Introduction to Nonnegative Matrix Factorization

Nicolas Gillis
Department of Mathematics and Operational Research
Faculté Polytechnique, Université de Mons
Rue de Houdain 9, 7000 Mons, Belgium
nicolas.gillis@umons.ac.be

Abstract

In this paper, we introduce and provide a short overview of nonnegative matrix factorization (NMF). Several aspects of NMF are discussed, namely, the application in hyperspectral imaging, geometry and uniqueness of NMF solutions, complexity, algorithms, and its link with extended formulations of polyhedra. In order to put NMF into perspective, the more general problem class of constrained low-rank matrix approximation problems is first briefly introduced.

1 Introduction

Constrained low-rank matrix approximation (CLRMA) is becoming more and more popular because it is able to extract pertinent information from large data sets; see, for example, the recent survey [87]. CLRMA is equivalent to *linear dimensionality reduction*. Given a set of n data points $m_j \in \mathbb{R}^p$ ($j = 1, 2, \dots, n$), the goal is to find a set of r basis vectors $u_k \in \mathbb{R}^p$ ($k = 1, 2, \dots, r$) and the corresponding weights v_{kj} so that for all j, $m_j \approx \sum_{k=1}^{r} v_{kj}\, u_k$. This problem is equivalent to the low-rank approximation of matrix M, with

$$M = [m_1\ m_2\ \dots\ m_n] \approx [u_1\ u_2\ \dots\ u_k][v_1\ v_2\ \dots\ v_n] = UV,$$

where each column of M is a data point, each column of U is a basis vector, and each column of V provides the coordinates of the corresponding column of M in the basis U. In other words, each column of M is approximated by a linear combination of the columns of U.

In practice, when dealing with such models, two key choices exist:

1. *Measure of the error $M - UV$.* Using the standard least-squares error, $\|M - UV\|_F^2 = \sum_{i,j}(M - UV)_{i,j}^2$, leads to principal component analysis (PCA) that can be solved by using the singular value decomposition (SVD). Surprisingly, one can show that the optimization problem in variables (U, V) has no spurious local minima (i.e., all local minima are global), which explains why it can be solved efficiently despite the error being nonconvex. Note that the resulting problem can be reformulated as a semidefinite program (SDP) by using the Ky Fan 2-k-Norm [29, Prop. 2.9].

   If data is missing or if weights are assigned to the entries of M, the problem can be cast as a weighted low-rank matrix approximation (WLRA) problem with error $\sum_{i,j} W_{i,j}(M - UV)_{i,j}^2$ for some nonnegative weight matrix W, where $W_{i,j} = 0$ when the entry (i, j) is missing [86].

Note that if W contains entries only in $\{0, 1\}$, then the problem is also referred to as PCA with missing data or low-rank matrix completion with noise.

WLRA is widely used for recommender systems [61] that predict the preferences of users for a given product based on the product's attributes and user preferences.

If the sum of the absolute values of the entries of the error $\sum_{i,j} |M - UV|_{i,j}$ is used, we obtain yet another variant more robust to outliers (sometimes referred to as robust PCA [15]). It can be used, for example, for background subtraction in video sequences where the noise (the moving objects) is assumed to be sparse while the background has low rank.

2. *Constraints that the factors U and V should satisfy.* These constraints depend on the application at hand and allow for meaningful interpretation of the factors. For example, k-means[1] is equivalent to requiring the factor V to have a single nonzero entry in each column that is equal to one, so that the columns of U are cluster centroids. Another widely used variant is sparse PCA, which requires that the factors (U and/or V) be sparse [28, 57, 69], thus yielding a more compact and easily interpretable decomposition (e.g., if V is sparse, each data point is the linear combination of only a few basis elements). Imposing componentwise nonnegativity on both factors U and V leads to nonnegative matrix factorization (NMF). For example, in document analysis where each column of M corresponds to a document (a vector of word counts), these nonnegativity constraints allow one to interpret the columns of the factor U as topics, and the columns of the factor V indicate in which proportion each document discusses each topic [64]. In this paper, we focus on this particular variant of CLRMA.

