Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Linear Algebra and its Applications 416 (2006) 270–287
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Nonnegative matrix factorization and I-divergence alternating minimization ☆

Lorenzo Finesso [a], Peter Spreij [b,*]

[a] *ISIB-CNR, Corso Stati Uniti, 4, 35127 Padova, Italy*
[b] *Korteweg-de Vries, Institute for Mathematics, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands*

Received 21 December 2004; accepted 23 November 2005
Available online 10 January 2006
Submitted by M. Neumann

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, nontrivial, 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.

Keywords: Nonnegative matrix factorization; Approximate factorization; I-divergence; Alternating minimization; Lifting; Auxiliary function; Convergence

☆ The authors have been supported in part by the European Community's Human Potential Programme under contract HPRN-CT-2000-00100, DYNSTOCH.
* Corresponding author. Tel.: +31 20 525 6070; fax: +31 20 525 5101.
E-mail addresses: finesso@isib.cnr.it (L. Finesso), spreij@science.uva.nl (P. Spreij).

0024-3795/$ - see front matter
doi:10.1016/j.laa.2005.11.012

1. Introduction

The approximate Nonnegative Matrix Factorization (NMF) of nonnegative matrices is a data analysis technique only recently introduced [9,14]. 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, $V \approx WH$. In [9] EM like algorithms for the construction of a factorization have been proposed. The algorithms have been later derived in [10] by using an *ad-hoc* auxiliary function, a common approach in deriving EM algorithms. In [14] the connection with the classic alternating minimization of the I-divergence [2] 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 [9]. There are alternative approaches to approximate nonnegative matrix factorization. For instance, recently, see [3], results have been obtained for the approximate factorization (w.r.t. the Frobenius norm) of *symmetric* nonnegative matrices.

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 [14] is a clear indication of this point of view, but a complete analysis of the relations between NMF and PCA is still lacking. 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 [6].

This paper is organized as follows. In Section 2 we pose the approximate nonnegative matrix factorization problem, define the I-divergence between matrices and discuss the solution proposed in [9,10]. In Section 3 we pave the way for the alternating minimization algorithm presenting the properly lifted version of the minimization problem and solving the two partial minimizations in the style of Csiszár and Tusnády [2]. In Section 4 we construct the alternating minimization algorithm and compute the iteration gain. One of the advantages of working with the lifted problem is that it sheds a new light also on the derivation of the algorithm via auxiliary functions given in [10]. In Section 5 we will use the results of Section 3 to construct a very natural auxiliary function to solve the original problem. A discussion of the convergence properties of the algorithm is given in Section 6. In the concluding Section 7 we establish a connection between the approximate NMF problem and the Archetypal Analysis algorithm of Cutler and Breiman [4]. The present paper is an extended version of [7].

2. Preliminaries and problem statement

The NMF is a long standing problem in linear algebra [8,12]. It can be stated as follows. Given $V \in \mathbb{R}_+^{m\times n}$, and $1 \leqslant k \leqslant \min\{m, n\}$, 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 $\mathrm{rank}(V) \leqslant \mathrm{prank}(V) \leqslant \min\{m, n\}$. 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 $k > \mathrm{rank}(V)$ have been constructed in the literature [8]. The prank has been characterized only for special classes of matrices [12] and algorithms for the construction of a NMF of a general positive matrix are not known.

The *approximate* NMF has been recently introduced in [9] 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 [9], and in this paper, the approximation is to be understood in the sense of minimum I-divergence. For two nonnegative numbers p and q the I-divergence is defined as

$$D(p\|q) = p\log\frac{p}{q} - p + q,$$

with the conventions $0/0 = 0, 0\log 0 = 0$ and $p/0 = \infty$ for $p > 0$. From the inequality $x\log x \geqslant x - 1$ it follows that $D(p\|q) \geqslant 0$ with equality iff $p = q$. For two nonnegative matrices $M = (M_{ij})$ and $N = (N_{ij})$, of the same size, the I-divergence is defined as

$$D(M\|N) = \sum_{ij} D(M_{ij}\|N_{ij}).$$

Again it follows that $D(M\|N) \geqslant 0$ with equality iff $M = N$. For nonnegative vectors or tensors of the same size a similar definition applies.

The problem of approximate NMF is to find for given V and a fixed number k (often referred to as the *inner size* of the factorization)

$$\arg\min_{W,H} D(V\|WH). \tag{1}$$

The function $D : (W, H) \to D(V\|WH)$ will sometimes be referred to as the *objective function*. The *domain* of D is the set of pairs (W, H) with nonnegative entries. The *interior* of the domain is the subset of pairs (W, H) with positive (> 0) entries, whereas pairs on the *boundary* have at least one entry equal to zero.

Although the objective function $(W, H) \mapsto D(V\|WH)$ is easily seen to be convex in W and H separately, it is not jointly convex in the two variables. Hence $(W, H) \mapsto D(V\|WH)$ may have several (local) minima and saddle points, that may prevent numerical minimization algorithms to converge to the global minimizer. However $D(V\|WH)$ cannot have a local maximum in an interior point (W_0, H_0), because then also $W \mapsto D(V\|WH_0)$ would have a local maximum in W_0, which contradicts convexity. Local maxima at the boundary are not a priori excluded.

It is not immediately obvious that the approximate NMF problem admits a solution. The following result is therefore relevant.

Proposition 2.1. *The minimization problem* (1) *has a solution.*

The proof of this proposition is deferred to Section 4.

Notice that, increasing the inner size from k to $k + 1$, the optimal value of the objective function decreases. This follows from the fact that one can trivially embed the factorization problem with inner size k into the problem with inner size $k + 1$ simply adding a zero last column to the optimal W and an arbitrary last row to the optimal H of the problem with inner size k. Unfortunately, unlike the SVD of a matrix, the best approximations with increasing k are not embedded one into another. For increasing k the computations are to be carried out anew.

Although, according to Proposition 2.1, a solution to the minimization problem exists, it will certainly not be 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, first we exclude without loss of generality the case where H has one or more zero rows, since we would then in fact try to minimize the I-divergence with inner size smaller than k. Let h be the diagonal matrix with elements $h_i = \sum_j H_{ij}$, then $WH = \widetilde{W}\widetilde{H}$ with $\widetilde{W} = Wh$, $\widetilde{H} = h^{-1}H$ and $\widetilde{H}$ is by construction row stochastic. The convention that H is row stochastic still does not rule out non-uniqueness. Think e.g. of post-multiplying W with a permutation matrix Π and pre-multiplying H with Π^{-1}.

Let e_n ($e_n^\top$) be the column (row) vector of size n whose elements are all equal to one. Given k, the (constrained) problem we will look at from now on is

$$\min_{W,H:He_m=e_k} D(V\|WH). \tag{2}$$

For the sake of brevity we will often write e for a vector of 1's of generic size. The constraint in the previous problem will then read as $He = e$.

To carry out the minimization numerically, Lee and Seung [9,10] proposed the following iterative algorithm. Denoting by W^t and H^t the matrices at step t, the update equations are

$$W_{il}^{t+1} = W_{il}^t \sum_j \frac{H_{lj}^t V_{ij}}{(W^t H^t)_{ij}}, \tag{3}$$

$$H_{lj}^{t+1} = H_{lj}^t \sum_i \frac{W_{il}^t V_{ij}}{(W^t H^t)_{ij}} \Bigg/ \sum_{ij} \frac{W_{il}^t H_{lj}^t V_{ij}}{(W^t H^t)_{ij}}. \tag{4}$$

The initial condition (W^0, H^0) will always be assumed to be in the interior of the domain. Only a partial justification for this algorithm is given in [10], although the update steps (3) and (4) are like those in the EM algorithm, known from statistics, see [5]. Likewise the convergence properties of the algorithm are unclear. In the next section the minimization problem will be cast in a different way to provide more insight in the specific form of the update equations and on the convergence properties of the algorithm.

