)$. Minimizing $D(\mathbf{P}||\mathbf{Q})$ w.r.t. $\mathbf{P}$ under the condition that the marginal of $\mathbf{P}$ is given by P is thus equivalent to minimizing the I-divergence between the conditional distributions $P^{X|Y}$ and $Q^{X|Y}$, and this clearly happens for $P^{X|Y} = Q^{X|Y}$.
The interpretation of (7) is less straightforward. However, refining (7), we have parallel to (11)

$$\begin{aligned}D(\mathbf{P}||\mathbf{Q}) =& \mathbb{E}_{\mathbb{P}} D(P^{Y_+|X,Y_-}||P^{Y_+|X}) \\ &+ D(P^{Y_-,X}||Q^{Y_-,X}) + \mathbb{E}_{\mathbb{P}} D(P^{Y_+|X}||Q^{Y_+|X}).\end{aligned}$$

Hence the minimization problem here is to minimize the I-divergence between the distributions of (Y_-, X) under $\mathbb{P}$ and $\mathbb{Q}$ and the I-divergence between the conditional probability measures $P^{Y_+|X}$ and $Q^{Y+|X}$. This explains the form of the optimal solution $\mathbf{Q}^*(\mathbf{P})$.

The next proposition shows that the original minimization of $D(P||Q)$ over nonnegative matrices Q for a given nonnegative matrix P is equivalent to a double minimization over the sets $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{Q}}$.

Proposition 4.3 *Let P be given. It holds that*

$$\min_{Q \in \mathcal{Q}} D(P||Q) = \min_{\mathbf{P} \in \boldsymbol{\mathcal{P}}, \mathbf{Q} \in \boldsymbol{\mathcal{Q}}} D(\mathbf{P}||\mathbf{Q}).$$

Proof. With $\mathbf{P}^* = \mathbf{P}^*(Q)$, the optimal solution of the partial minimization over $\boldsymbol{\mathcal{P}}$, we have

$$\begin{aligned}D(\mathbf{P}||\mathbf{Q}) &\geq D(\mathbf{P}^*||\mathbf{Q}) \\ &= D(P||Q) \\ &\geq \min_{Q \in \mathcal{Q}} D(P||Q).\end{aligned}$$

It follows that $\min_{\mathbf{P} \in \boldsymbol{\mathcal{P}}, \mathbf{Q} \in \boldsymbol{\mathcal{Q}}} D(\mathbf{P}||\mathbf{Q}) \geq \min_{Q \in \mathcal{Q}} D(P||Q)$.
Conversely, let $Q^* \in \mathcal{Q}$ be the minimizer of $D(P||Q)$ and let $\mathbf{Q}$ be a corresponding element in $\boldsymbol{\mathcal{Q}}$. Furthermore, let $\mathbf{P} \in \boldsymbol{\mathcal{P}}$ be arbitrary. Then we have

$$\begin{aligned}D(P||Q^*) &\geq D(\mathbf{P}^*(\mathbf{Q})||\mathbf{Q}) \\ &\geq \min_{\mathbf{P} \in \boldsymbol{\mathcal{P}}, \mathbf{Q} \in \boldsymbol{\mathcal{Q}}} D(\mathbf{P}||\mathbf{Q}),\end{aligned}$$

which shows the other inequality. □

5 Alternating minimization algorithm

The results of the previous section are aimed at setting up an alternating minimization algorithm for obtaining $\min_Q D(P||Q)$, where P is a given nonnegative matrix. In view of proposition 4.3 we can lift this problem to the $(\boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{Q}})$ space.

Starting with an arbitrary $\mathbf{Q}_1 \in \boldsymbol{\mathcal{Q}}$ with strictly positive elements, we adopt the following recursive scheme

$$\mathbf{Q}_n \to \mathbf{P}_n \to \mathbf{Q}_{n+1} \to \mathbf{P}_{n+1}, \tag{12}$$

where $\mathbf{P}_n = \mathbf{P}^*(\mathbf{Q}_n)$, $\mathbf{Q}_{n+1} = \mathbf{Q}^*(\mathbf{P}_n)$ and $\mathbf{P}_{n+1} = \mathbf{P}^*(\mathbf{Q}_{n+1})$. The two Pythagorean rules from lemma 4.1 now take the forms

$$\begin{aligned} D(\mathbf{P}_n||\mathbf{Q}_{n+1}) &= D(\mathbf{P}_n||\mathbf{P}_{n+1}) + D(\mathbf{P}_{n+1}||\mathbf{Q}_{n+1}) \\ D(\mathbf{P}_n||\mathbf{Q}_n) &= D(\mathbf{P}_n||\mathbf{Q}_{n+1}) + D(\mathbf{Q}_{n+1}||\mathbf{Q}_n). \end{aligned}$$