CLRMA problems are at the heart of many fields of applied mathematics and computer science, including, statistics and data analysis [56], machine learning and data mining [30], signal and image processing [1], graph theory [22], numerical linear algebra, and systems theory and control [72]. The good news for the optimization community is that these CLRMA models lead to a wide variety of theoretical and algorithmic challenges for optimizers: Can we solve these problems? Under which conditions? What is the most appropriate model for a given application? Which algorithm should we use in which situation? What type of guarantees can we provide?

CLRMA problems can be formulated in the following way:

$$\min_{U \in \Omega_U, V \in \Omega_V} \|M - UV\|. \tag{1}$$

As an introduction, below we discuss several aspects of (1).

Complexity. As soon as the norm $\|\cdot\|$ is not the Frobenius norm or the feasible domain has constraints (i.e., $\Omega_U \neq \mathbb{R}^{p \times r}$ or $\Omega_V \neq \mathbb{R}^{r \times n}$), the problem becomes difficult in most cases. For example, WLRA, robust PCA, NMF, and sparse PCA are all NP hard [44, 50, 90, 74]. An active direction of research is developing approximation algorithms for such problems; see, for example, [26] for the norm $\sum_{j=1}^{n} \|M(:,j) - UV(:,j)\|_2^p$ (for $p = 2$, this is PCA), [79] for WLRA, and [85] for the componentwise ℓ_1-norm.

Convexification. Under some conditions on the matrix M, convexification approaches can lead to optimality guarantees. When there are no constraints ($\Omega_U = \mathbb{R}^{p \times r}$, $\Omega_V = \mathbb{R}^{r \times n}$), (1) can be

[1] k-means is the problem of finding a set of centroids u_k such that the sum of the distances between each data point and the closest centroid is minimized.

equivalently rewritten as

$$\min_X \|M - X\| \text{ such that } \operatorname{rank}(X) = r.$$

From X, a solution (U, V) can be obtained by factorizing X (e.g., using the SVD). The most widely used convex models are based on minimizing the nuclear norm of X:

$$\min_X \|M - X\| + \lambda \|X\|_*, \tag{2}$$

where $\lambda > 0$ is a penalty parameter and $\|X\|_* = \sum_{i=1}^{\min(n,p)} \sigma_i(X) = \|\sigma(X)\|_1$, $\sigma(X)$ being the vector of singular values of X. This problem can be written as a semidefinite program; see [80] and the references therein.

When the matrix M satisfies some conditions depending on the model (in particular, M has to be close to a low-rank matrix), the optimal solution to (2) can be guaranteed to recover the solution of the original problem; examples include PCA with missing data [80] and robust PCA [19, 15].

As far as we know, these approaches have two drawbacks. First, if the input matrix M does not satisfy the required conditions, which is often the case in practice (e.g., for recommender systems and document classification where the input matrix is usually not close to a low-rank matrix), it is unclear whether the quality of the solution to (2) will be satisfactory. Second, the number of variables is much larger than in (1), namely, mn vs. $r(m + n)$. For large-scale problems, even first-order methods might be too costly. A possible way to handle the large positive semidefinite matrix is to (re)factor it in the SDP as the product of two matrices; this is sometimes referred to as the Burer-Monteiro approach [14]. In fact, in many cases, any stationary point can be guaranteed to be a global minimum [12, 66]; see also [65] for a survey. This is currently an active area of research: trying to identify nonconvex problems for which optimal solutions can be guaranteed to be computed efficiently (see the end of the next paragraph for other examples).