We will now show that the V matrix in the approximate NMF problem can always be taken as a probability matrix P i.e. such that $P_{ij} \geqslant 0$, $\sum_{ij} P_{ij} = 1$. This will pave the way for the probabilistic interpretation of the exact and approximate NMF problems to be given later.

Let $P = \frac{1}{e^\top V e} V$, $Q_- = \frac{1}{e^\top W e} W$, $w = e^\top W e$ and $Q_+ = H$. Notice that $e^\top P e = e^\top Q_- e = 1$ and $Q_+ e = e$. Using the definition of divergence and elementary computations, we obtain the decomposition

$$D(V\|WH) = e^\top V e D(P\|Q_- Q_+) + D(e^\top V e\|w).$$

Hence, since the number $e^\top V e$ is known, minimizing $D(V\|WH)$ w.r.t. (W, H) is equivalent to minimizing $D(P\|Q_- Q_+)$ w.r.t. (Q_-, Q_+) and $D(e^\top V e\|w)$ w.r.t. w. The minimizers of the three problems satisfy the relations $W^* = e^\top V e Q_-^*$, $H^* = Q_+^*$, and $w^* = e^\top V e$. Minimizing $D(V\|WH)$ is therefore equivalent to minimizing $D(P\|Q_- Q_+)$. This enables us to give the problem a probabilistic interpretation. Indeed,

$$D(P\|Q_- Q_+) = \sum_{ij} D(P_{ij}\|(Q_- Q_+)_{ij}) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{(Q_- Q_+)_{ij}}, \tag{5}$$

which is the usual I-divergence (Kullback–Leibler distance) between (finite) probability measures. This will be exploited in later sections. From now on we will always consider the following problem. Given the probability matrix P and the integer k find

$$\min_{Q_-, Q_+ : Q_+ e = e} D(P\|Q_- Q_+).$$

For typographical reasons we often, but not always, denote the entries of P by $P(ij)$ instead of P_{ij} and likewise for other matrices.

The minimization algorithm is easily seen to be *invariant under the previous normalizations*. Let $Q_-^t = \frac{W^t}{e^\top W^t e}$ and $Q_-^t = H^t$. Substitute the definitions of (P, Q_-^t, Q_+^t) into (3) and (4) and

use the easily verified fact that $e^\top W^t e = e^\top V e$ for $t \geqslant 1$ to obtain the update equations in the new notations

$$Q_-^{t+1}(il) = Q_-^t(il) \sum_j \frac{Q_+^t(lj)P(ij)}{(Q_-^t Q_+^t)(ij)}, \tag{6}$$

$$Q_+^{t+1}(lj) = Q_+^t(lj) \sum_i \frac{Q_-^t(il)P(ij)}{(Q_-^t Q_+^t)(ij)} \Bigg/ \sum_{ij} \frac{Q_-^t(il)Q_+^t(lj)P(ij)}{(Q_-^t Q_+^t)(ij)}. \tag{7}$$

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.

3.1. Setup

Let be given a probability matrix P (i.e. $P(ij) \geqslant 0, \sum_{ij} P(ij) = 1$) and an integer $k \leqslant \min\{m, n\}$. We introduce the following sets:

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

$$\mathscr{Q} = \Big\{ \mathbf{Q} \in \mathbb{R}_+^{m\times k\times n} : \mathbf{Q}(ilj) = Q_-(il)Q_+(lj), \\ Q_-, Q_+ \geqslant 0, Q_+e = e, e^\top Q_- e = 1 \Big\},$$

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

The interpretation of the sets $\mathscr{P}$, $\mathscr{Q}$, $\mathscr{Q}$ is given next.

Suppose one is given random variables (Y_-, X, Y_+), taking values in $\{1, \ldots, m\} \times \{1, \ldots, k\} \times \{1, \ldots, n\}$. For convenience we can think of the r.v.'s as defined on the canonical measurable space $(\Omega, \mathscr{F})$, where Ω is the set of all triples (i, l, j) and $\mathscr{F}$ is 2^Ω. For $\omega = (i, l, j)$ we have the identity mapping $(Y_-, X, Y_+)(\omega) = (i, l, j)$. If $\mathbb{R}$ a given probability measure on this space, then the distribution of the triple (Y_-, X, Y_+) under $\mathbb{R}$ is given by the *tensor* $\mathbf{R}$ defined by

$$\mathbf{R}(ilj) = \mathbb{R}(Y_- = i, X = l, Y_+ = j). \tag{8}$$

Conversely, a given tensor $\mathbf{R}$ defines a probability measure $\mathbb{R}$ on $(\Omega, \mathscr{F})$. We will use the notation D both for I-divergence between tensors and matrices and for the Kullback–Leibler divergence between probabilities. If $\mathbf{P}$, $\mathbf{Q}$ are tensors related to probability measures $\mathbb{P}$ and $\mathbb{Q}$ like in (8) we obviously have $D(\mathbf{P}\|\mathbf{Q}) = D(\mathbb{P}\|\mathbb{Q})$.

The sets $\mathscr{P}$, $\mathscr{Q}$ correspond to subsets of the set of all measures on $(\Omega, \mathscr{F})$. In particular $\mathscr{P}$ corresponds to the subset of all measures whose $Y = (Y_-, Y_+)$ marginal coincides with the given P, while $\mathscr{Q}$ corresponds to the subset of measures under which Y_- and Y_+ are conditionally independent given X. The first assertion is evident by the definition of $\mathscr{P}$. To prove the second assertion notice that if $\mathbb{Q}(Y_- = i, X = l, Y_+ = j) = \mathbf{Q}(ilj) = Q_-(il)Q_+(lj)$, then summing over j one gets $\mathbb{Q}(Y_- = i, X = l) = Q_-(il)$ (since $Q_+e = e$) and similarly $\mathbb{Q}(Y_+ = j|X =$

$l) = Q_+(lj)$. It follows that $\mathbb{Q}(Y_- = i, X = l, Y_+ = j) = \mathbb{Q}(Y_- = i, X = l)\mathbb{Q}(Y_+ = j|X = l)$ which is equivalent to

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

i.e., Y_-, Y_+ are conditionally independent given X.

Finally the set $\mathscr{Q}$ is best interpreted algebraically as the set of $m \times n$ probability matrices that admit exact NMF of size k.

The following observation (taken from [11]) motivates our approach.

Lemma 3.1. *P admits exact factorization of inner size k iff $\mathscr{P} \cap \mathscr{Q} \neq \emptyset$.*

Proof. If $\mathscr{P} \cap \mathscr{Q} \neq \emptyset$ then there exists a matrix $\mathbf{Q} \in \mathscr{Q}$ which also belongs to $\mathscr{P}$, therefore $P = Q_- Q_+$. Conversely, if we have $P = Q_- Q_+$ with inner size k, then the tensor $\mathbf{P}$ given by $\mathbf{P}(ilj) = Q_-(il)Q_+(lj)$ clearly belongs to $\mathscr{P}$. As in Section 2 we can w.l.o.g. assume that $Q_+ e = e$, so that $\mathbf{P}$ belongs to $\mathscr{Q}$ as well. □

We are now ready to give a natural probabilistic interpretation to the exact NMF problem. The probability matrix P admits exact NMF $P = Q_- Q_+$ iff there exists at least one measure on $(\Omega, \mathscr{F})$ whose $Y = (Y_-, Y_+)$ marginal is P and at the same time making Y_- and Y_+ conditionally independent given X.