Addition of these two equations results in

$$D(\mathbf{P}_n||\mathbf{Q}_n) = D(\mathbf{P}_n||\mathbf{P}_{n+1}) + D(\mathbf{P}_{n+1}||\mathbf{Q}_{n+1}) + D(\mathbf{Q}_{n+1}||\mathbf{Q}_n),$$

and together with (9) this becomes

$$D(P||Q_n) = D(\mathbf{P}_n||\mathbf{P}_{n+1}) + D(P||Q_{n+1}) + D(\mathbf{Q}_{n+1}||\mathbf{Q}_n). \tag{13}$$

This equation also shows that $D(P||Q_n) \geq D(P||Q_{n+1})$. The procedure outlined in equation (12) will be made explicit, using equations (4), (6) and (5). Since it is our aim to apply the above results to the problem as sketched in section 2, we now turn back to the notation of that section. So, instead of Q_- we write W, instead of Q_+ we write H, instead of Q we write WH, of course these will be endowed with superscript indices n and $n+1$ below, and P becomes V again. From (12) we get $\mathbf{Q}_{n+1} = \mathbf{Q}^*(\mathbf{P}^*(\mathbf{Q}_n))$ and combining this with the substitution of (4) into (5) we obtain–in the original notation–

$$W_{il}^{n+1} = \sum_j \frac{W_{il}^n H_{lj}^n V_{ij}}{(W^n H^n)_{ij}},$$

which is just (2). Of course (3) can be derived similarly.

6 Discussion of the algorithm

In the previous section we have shown that the update rules (2) and (3) are the result of an alternating minimization procedure. The convergence properties of the algorithm can be studied using the general results of [1]. Due to the similarity with the EM algorithm one may expect similar convergence properties, see [11].

At each iteration the I-divergence between V and the $W^n H^n$ is reduced, equivalently the sequence $F(W^n, H^n)$ is increasing. This follows from equation (13). Secondly, once the algorithm reaches a stationary (W, H)-point of F (the partial derivatives vanish here), the updated values are exactly equal to the given values. This can be immediately seen by computing the fist order necessary conditions for a stationary point and comparing these to the update formulas.

Moreover, as long as the algorithm does not reach a stationary point there will always be a strict increase in the objective function F. In the third place, all the W^n and H^n evolve in a compact set. For the H^n this is trivial, since they are nonnegative row stochastic matrices. For the W^n this follows from (2), since $W_{il}^{n+1} \leq \sum_j V_{ij}$ (starting the algorithm with matrices that have strictly positive elements ensures that all W^n and H^n have strictly positive elements). A detailed account of the properties of the algorithm is deferred to another publication.

References

[1] I. Csiszár and G. Tusnády (1984), Information geometry and alternating minimization procedures, *Statistics & Decisons, supplement issue* **1**, 205–237.

[2] A. Cutler and L. Breiman (1994), Archetypal analysis, *Technometrics* **36**, 338–347.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm. With discussion. *J. Roy. Statist. Soc. Ser. B* **39** no. 1, 1–38.

[4] L. Finesso and P.J.C. Spreij (2002), Approximate realization of finite Hidden Markov Chains, *Proceedings of the 2002 IEEE Information Theory Workshop* Bangalore, India.

[5] M. Hazewinkel (1984), On positive vectors, positive matrices and the specialization order, *CWI report PM-R8407*.

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

[7] D.D. Lee and H.S. Sebastian Seung (2001), Algorithms for non-negative matrix factorization. (working paper).

[8] G. Picci and J.H. van Schuppen (1984), On the weak finite stochastic realization problem, Springer LNCIS, vol. 58, 237–242.

[9] G. Picci, J.M. van den Hof, J.H. van Schuppen (1998), Primes in several classes of the positive matrices, *Linear Algebra Appl.* **277** 149–185

[10] J.A. O'Sullivan (2000), Properties of the information value decomposition, *Proceedings ISIT 2000, Sorrento, Italy*, 491.

[11] C.J. Wu (1983), On the convergence properties of the EM algorithm, *Ann. Stat.*, vol. **11**, No. 1, 95–103.