Nonconvex approaches. One can tackle (1) in many ways using standard nonlinear optimization schemes. The most straightforward and popular way is to use a two-block coordinate descent method (in particular if Ω_U and Ω_V are convex sets since the subproblems in U and V are convex):

0. Initialize (U, V).

1. $U \leftarrow X$, where X solves exactly or approximately $\min_{X \in \Omega_U} \|M - XV\|$.

2. $V \leftarrow Y$, where Y solves exactly or approximately $\operatorname{argmin}_{Y \in \Omega_V} \|M - UY\|$.

This simple scheme can be implemented in different ways. The subproblems are usually not solved up to high precision; for example, a few steps of a (fast) gradient method can be used. These methods can in general be guaranteed to converge to a stationary point of (1) [16]. More sophisticated schemes include Riemannian optimization techniques [11, 89]. Many methods based on randomization have also been developed recently; see the surveys [71, 91].

Alternating and local minimization were shown to lead to optimal solutions under assumptions similar to those needed for convexification-based approaches; see, for example, [59, 55] for PCA with missing data, [2] for (a variant of) sparse PCA, and [78] for robust PCA. Recently, [7, 38] showed that PCA with missing data has no spurious local minima (under appropriate conditions).

Outline of the paper. In the rest of this paper, we focus on a particular CLRMA problem, namely, nonnegative matrix factorization (NMF), with $\|\cdot\| = \|\cdot\|_F^2$, $\Omega_U = \mathbb{R}_+^{p \times r}$, and $\Omega_V = \mathbb{R}_+^{r \times n}$. As opposed to other CLRMA variants (such as robust PCA, sparse PCA, and PCA with missing data), as far as we know, no useful convexification approach exists.

The goal of this paper is not to provide an exhaustive survey but rather to provide a brief introduction, focusing only on several aspects of NMF (obviously biased toward our own interests). In particular, we address the application of NMF for hyperspectral imaging, the geometric interpretation of NMF, complexity issues, algorithms, and the nonnegative rank and its link with extended formulations of polyhedra.

2 Nonnegative Matrix Factorization

The standard NMF problem can be formulated as follows

$$\min_{U \in \mathbb{R}^{p \times r}, V \in \mathbb{R}^{r \times n}} \|M - UV\|_F^2 \text{ such that } U, V \geq 0. \tag{3}$$

As mentioned in the introduction, these nonnegativity constraints allow interpreting the basis elements in the same way as the data (e.g., as image, or vector of word counts) while the nonnegativity of V allows interpreting the weights as activation coefficients. We describe in detail in the next section a particular application, namely, blind hyperspectral unmixing, where the nonnegativity of U and V has a physical interpretation.

The nonnegativity constraints also naturally lead to sparse factors. In fact, the first-order optimality conditions of a problem of the type $\min_{x \geq 0} f(x)$ are $x_i \geq 0$, $\nabla_i f(x) \geq 0$ and $\nabla_i f(x) x_i = 0$ for all i. Hence stationary points of (3) are expected to have zero entries. This property of NMF enhances its interpretability and provides a better compression compared with unconstrained variants.

We refer to the problem of finding an exact factorization, that is, finding $U \geq 0$ and $V \geq 0$ such that $M = UV$, as "exact NMF." The minimum r such that an exact NMF exists is the nonnegative rank of M, denoted $\text{rank}_+(M)$. We have that $\text{rank}(M) \leq \text{rank}_+(M) \leq \min(m, n)$ (since $M = MI = IM$, where I is the identify matrix).

NMF has been used successfully in many applications; see, for example, [25, 42] and the references therein. In the next section we focus on one particular application, namely, blind hyperspectral unmixing.