Having shown that the exact NMF factorization $P = Q_- Q_+$ is equivalent to $\mathscr{P} \cap \mathscr{Q} \neq \emptyset$ it is not surprising that the approximate NMF, corresponding to $\mathscr{P} \cap \mathscr{Q} = \emptyset$, can be viewed as a double minimization over the sets $\mathscr{P}$ and $\mathscr{Q}$.

Proposition 3.2. *Let P be given. The function $(\mathbf{P}, \mathbf{Q}) \mapsto D(\mathbf{P}\|\mathbf{Q})$ attains a minimum on $\mathscr{P} \times \mathscr{Q}$ and it holds that*

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

The proof will be given in Section 3.2.

Remark 3.3. Let $\mathbf{P}^*$ and $\mathbf{Q}^*$ be the minimizing elements in Proposition 3.2. If there is l_0 such that $\sum_{ij} \mathbf{P}^*(il_0j) = 0$, then all $\mathbf{Q}^*(il_0j)$ are zero as well. Similarly, if there is l_0 such that $\sum_{ij} \mathbf{Q}^*(il_0j) = 0$, then all $\mathbf{P}^*(il_0j)$ are zero as well. In each (and hence both) of these cases the optimal approximate factorization $Q_-^* Q_+^*$ of P is of inner size less than k (delete the column corresponding to l_0 from Q_-^* and the corresponding row of Q_+^*).

3.2. Two partial minimization problems

In the next section we will construct the algorithm for the solution of the double minimization problem

$$\min_{\mathbf{P} \in \mathscr{P}, \mathbf{Q} \in \mathscr{Q}} D(\mathbf{P}\|\mathbf{Q}),$$

of Proposition 3.2, as an alternating minimization algorithm over the two sets $\mathscr{P}$ and $\mathscr{Q}$. This motivates us to consider here two partial minimization problems. In the first one, given $\mathbf{Q} \in \mathscr{Q}$

we minimize the I-divergence $D(\mathbf{P}\|\mathbf{Q})$ over $\mathbf{P} \in \mathscr{P}$. In the second problem, given $\mathbf{P} \in \mathscr{P}$ we minimize the I-divergence $D(\mathbf{P}\|\mathbf{Q})$ over $\mathbf{Q} \in \mathscr{Q}$.

Let us start with the first problem. The unique solution $\mathbf{P}^* = \mathbf{P}^*(\mathbf{Q})$ can easily be computed analytically and is given by

$$\mathbf{P}^*(ilj) = \frac{\mathbf{Q}(ilj)}{Q(ij)} P(ij), \tag{9}$$

where $Q(ij) = \sum_l \mathbf{Q}(ilj)$. We also adopt the convention to put $\mathbf{P}^*(ilj) = 0$ if $Q(ij) = 0$, which ensures that, viewed as measures, $\mathbf{P}^* \ll \mathbf{Q}$.

Now we turn to the second partial minimization problem. The unique solution $\mathbf{Q}^* = \mathbf{Q}^*(\mathbf{P})$ to this problem can also be easily computed analytically and is given by

$$Q_-^*(il) = \sum_j \mathbf{P}(ilj), \tag{10}$$

$$Q_+^*(lj) = \frac{\sum_i \mathbf{P}(ilj)}{\sum_{ij} \mathbf{P}(ilj)}, \tag{11}$$

where we assign arbitrary values to the $Q_+^*(lj)$ (complying with the constraint $Q_+ e = e$) for those l with $\sum_{ij} \mathbf{P}(ilj) = 0$.

The two partial minimization problems and their solutions have a nice probabilistic interpretation.

In the first minimization problem, one is given a distribution $\mathbf{Q}$, which makes the pair $Y = (Y_-, Y_+)$ conditionally independent given X, and finds the best approximation to it in the set $\mathscr{P}$ of distributions with the marginal of Y given by P. Let $\mathbf{P}^*$ denote the optimal distribution of (Y_-, X, Y_+). Eq. (9) can then be interpreted, in terms of the corresponding measures, as

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

Notice that the conditional distributions of X given Y under $\mathbb{P}^*$ and $\mathbb{Q}$ are the same. We will see below that this is not a coincidence.

In the second minimization problem, one is given a distribution $\mathbf{P}$, with the marginal of Y given by P and finds the best approximation to it in the set $\mathscr{Q}$ of distributions which make $Y = (Y_-, Y_+)$ conditionally independent given X. Let $\mathbf{Q}^*$ denote the optimal distribution of (Y_-, X, Y_+). Eqs. (10) and (11) can then be interpreted, in terms of the corresponding measures, as

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

and

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

We see that the optimal solution $\mathbb{Q}^*$ is such that the marginal distributions of (X, Y_-) under $\mathbb{P}$ and $\mathbb{Q}^*$ coincide as well as the conditional distributions of Y_+ given X under $\mathbb{P}$ and $\mathbb{Q}^*$. Again, this is not a coincidence, as we will explain below.

Remark 3.4. As a side remark we notice that the minimization of $D(\mathbf{Q}\|\mathbf{P})$ over $\mathbf{P} \in \mathscr{P}$ for a given $\mathbf{Q} \in \mathscr{Q}$ yields the same solution $\mathbf{P}^*$. A similar result does not hold for the second minimization problem. This remark is not relevant for what follows.

We can now state the so called *Pythagorean rules* for the two partial minimization problems. This terminology was introduced by Csiszár [1].

Lemma 3.5. *For fixed* $\mathbf{Q}$ *and* $\mathbf{P}^* = \mathbf{P}^*(\mathbf{Q})$ *it holds that, for any* $\mathbf{P} \in \mathscr{P}$,

$$D(\mathbf{P}\|\mathbf{Q}) = D(\mathbf{P}\|\mathbf{P}^*) + D(\mathbf{P}^*\|\mathbf{Q}), \tag{12}$$

moreover

$$D(\mathbf{P}^*\|\mathbf{Q}) = D(P\|Q), \tag{13}$$

where

$$Q(ij) = \sum_l \mathbf{Q}(ilj). \tag{14}$$

For fixed $\mathbf{P}$ *and* $\mathbf{Q}^* = \mathbf{Q}^*(\mathbf{P})$ *it holds that, for any* $\mathbf{Q} \in \mathscr{Q}$,

$$D(\mathbf{P}\|\mathbf{Q}) = D(\mathbf{P}\|\mathbf{Q}^*) + D(\mathbf{Q}^*\|\mathbf{Q}). \tag{15}$$

Proof. To prove the first rule we compute

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

The first rule follows. To prove the relation (13) insert Eq. (9) into $D(\mathbf{P}^*\|\mathbf{Q})$ and sum over l to get

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

To prove the second rule 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 observe that $\mathbf{Q}(il\cdot) = Q_-(il)$, and $\mathbf{Q}(j|l) = Q_+(lj)/\sum_j Q_+(lj)$, and $Q_-^*(il) = \mathbf{P}(il\cdot)$ and $Q_+^*(lj) = \mathbf{P}(j|l)$. We now compute

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

The second rule follows. □

With the aid of the relation (13) we can now prove Proposition 3.2.

Proof of Proposition 3.2. With $\mathbf{P}^* = \mathbf{P}^*(\mathbf{Q})$, the optimal solution of the partial minimization over $\mathscr{P}$, we have

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

It follows that $\inf_{\mathbf{P}\in\mathscr{P},\mathbf{Q}\in\mathscr{Q}} D(\mathbf{P}\|\mathbf{Q}) \geqslant \min_{Q\in\mathscr{Q}} D(P\|Q)$.

Conversely, let $\mathbf{Q}$ in $\mathscr{Q}$ be given and let Q be defined by $Q(ij) = \sum_l \mathbf{Q}(ilj)$. From

$$\begin{aligned} D(P\|Q) &= D(\mathbf{P}^*(\mathbf{Q})\|\mathbf{Q}) \\ &\geqslant \inf_{\mathbf{P}\in\mathscr{P},\mathbf{Q}\in\mathscr{Q}} D(\mathbf{P}\|\mathbf{Q}), \end{aligned}$$

we obtain

$$\min_{Q\in\mathscr{Q}} D(P\|Q) \geqslant \inf_{\mathbf{P}\in\mathscr{P},\mathbf{Q}\in\mathscr{Q}} D(\mathbf{P}\|\mathbf{Q}).$$

Finally we show that we can replace the infima by minima. Let Q_-^* and Q_+^* be such that $(Q_-, Q^+) \mapsto D(P\|Q_-Q^+)$ is minimized (their existence is guaranteed by Proposition 2.1). Let $\mathbf{Q}^*$ be a corresponding element in $\mathscr{Q}$ and $\mathbf{P}^* = \mathbf{P}^*(\mathbf{Q}^*)$. Then $D(\mathbf{P}^*\|\mathbf{Q}^*) = D(P\|Q_-^*Q_+^*)$ and the result follows. □

For a probabilistic derivation of the solutions of the two partial minimization problems and of their corresponding Pythagorean rules, we use a general result (Lemma 3.6 below) on the I-divergence between two joint laws of any random vector (U, V). We denote the law of (U, V) under arbitrary probability measures $\mathbb{P}$ and $\mathbb{Q}$ by $\mathbb{P}^{U,V}$ and $\mathbb{Q}^{U,V}$. The conditional distributions of U given V are summarized by the matrices $\mathbb{P}^{U|V}$ and $\mathbb{Q}^{U|V}$, with the obvious convention $\mathbb{P}^{U|V}(ij) = \mathbb{P}(U = j|V = i)$ and likewise for $\mathbb{Q}^{U|V}$.

Lemma 3.6. *It holds that*

$$D(\mathbb{P}^{U,V}\|\mathbb{Q}^{U,V}) = \mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{U|V}\|\mathbb{Q}^{U|V}) + D(\mathbb{P}^{V}\|\mathbb{Q}^{V}), \tag{16}$$

where

$$D(\mathbb{P}^{U|V}\|\mathbb{Q}^{U|V}) = \sum_j P(U = j|V) \log \frac{P(U = j|V)}{Q(U = j|V)}.$$

If moreover $V = (V_1, V_2)$, and U, V_2 are conditionally independent given V_1 under $\mathbb{Q}$, then the first term on the RHS of (16) can be written as

$$\mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{U|V}\|\mathbb{Q}^{U|V}) = \mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{U|V}\|\mathbb{P}^{U|V_1}) + \mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{U|V_1}\|\mathbb{Q}^{U|V_1}). \tag{17}$$

Proof. It follows from elementary manipulations. □

The first minimization problem can be solved probabilistically as follows. Given $\mathbf{Q}$ we are to find its best approximation within $\mathscr{P}$. Let $\mathbb{Q}$ correspond to the given $\mathbf{Q}$ and $\mathbb{P}$ correspond to the generic $\mathbf{P} \in \mathscr{P}$. Choosing $U = X$, $V = Y = (Y_-, Y_+)$ in Lemma 3.6, and remembering that $\mathbb{P}^Y$ is determined by P for all $\mathbf{P} \in \mathscr{P}$, Eq. (16) now reads

$$D(\mathbf{P}\|\mathbf{Q}) = \mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{X|Y}\|\mathbb{Q}^{X|Y}) + D(P\|Q), \tag{18}$$

where the matrix Q is as in (14). The problem is equivalent to the minimization of $\mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{X|Y} \| \mathbb{Q}^{X|Y})$ w.r.t. $\mathbf{P} \in \mathscr{P}$, which is attained (with value 0) at $\mathbb{P}^*$ with $\mathbb{P}^{*X|Y} = \mathbb{Q}^{X|Y}$ and $\mathbb{P}^{*Y} = P$. To derive probabilistically the corresponding Pythagorean rule, we apply (16) with $\mathbb{P}^*$ instead of $\mathbb{Q}$. We obtain, using $\mathbb{P}^Y = \mathbb{P}^{*Y}$,

$$D(\mathbb{P}^{X,Y} \| \mathbb{P}^{*X,Y}) = \mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{X|Y} \| \mathbb{P}^{*X|Y}). \tag{19}$$

Since also

$$\mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{X|Y} \| \mathbb{Q}^{X|Y}) = \mathbb{E}_{\mathbb{P}} D(\mathbb{P}^{X|Y} \| \mathbb{P}^{*X|Y}), \tag{20}$$

we combine Eqs. (19) and (20) and insert the result into (18). Recognizing the fact that $D(\mathbf{P} \| \mathbf{P}^*) = D(\mathbb{P}^{X,Y} \| \mathbb{P}^{*X,Y})$, and using $D(\mathbf{P}^* \| \mathbf{Q}) = D(P \| Q)$ according to (13), we then identify (18) as the first Pythagorean rule (12).

The treatment of the second minimization problem follows a similar pattern. Given $\mathbf{P}$ we are to find its best approximation within $\mathscr{Q}$. Let $\mathbb{P}$ correspond to the given $\mathbf{P}$ and $\mathbb{Q}$ correspond to the generic $\mathbf{Q} \in \mathscr{Q}$. Choosing $U = Y_+$, $V_1 = X$ and $V_2 = Y_-$ in Lemma 3.6, and remembering that under any $\mathbf{Q} \in \mathscr{Q}$ the r.v. Y_-, Y_+ are conditionally independent given X, Eq. (16) refined with (17) now reads

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

The problem is equivalent to the minimizations of the second and third I-divergences on the RHS w.r.t. $\mathbf{Q} \in \mathscr{Q}$, which are attained (both with value 0) at $\mathbb{Q}^*$ with $\mathbb{Q}^{*Y_+|X} = \mathbb{P}^{Y_+|X}$ and $\mathbb{Q}^{*Y_-,X} = \mathbb{P}^{Y_-,X}$. Note that X has the same distribution under $\mathbb{P}$ and $\mathbb{Q}^*$. To derive probabilistically the corresponding Pythagorean rule we notice that

$$D(\mathbf{P} \| \mathbf{Q}) - D(\mathbf{P} \| \mathbf{Q}^*) = \mathbb{E}_{\mathbb{Q}^*} D(\mathbb{Q}^{*Y_+|X} \| \mathbb{Q}^{Y_+|X}) + D(\mathbb{Q}^{*Y_-,X} \| \mathbb{Q}^{Y_-,X}). \tag{21}$$

In the right hand side of (21) we can, by conditional independence, replace $\mathbb{E}_{\mathbb{Q}^*} D(\mathbb{Q}^{*Y_+|X} \| \mathbb{Q}^{Y_+|X})$ with $\mathbb{E}_{\mathbb{Q}^*} D(\mathbb{Q}^{*Y_+|X,Y_-} \| \mathbb{Q}^{Y_+|X,Y_-})$. By yet another application of (16), we thus see that $D(\mathbf{P} \| \mathbf{Q}) - D(\mathbf{P} \| \mathbf{Q}^*) = D(\mathbf{Q}^* \| \mathbf{Q})$, which is the second Pythagorean rule (15).