3 Hyperspectral Imaging

A grayscale image is an image in which the value of each pixel is a single sample. An RGB image has three channels (red, green, and blue) and allows a color image to be reconstructed as it is perceived by an human eye. A hyperspectral image is an image for which usually each pixel has between 100 and 200 channels, corresponding to the reflectance (fraction of light reflected by that pixel) at different wavelengths. The wavelengths measured in a hyperspectral image depend on the camera used and are usually chosen depending on the application at hand. The advantage of hyperspectral images is that they contain much more information, some of it blind to the human eye, that allows one to identify and characterize the materials present in a scene much more precisely; see Figure 1 for an illustration. Its numerous applications include agriculture, eye care, food processing, mineralogy, surveillance, physics, astronomy, chemical imaging, and environmental science; see, for example, `https://en.wikipedia.org/wiki/Hyperspectral_imaging` or `http://sciencenordic.com/lengthy-can-do-list-colour-camera`.

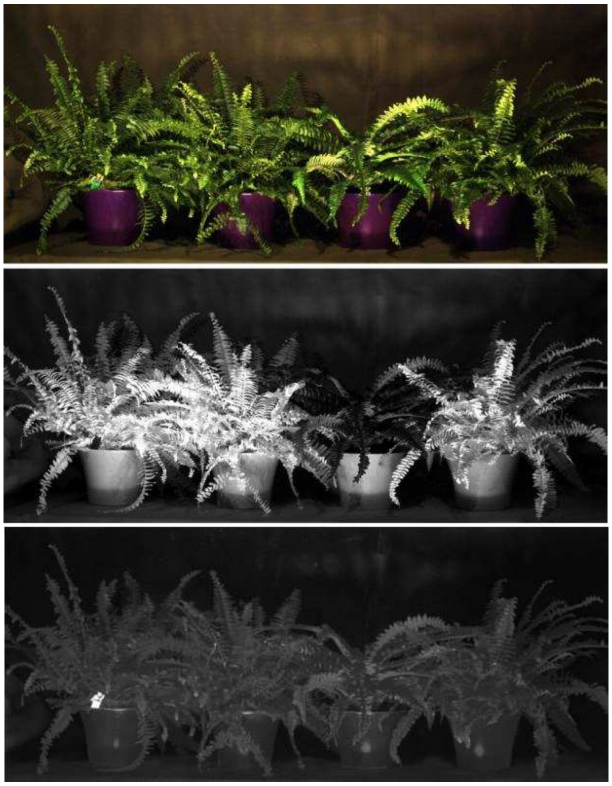

Figure 1: From top to bottom: (1) RGB image of four plants: can you identify the artificial one? (2) Grayscale image at a wavelength that is blind to the naked eye (namely, 770 nm, infrared) and allows identifying the artificial plant (plants have a high reflectance at infrared wavelengths, as opposed to the artificial material). (3) Analysis of the image allows finding a small target, a LEGO figure within the plants. Source: sciencenordic.com, Photo: Torbjørn Skauli, FFI.

Assume a scene is being imaged by a hyperspectral imager using p wavelengths (that is, p channels) and n pixels. Let us construct the matrix $M \in \mathbb{R}^{p \times n}_{+}$ such that $M(i,j)$ is the reflectance of the jth pixel at the ith wavelength. Each column of M therefore corresponds to the so-called spectral signature of a pixel, while each row corresponds to a vectorized image at a given wavelength. Given such an image, an important goal in practice is to (1) identify the constitutive materials present in the image, called endmembers (e.g., grass, trees, road surfaces, roof tops) and (2) classify the pixels accordingly, that is, identify which pixels contain which materials and in which quantity. In fact, the resolution of most hyperspectral images is low, and hence most pixels will contain several materials. If a library or dictionary of spectral signatures of materials present in the image is not available, this problem is referred to as blind hyperspectral unmixing (blind HU): the goal is to identify the endmembers and quantify the abundances of the endmembers in each pixel.