4. 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 3.2 we can lift this problem to the $\mathscr{P} \times \mathscr{Q}$ space. Starting with an arbitrary $\mathbf{Q}^0 \in \mathscr{Q}$ with positive elements, we adopt the following alternating minimization scheme

$$\to \mathbf{Q}^t \to \mathbf{P}^t \to \mathbf{Q}^{t+1} \to \mathbf{P}^{t+1} \to \tag{22}$$

where $\mathbf{P}^t = \mathbf{P}^*(\mathbf{Q}^t)$, $\mathbf{Q}^{t+1} = \mathbf{Q}^*(\mathbf{P}^t)$.

To relate this algorithm to the one of Section 2 (formulas (6) and (7)) we combine two steps of the alternating minimization at a time. From (22) we get

$$\mathbf{Q}^{t+1} = \mathbf{Q}^*(\mathbf{P}^*(\mathbf{Q}^t)).$$

Computing the optimal solutions according to (9), (10) and (11) one gets from here the formulas (6) and (7) of Section 2.

The Pythagorean rules allow us to easily compute the update gain $D(P \| Q^t) - D(P \| Q^{t+1})$ of the algorithm.

Proposition 4.1. *The update gain at each iteration of the algorithm (22) in terms of the matrices Q^t is given by*

$$D(P\|Q^t) - D(P\|Q^{t+1}) = D(\mathbf{P}^t\|\mathbf{P}^{t+1}) + D(\mathbf{Q}^{t+1}\|\mathbf{Q}^t). \tag{23}$$

Proof. The two Pythagorean rules from Lemma 3.5 now take the forms

$$D(\mathbf{P}^t\|\mathbf{Q}^t) = D(\mathbf{P}^t\|\mathbf{Q}^{t+1}) + D(\mathbf{Q}^{t+1}\|\mathbf{Q}^t),$$
$$D(\mathbf{P}^t\|\mathbf{Q}^{t+1}) = D(\mathbf{P}^t\|\mathbf{P}^{t+1}) + D(\mathbf{P}^{t+1}\|\mathbf{Q}^{t+1}).$$

Addition of these two equations results in

$$D(\mathbf{P}^t\|\mathbf{Q}^t) = D(\mathbf{P}^t\|\mathbf{P}^{t+1}) + D(\mathbf{P}^{t+1}\|\mathbf{Q}^{t+1}) + D(\mathbf{Q}^{t+1}\|\mathbf{Q}^t)$$

and since $D(\mathbf{P}^t\|\mathbf{Q}^t) = D(P\|Q^t)$ from (13), the result follows. □

Remark 4.2. If one starts the algorithm with matrices (Q_-^0, Q_+^0) in the interior of the domain, the iterations will remain in the interior. Suppose that, at step n, the update gain is zero. Then, from (23), we get that $D(\mathbf{Q}^{t+1}\|\mathbf{Q}^t) = 0$. Hence the tensors $\mathbf{Q}^{t+1}$ and $\mathbf{Q}^t$ are identical. From this it follows by summation that $Q_-^{t+1} = Q_-^t$. But then we also have the equality $Q_-^t(il)Q_+^{t+1}(lj) = Q_-^t(il)Q_+^t(lj)$ for all i, l, j. Since all $Q_-^t(il)$ are positive, we also have $Q_+^{t+1} = Q_+^t$. Hence, the updating formulas strictly decrease the objective function until the algorithm reaches a fixed point.

We close this section with the proof of Proposition 2.1 in which we use the result of Proposition 4.1.

Proof of Proposition 2.1. We first prove that there exists a pair of matrices (W, H) with $He_m = e_k$ and $We_k = Ve_n$ for which $D(V\|WH)$ is finite. Put $W = \frac{1}{k}Ve_ne_k^\top$ and $H = \frac{1}{e_m^\top Ve_n}e_ke_m^\top V$. Note that indeed $He_m = e_k$ and $We_k = Ve_n$ and that all elements of W and H, and hence those of WH, are positive, $D(V\|WH)$ is therefore finite.

Next we show that we can restrict ourselves to minimization over a compact set $\mathscr{K}$ of matrices. Specifically, we will show that for all positive matrices W and H, there exist positive matrices W' and H' with $(W', H') \in \mathscr{K}$ such that $D(V\|W'H') \leqslant D(V\|WH)$. We choose for arbitrary W^0 and H^0 the matrices W^1 and H^1 according to (3) and (4). It follows from Proposition 4.1 that indeed $D(V\|W^1H^1) \leqslant D(V\|W^0H^0)$. Moreover, it is immediately clear from (3) and (4) that we have $W^1e = Ve$ and $H^1e = e$. Hence, it is sufficient to confine search to the compact set $\mathscr{L}$ where $He = e$ and $We = Ve$.

Fix a pair of indices i, j. Since we can compute the divergence elementwise we have the trivial estimate

$$D(V\|WH) \geqslant V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}.$$

Since for $V_{ij} > 0$ the function $d_{ij} : x \to V_{ij}\log\frac{V_{ij}}{x} - V_{ij} + x$ is decreasing on $(0, V_{ij})$, we have for any sufficiently small $\varepsilon > 0$ (of course $\varepsilon < V_{ij}$) that $d_{ij}(x) > d_{ij}(\varepsilon)$ for $x \leqslant \varepsilon$ and of course $\lim_{\varepsilon\to 0} d_{ij}(\varepsilon) = \infty$. Hence to find the minimum of d_{ij}, it is sufficient to look at $x \geqslant \varepsilon$. Let $\varepsilon_0 > 0$ and such that $\varepsilon_0 < \min\{V_{ij} : V_{ij} > 0\}$. Let $\mathscr{G}$ be the set of (W, H) such that $(WH)_{ij} \geqslant \varepsilon_0$ for all i, j with $V_{ij} > 0$. Then $\mathscr{G}$ is closed. Take now $\mathscr{K} = \mathscr{L} \cap \mathscr{G}$, then $\mathscr{K}$ is the compact

set we are after. Let us observe that $\mathscr{K}$ is non-void for sufficiently small ε_0. Clearly the map $(W, H) \mapsto D(V\|WH)$ is continuous on $\mathscr{K}$ and thus attains its minimum. □

5. Auxiliary functions

Algorithms for recursive minimization can often be constructed by using *auxiliary functions*. For the problem of minimizing the divergence $D(V\|WH)$, some such functions can be found in [10] and they are analogous to functions that are used when studying the EM algorithm, see [15]. The choice of an auxiliary function is usually based on *ad hoc* reasoning, like for instance finding a Lyapunov function for studying the stability of the solutions of a differential equation. We show in this section that the lifted version of the divergence minimization problem leads in a natural way to useful auxiliary functions. Let us first explain what is meant by an auxiliary function.

Suppose one wants to minimize a function $x \mapsto F(x)$, defined on some domain. The function $(x, x') \mapsto G(x, x')$ is an auxiliary function for F if

$$G(x, x') \geqslant F(x'), \quad \forall x, x',$$
$$G(x, x) = F(x), \quad \forall x.$$

If we define (assuming that the arg min below exists and is unique)

$$x' = x'(x) = \arg\min G(x, \cdot), \tag{24}$$

then we have

$$F(x') \leqslant G(x, x') \leqslant G(x, x) = F(x)$$

and hence the value of F decreases by replacing x with x'. A recursive procedure to find the minimum of F can be based on the recipe (24) by taking $x = x^t$ and $x' = x^{t+1}$. To be useful an auxiliary function G must allow for a simple computation or characterization of $\arg\min G(x, \cdot)$.

We consider now the minimization of $D(P\|Q)$ and its lifted version, the minimization of $D(\mathbf{P}\|\mathbf{Q})$ as in Section 3. In particular, with reference to the alternating minimization scheme (22), with the notations of Section 4, we know that $\mathbf{Q}^{t+1}$ is found by minimizing $\mathbf{Q}' \mapsto D(\mathbf{P}^*(\mathbf{Q}^t)\|\mathbf{Q}')$. This strongly motivates the choice of the function

$$(\mathbf{Q}, \mathbf{Q}') \mapsto G(\mathbf{Q}, \mathbf{Q}') = D(\mathbf{P}^*(\mathbf{Q})\|\mathbf{Q}')$$

as an auxiliary function for minimizing $D(P\|Q)$ w.r.t. Q.