The simplest and most popular model is the linear mixing model (LMM). It assumes that the spectral signature of a pixel equals the weighted linear combination of the spectral signatures of the endmembers it contains, where the weight is given by the abundances. Physically, the reflectance of a pixel will be proportional to the materials it contains: for example, if a pixel contains 30% of aluminum and 70% of copper, its spectral signature will be equal to 0.3 times the spectral signature of the aluminum plus 0.7 times the spectral signature of the copper. In practice, this model is only approximate because of imperfect conditions (measurement noise, light reflecting off several times before being measured, atmospheric distortion, etc.). We refer the reader to [8, 70] for recent surveys on (blind) HU techniques and to [84] for an introduction to hyperspectral imaging.

If we use the LMM and assume that the image contains r endmembers whose spectral signatures are given by the columns of the matrix $U \in \mathbb{R}^{m \times r}_{+}$, we have for all j

$$M(:, j) = \sum_{k=1}^{r} v_{kj} U(:, k) = UV(:, j),$$

where $v_{kj} \geq 0$ is the abundance of the kth endmember in the jth pixel. Therefore, blind HU boils down to the NMF of matrix M; see Figure 2 for an illustration.

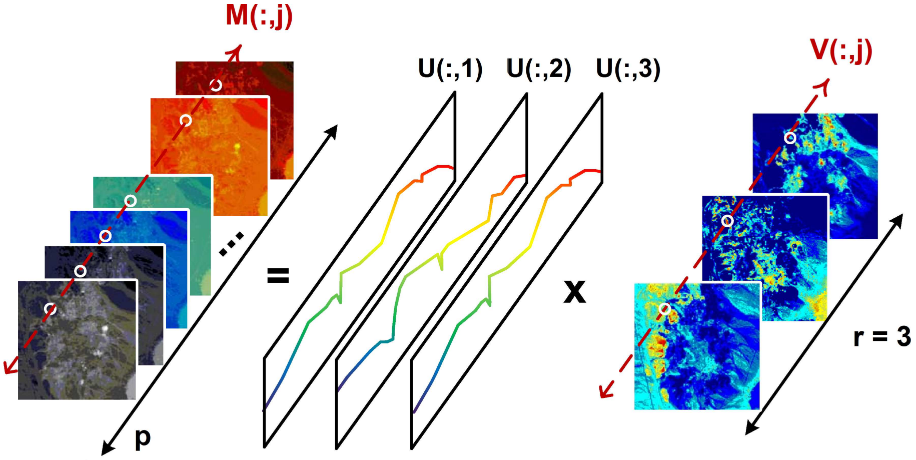

Figure 2: Illustration of the decomposition of a hyperspectral image with three endmembers [70]. On the left, the hyperspectral image M; in the middle, the spectral signatures of the three endmembers as the columns of matrix U; on the right, the abundances of each material in each pixel (referred to as the abundance maps).

Using a standard NMF algorithm, that is, an algorithm that tries to solve (3), will in general not lead to the sought decomposition. The reason is that the solution of NMF is highly nonunique, as discussed later. In practice, a meaningful solution is achieved usually by using additional constraints/penalty terms, including: the sum-to-one constraints on the abundances ($\sum_{k=1}^{r} v_{kj} = 1 \, \forall j$), sparsity of V (because most pixels contain only a few endmembers), piecewise smoothness of the columns of U (since they correspond to spectral signatures), and spatial coherence of the rows of V (because neighboring pixels are more likely to contain the same endmembers). Numerous constrained variants of NMF exist that we do not discuss here; see, for example, [25, 42] and the references therein.

4 Geometry and Uniqueness

NMF has a nice geometric interpretation, which is crucial to consider in order to understand the nonuniqueness of the solutions. As discussed subsequently, it also allows one to develop efficient algorithms and is closely related to the extended formulations of polyhedra.