Using the decomposition of the divergence in Eq. (16) we can rewrite G as

$$G(\mathbf{Q}, \mathbf{Q}') = D(\mathbb{P}^{*Y}\|\mathbb{Q}'^{Y}) + \mathbb{E}_{\mathbb{P}^*} D(\mathbb{P}^{*X|Y}\|\mathbb{Q}'^{X|Y}). \tag{25}$$

Since $\mathbb{P}^{*X|Y} = \mathbb{Q}^{X|Y}$, and $\mathbb{P}^{*Y} = P$ we can rewrite (25) as

$$G(\mathbf{Q}, \mathbf{Q}') = D(P\|\mathbb{Q}'^{Y}) + \mathbb{E}_P D(\mathbb{Q}^{X|Y}\|\mathbb{Q}'^{X|Y}). \tag{26}$$

From (26) it follows that $G(\mathbf{Q}, \mathbf{Q}') \geqslant D(P\|Q')$, and that $G(\mathbf{Q}, \mathbf{Q}) = D(P\|Q)$, precisely the two properties that define an auxiliary function for $D(P\|Q)$.

In [10] one can find two auxiliary functions for the original minimization problem $D(V\|WH)$. One function is for minimization over H with fixed W, the other for minimization over W with fixed H. To show the connection with the function G defined above, we first make the dependence of G on Q_-, Q_+, Q'_-, Q'_+ explicit by writing $G(\mathbf{Q}, \mathbf{Q}')$ as $G(Q_-, Q_+, Q'_-, Q'_+)$.

The auxiliary function for minimization with fixed Q_- can then be taken as

$$Q'_+ \mapsto G^+_{\mathbf{Q}}(Q'_+) = G(Q_-, Q_+, Q_-, Q'_+),$$

whereas the auxiliary function for minimization with fixed Q_+ can be taken as

$$Q'_- \mapsto G^-_{\mathbf{Q}}(Q'_-) = G(Q_-, Q_+, Q'_-, Q_+)$$

The functions $G^+_{\mathbf{Q}}$ and $G^-_{\mathbf{Q}}$ correspond to the auxiliary functions in [10], where they are given in an explicit form, but where no rationale for them is given.

For the different auxiliary functions introduced above, we will now compute the update gains and compare these expressions with (23).

Lemma 5.1. *Consider the auxiliary functions $G, G^-_{\mathbf{Q}}, G^+_{\mathbf{Q}}$ above. Denote by Q'_- and Q'_+ the minimizers of the auxiliary functions in all three cases. The following equalities hold*:

$$D(P\|Q_-Q_+) - G^-_{\mathbf{Q}}(Q'_-) = D(\mathbb{Q}'^{Y_-,X}\|\mathbb{Q}^{Y_-,X}) \tag{27}$$

$$D(P\|Q_-Q_+) - G^+_{\mathbf{Q}}(Q'_+) = \mathbb{E}_{\mathbb{P}^*} D(\mathbb{Q}'^{Y_+|X}\|\mathbb{Q}^{Y_+|X}) \tag{28}$$

$$D(P\|Q_-Q_+) - G(Q_-, Q_+, Q'_-, Q'_+) = D(\mathbb{Q}'^{Y_-,X}\|\mathbb{Q}^{Y_-,X}) + \mathbb{E}_{\mathbb{Q}'} D(\mathbb{Q}'^{Y_+|X}\|\mathbb{Q}^{Y_+|X}). \tag{29}$$

Proof. We prove (29) first. The other two follow from this. A simple computation, valid for any Q_- and Q_+, yields

$$\begin{aligned} &D(P\|Q_-Q_+) - G(Q_-, Q_+, Q'_-, Q'_+) \qquad (30)\\ &= \sum_{ij} P(ij) \sum_l \frac{\mathbf{Q}(ilj)}{Q(ij)} \left(\log \frac{Q'_-(il)}{Q_-(il)} + \log \frac{Q'_+(lj)}{Q_+(lj)} \right) \\ &= \sum_{il} \left(\sum_j \frac{P(ij)\mathbf{Q}(ilj)}{Q(ij)} \right) \log \frac{Q'_-(il)}{Q_-(il)} + \sum_{lj} \left(\sum_i \frac{P(ij)\mathbf{Q}(ilj)}{Q(ij)} \right) \log \frac{Q'_+(lj)}{Q_+(lj)} \qquad (31) \end{aligned}$$

Now we exploit the known formulas (6) and (7) for the optimizing Q'_- and Q'_+. The first term in (31) becomes in view of (6) (or, equivalently, in view of (9) and (10))

$$\sum_{il} Q'_-(il) \log \frac{Q'_-(il)}{Q_-(il)},$$

which gives the first term on the RHS of (29). Similarly, the second term in (31) can be written in view of (7) as

$$\sum_l \left(\sum_{ij} \mathbf{Q}'(ilj) \right) \sum_j Q'_+(lj) \log \frac{Q'_+(lj)}{Q_+(lj)},$$

which yields the second term on the RHS of formula (29). Formulas (27) and (28) are obtained similarly, noticing that optimization of $G^+_{\mathbf{Q}}$ and $G^-_{\mathbf{Q}}$ separately yield the same Q'_+, respectively Q'_-, as those obtained by minimization of G. □

Remark 5.2. Notice that although for instance $G^-_{\mathbf{Q}}(Q'_-) \geqslant D(P\|Q'_-Q'_+)$ for all Q'_- and Q'_+, we have for the optimal Q'_- that $G^-_{\mathbf{Q}}(Q'_-) \leqslant D(P\|Q_-Q_+)$.

Corollary 5.3. *The update gain of the algorithm* (6), (7) *can be represented by*

$$D(P\|Q^t) - D(P\|Q^{t+1}) = D\left(\mathbb{Q}^{t+1^{Y_-,X}}\|\mathbb{Q}^{t^{Y_-,X}}\right) + \mathbb{E}_{\mathbb{Q}^{t+1}} D\left(\mathbb{Q}^{t+1^{Y_+|X}}\|\mathbb{Q}^{t^{Y_+|X}}\right) + \mathbb{E}_P D\left(\mathbb{Q}^{t^{X|Y}}\|\mathbb{Q}^{t+1^{X|Y}}\right). \tag{32}$$

Proof. Write

$$D(P\|Q^t) - D(P\|Q^{t+1}) = D(P\|Q^t) - G(\mathbf{Q}^t, \mathbf{Q}^{t+1}) + G(\mathbf{Q}^t, \mathbf{Q}^{t+1}) - D(P\|Q^{t+1})$$

and use Eqs. (25) and (29). □

We return to the update formula (23). A computation shows the following equalities.

$$D(\mathbf{P}^t\|\mathbf{P}^{t+1}) = \mathbb{E}_P D(\mathbb{Q}^{t^{X|Y}}\|\mathbb{Q}^{t+1^{X|Y}}) \tag{33}$$

$$D(\mathbf{Q}^{t+1}\|\mathbf{Q}^t) = D(\mathbb{Q}^{t+1^{Y_-,X}}\|\mathbb{Q}^{t^{Y_-,X}}) + \mathbb{E}_{\mathbb{Q}^{t+1}} D(\mathbb{Q}^{t+1^{Y_+|X}}\|\mathbb{Q}^{t^{Y_+|X}}). \tag{34}$$

In Eq. (33) we recognize the second term in the auxiliary function, see (26). Eq. (34) corresponds to Eq. (29) of Lemma 5.1 and we see that formula (23) is indeed the same as (32).

The algorithm (6), (7) is to be understood by using these two equations simultaneously. As an alternative one could first use (6) to obtain Q_-^{t+1} and, instead of using Q_-^t, feed this result into (7) to obtain Q_+^{t+1}. If we do this, we can express the update gain of the first partial step, like in the proof of Corollary 5.3, by adding the result of Eq. (27) to the second summand of (26), with the understanding that $\mathbb{Q}'$ is now given by the $Q^{t+1}(ij)Q^t(lj)$. The update gain of the second partial step is likewise obtained by combining the result of (28) and the second summand of (26), with the understanding that now $\mathbb{Q}$ is to be interpreted as given by the $Q^{t+1}(ij)Q^t(lj)$. Of course, as another alternative, the order of the partial steps can be reversed. Clearly, the expressions for the update gains for these cases also result from working with the auxiliary functions $G_{\mathbf{Q}}^-$ and $G_{\mathbf{Q}}^+$, the Eqs. (27) and (28) and proceeding as in the proof of Corollary 5.3.

6. Convergence properties

In this section we study the convergence properties of the divergence minimization algorithm (6), (7).

The next theorem states that the sequences generated by the algorithm converge for every (admissible) initial value. Of course the limits will in general depend on the initial value.

Theorem 6.1. *Let $Q_-^t(il)$, $Q_+^t(lj)$ be generated by the algorithm* (6), (7) *and $\mathbf{Q}^t$ the corresponding tensors. Then the $Q_-^t(il)$ converge to limits $Q_-^\infty(il)$ and the $\mathbf{Q}^t$ converges to a limit $\mathbf{Q}^\infty$ in $\mathcal{Q}$. The $Q_+^t(lj)$ converge to limits $Q_+^\infty(lj)$ for all l with $\sum_i Q_+^\infty(il) > 0$.*

Proof. We first show that the Q_-^t and Q_+^t form convergent sequences. We start with Eq. (23). By summing over n we obtain

$$D(P\|Q^0) - D(P\|Q^t) = \sum_{k=1}^{t-1}\left(D(\mathbf{P}^s\|\mathbf{P}^{s+1}) + D(\mathbf{Q}^{s+1}\|\mathbf{Q}^s)\right).$$

It follows that $\sum_{k=1}^\infty D(\mathbf{P}^s\|\mathbf{P}^{s+1})$ and $\sum_{k=1}^\infty D(\mathbf{Q}^{s+1}\|\mathbf{Q}^s)$ are finite. Now we use that fact that for any two probability measures, the Kullback–Leibler divergence $D(\mathbb{P}\|\mathbb{Q})$ is greater than

or equal to their Hellinger distance $H(\mathbb{P}, \mathbb{Q})$, which is the L^2 distance between the square roots of corresponding densities w.r.t. some dominating measure, see [13, p. 368]. In our case we have $H(\mathbb{Q}^s, \mathbb{Q}^{s+1}) = \sum_{ilj}(\sqrt{\mathbf{Q}^{s+1}(ilj)} - \sqrt{\mathbf{Q}^s(ilj)})^2$. So we obtain that

$$\sum_{k=1}^{\infty} H(\mathbf{Q}^{s+1}, \mathbf{Q}^s) < \infty.$$

We therefore have that, pointwise, the tensors $\mathbf{Q}^t$ form a Cauchy sequence and hence have a limit $\mathbf{Q}^\infty$. We will show that $\mathbf{Q}^\infty$ belongs to $\mathscr{Q}$. Since the $\mathbf{Q}^t(ilj)$ converge to limits $\mathbf{Q}^\infty(ilj)$, by summation we have that the marginals $Q_-^t(il) = \mathbf{Q}^t(il\cdot)$ converge to limits $\mathbf{Q}^\infty(il\cdot)$ (we use the notation of the proof of Lemma 3.5), and likewise we have convergence of the marginals $\mathbf{Q}^t(\cdot lj)$ to $\mathbf{Q}^\infty(\cdot lj)$ and $\mathbf{Q}^t(\cdot l\cdot)$ to $\mathbf{Q}^\infty(\cdot l\cdot)$. Hence, if $\mathbf{Q}^\infty(\cdot l\cdot) > 0$, then the $Q_+^t(lj)$ converge to $Q_+^\infty(ij) := \mathbf{Q}^\infty(\cdot lj)/\mathbf{Q}^\infty(\cdot l\cdot)$ and we have $\mathbf{Q}^\infty(ilj) = \mathbf{Q}^\infty(il\cdot)Q_+^\infty(ij)$. Now we analyze the case where $\mathbf{Q}^\infty(\cdot l_0\cdot) = 0$ for some l_0. Since in this case both $\mathbf{Q}^\infty(il_0j)$ and $\mathbf{Q}^\infty(il_0\cdot)$ are zero, we have still have a factorization $\mathbf{Q}^\infty(il_0j) = Q_-^\infty(il_0)Q_+^\infty(l_0j)$, where we can assign to the $Q_+^\infty(l_0j)$ arbitrary values. Let L be the set of l for which $\sum_i Q_-^\infty(il) > 0$. Then $Q^\infty(ij) = \sum_{l\in L} Q_-^\infty(il)Q_+^\infty(lj)$ and the Q^t converge to Q^∞. This proves the theorem. □

Remark 6.2. Theorem 6.1 says nothing of the convergence of the $Q_+^t(lj)$ for those l where $\sum_i Q_-^\infty(il) = 0$. But their behavior is uninteresting from a factorization point of view. Indeed, since the l-th column of Q_-^∞ is zero, the values of the l-th row of Q_+^∞ are not relevant, since they do not appear in the product $Q_-^\infty Q_+^\infty$. As a matter of fact, we now deal with an approximate nonnegative factorization with a lower inner size. See also Remark 3.3.

In the next theorem we characterize the properties of the fixed points of the algorithm. Recall from Section 2 that the objective function has no local maxima in the interior of the domain.

Theorem 6.3. *If (Q_-, Q_+) is a limit point of the algorithm (6), (7) in the interior of the domain, then it is a stationary point of the objective function D. If (Q_-, Q_+) is a limit point on the boundary of the domain corresponding to an approximate factorization where none of the columns of Q_- is zero ($\sum_i Q_-(il) > 0$ for all l), then all partial derivatives $\frac{\partial D}{\partial Q_-(il)}$ and $\frac{\partial D}{\partial Q_+(lj)}$ are nonnegative.*

Proof. By computing the first order partial derivatives of the objective function, using the middle term of Eq. (5), we can rewrite the update Eqs. (6), (7) as

$$Q_-^{t+1}(il) = Q_-^t(il)\left(-\frac{\partial D^t}{\partial Q_-(il)} + 1\right) \tag{35}$$

and

$$Q_+^{t+1}(lj)\left(\sum_i Q_-^{t+1}(il)\right) = Q_+^t(lj)\left(-\frac{\partial D^t}{\partial Q_+(lj)} + \sum_i Q_-^t(il)\right), \tag{36}$$

where $\frac{\partial D^t}{\partial Q_-(il)}$ stands for the partial derivative $\frac{\partial D}{\partial Q_-(il)}$ evaluated at (Q_-^t, Q_+^t) and likewise for $\frac{\partial D^t}{\partial Q_+(lj)}$.

Let (Q_-, Q_+) be a limit point of the algorithm. Eqs. (35) and (36) become

$$Q_-(il) = Q_-il\left(-\frac{\partial D}{\partial Q_-(il)} + 1\right)$$

$$Q_+(lj)\left(\sum_i Q_-(il)\right) = Q_+(lj)\left(-\frac{\partial D}{\partial Q_+(lj)} + \sum_i Q_-(il)\right).$$

It follows that we then have the relations

$$Q_-(il)\frac{\partial D}{\partial Q_-(il)} = 0$$

and

$$Q_+(lj)\frac{\partial D}{\partial Q_+(lj)} = 0.$$

We first consider Q_-. Suppose that for some i and l we have $Q_-(il) > 0$, then necessarily $\frac{\partial D}{\partial Q_-(il)} = 0$. Suppose now that for some i, l we have $Q_-(il) = 0$ and that $\frac{\partial D}{\partial Q_-(il)} < 0$. Of course, by continuity, this partial derivative will be negative in a sufficiently small neighborhood of this limit point. Since we deal with a limit point of the algorithm, we must have infinitely often for the iterates that $Q_-^{t+1}(il) < Q_-^t(il)$. From (35) we then conclude that in these points we have $\frac{\partial D}{\partial Q_-(il)} > 0$. Clearly, this contradicts our assumption of a negative partial derivative, since eventually the iterates will be in the small neighborhood of the limit point, where the partial derivative is positive. Hence, we conclude that $\frac{\partial D}{\partial Q_-(il)} \geqslant 0$, if $Q_-(il) = 0$. The proof of the companion statement for the $Q_+(lj)$ is similar. If $Q_+(lj) > 0$, the corresponding partial derivative is zero. Let l be such that $Q_+(lj) = 0$ and suppose that we have that $\frac{\partial D}{\partial Q_+(lj)} < 0$. If we run the algorithm, then $\frac{\partial D^t}{\partial Q_+(lj)} / \sum_i Q_-^{t+1}(il)$ converges to a negative limit, whereas $\sum_i Q_-^t(il) / \sum_i Q_-^{t+1}(il)$ converges to one. Hence there is $\eta > 0$ such that eventually $\frac{\partial D^t}{\partial Q_+(lj)} / \sum_i Q_-^{t+1}(il) < -2\eta/3$ and $\sum_i Q_-^t(il) / \sum_i Q_-^{t+1}(il) > 1 - \eta/3$. Hence eventually we would have, see (36),

$$Q_+^{t+1}(lj) - Q_+^t(lj) = Q_+^t(lj)\left(-\frac{\frac{\partial D^t}{\partial Q_+(lj)}}{\sum_i Q_-^{t+1}(il)} + \frac{\sum_i Q_-^t(il)}{\sum_i Q_-^{t+1}(il)} - 1\right) > \eta/3,$$

which contradicts convergence of $Q_+^t(lj)$ to zero. □

Remark 6.4. If it happens that a limit point Q_- has a zero l-th column, then it can easily be shown that the partial derivatives $\frac{\partial D}{\partial Q_+(lj)}$ of D are zero. Nothing can be said of the values of the partial derivatives $\frac{\partial D}{\partial Q_-(il)}$ for such l. But, see also Remark 6.2, this case can be reduced to one with a lower inner size factorization, for which the assertion of Theorem 6.3 is valid.

Corollary 6.5. *The limit points of the algorithm with $\sum_i Q_-(il) > 0$ for all l are all Kuhn–Tucker points for minimization of D under the inequality constraints $Q_- \geqslant 0$ and $Q_+ \geqslant 0$.*

Proof. Consider the Lagrange function L defined by

$$L(Q_-, Q_+) = D(P\|Q_-Q_+) - \lambda \cdot Q_- - \mu \cdot Q_+,$$

where for instance the inner product $\lambda \cdot Q_-$ is to be read as $\sum_{il} \lambda_{il} Q_-(il)$ for $\lambda_{il} \in \mathbb{R}$. Let us focus on a partial derivative $\frac{\partial L}{\partial Q_-(il)}$ in a fixed point of the algorithm. The treatment of the other partial derivatives is similar. From the proof of Theorem 6.3 we know that in a fixed point we have $Q_-(il)\frac{\partial D}{\partial Q_-(il)} = 0$. Suppose that $Q_-(il) > 0$, then $\frac{\partial D}{\partial Q_-(il)} = 0$ and the Kuhn–Tucker conditions for this variable are satisfied with $\lambda_{il} = 0$. If $Q_-(il) = 0$, then we know from Theorem 6.3 that

$\frac{\partial D}{\partial Q_-(il)} \geqslant 0$. By taking $\lambda_{il} = \frac{\partial D}{\partial Q_-(il)} \geqslant 0$, we see that also here the Kuhn–Tucker conditions are satisfied. □

Remark 6.6. Wu [15] has a number of theorems that characterize the limit points of the closely related EM algorithm, or generalized EM algorithm. These are all consequence of a general convergence result in Zangwill [16]. The difference of our results with his is, that we also *have to* consider possible limit points on the boundary, whereas Wu's results are based on the assumption that all limit points lie in the interior of the domain.

7. Relation with other minimization problems

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 [4]. 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 $\|A\|_\Sigma = (\mathrm{tr}(\mathrm{A}^{\mathrm{T}}\Sigma\mathrm{A}))^{1/2}$. 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.

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

Acknowledgments

An anonymous referee is gratefully acknowledged for helping us to improve the quality of the presentation and for suggesting to us to investigate the boundary behavior of the algorithm, similar to what has been reported in [3].

References

[1] I. Csiszár, I-divergence geometry of probability distributions and minimization problems, Ann. Prob. 3 (1975) 146–158.

[2] I. Csiszár, G. Tusnády, Information geometry and alternating minimization procedures, Statist. Decisions 1 (supplement issue) (1984) 205–237.
[3] M. Catral, L. Han, M. Neumann, R.J. Plemmons, On reduced rank nonnegative matrix factorization for symmetric nonnegative matrices, Linear Algebra Appl. 393 (2004) 107–126.
[4] A. Cutler, L. Breiman, Archetypal analysis, Technometrics 36 (1994) 338–347.
[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. With discussion, J. Roy. Statist. Soc. Ser. B 39 (1) (1977) 1–38.
[6] L. Finesso, P.J.C. Spreij, Approximate realization of finite Hidden Markov Chains, in: Proceedings of the 2002 IEEE Information Theory Workshop, 90–93, Bangalore, India, 2002.
[7] L. Finesso, P.J.C. Spreij, Approximate nonnegative matrix factorization via alternating minimization, in: Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), Leuven, July 5–9, 2004. Available from: <http://www.mtns2004.be/database/papersubmission/upload/184.pdf>.
[8] M. Hazewinkel, On positive vectors, positive matrices and the specialization order, CWI report PM-R8407 (1984).
[9] D.D. Lee, H.S. Sebastian Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[10] D.D. Lee, H.S. Sebastian Seung, Algorithms for non-negative matrix factorization, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Advances in Neural and Information Processing Systems, vol. 13, MIT Press, 2001, pp. 556–562.
[11] G. Picci, J.H. van Schuppen, On the weak finite stochastic realization problem, Springer LNCIS, vol. 58, 1984, pp. 237–242.
[12] G. Picci, J.M. van den Hof, J.H. van Schuppen, Primes in several classes of the positive matrices, Linear Algebra Appl. 277 (1998) 149–185.
[13] A.N. Shiryaev, Probability, second ed., Springer, 1996.
[14] J.A. O'Sullivan, Properties of the information value decomposition, in: Proceedings ISIT 2000, Sorrento, Italy, 2000, p. 491.
[15] C.J. Wu, On the convergence properties of the EM algorithm, Ann. Stat. 11 (1) (1983) 95–103.
[16] W.I. Zangwill, Nonlinear Programming, A Unified Approach, Prentice Hall, 1969.