Let us consider the exact case, that is, $M = UV$. Without loss of generality, (i) the zero columns of M and U can be removed, and (ii) the columns of M and U can be normalized so that the entries of each column sum to one:

$$MD_M^{-1} = UD_U^{-1}D_U V D_M^{-1},$$

where D_M and D_U are diagonal matrices with $D_M(j,j) = \|M(:,j)\|_1$ and $D_U(j,j) = \|U(:,j)\|_1$, respectively. Since we have $M(:,j) = \sum_{k=1}^{r} U(:,k)V(k,j) = UV(:,j)$, this normalization implies that the columns of V also have their entries summing to one, that is, $\|V(:,j)\|_1 = 1$ for all j. Thus that, after normalization, the columns of M belong to the convex hull of the columns of U:

$$M(:,j) \;\in\; \mathrm{conv}(U) \;\subseteq\; \Delta^p = \{x \in \mathbb{R}^p | x \geq 0, \|x\|_1 = 1\} \quad \forall j,$$

where $\mathrm{conv}(U) = \{Ux | x \geq 0, \|x\|_1 = 1\}$. Therefore, the exact NMF problem is equivalent to finding a polytope, $\mathrm{conv}(U)$, nested between two given polytopes, $\mathrm{conv}(M)$ and the unit simplex Δ^p. The dimension of the inner polytope, $\mathrm{conv}(M)$, is $\mathrm{rank}(M) - 1$, while the dimension of the outer polytope, Δ^p, is $p - 1$. The dimension of the nested polytope $\mathrm{conv}(U)$ is not known in advance. When the three polytopes (inner, nested, and outer) have the same dimension, this problem is well known in computational geometry and is referred to as the nested polytope problem (NPP) [27].

If $\mathrm{rank}(M) = \mathrm{rank}(U)$, the column spaces of M and U must coincide, and the outer polytope can be restricted to $\Delta^p \cap \mathrm{col}(M)$, in which case the inner, nested, and outer polytopes have the same dimension. If we impose explicitly this additional constraint ($\mathrm{rank}(M) = \mathrm{rank}(U)$) on the exact NMF problem, we can prove that NPP and this restricted variant of exact NMF are equivalent, that is, they can be reduced to one another [46, 20].

To illustrate, we present a simple example with nested hexagons; this is similar to the example presented in [76]. Let $a > 1$, and let M_a be the matrix

$$\frac{1}{a}\begin{pmatrix} 1 & a & 2a-1 & 2a-1 & a & 1 \\ 1 & 1 & a & 2a-1 & 2a-1 & a \\ a & 1 & 1 & a & 2a-1 & 2a-1 \\ 2a-1 & a & 1 & 1 & a & 2a-1 \\ 2a-1 & 2a-1 & a & 1 & 1 & a \\ a & 2a-1 & 2a-1 & a & 1 & 1 \end{pmatrix}. \tag{4}$$

The restricted exact NMF problem for M_a involves two nested hexagons (recall that we restrict the polytopes to be in the intersection between the column space of M_a and Δ^p, which has dimension 2 since $\mathrm{rank}(M_a) = 3$). Each facet of the outer polytope corresponds to a facet of the nonnegative orthant, that is, to a nonnegativity constraint. The inner hexagon is smaller than the outer one with a ratio of $\frac{a-1}{a}$.

For $a = 2$, the inner hexagon is twice as small as the outer one, and we can fit a triangle between the two so that $\mathrm{rank}_+(M_a) = 3$; see Figure 3 (top). For any $a > 2$, $\mathrm{rank}_+(M_a) \geq 4$ because no triangle can fit between the two hexagons. For $a = 3$, the inner hexagon is 2/3 smaller than the outer one, and we can fit a rectangle between the two and $\mathrm{rank}_+(M_a) = 4$; see Figure 3 (bottom). This implies that $\mathrm{rank}_+(M_a) = 4$ for all $2 < a \leq 3$.

For any $a > 3$, $\mathrm{rank}_+(M_a) = 5$. Surprisingly, the nonnegative rank of M_a is always no more than 5 (even when a tends to infinity, in which case the inner and outer hexagons coincide) because there exists a three-dimensional polytope within Δ^6 with 5 vertices that contains the outer polytope; see

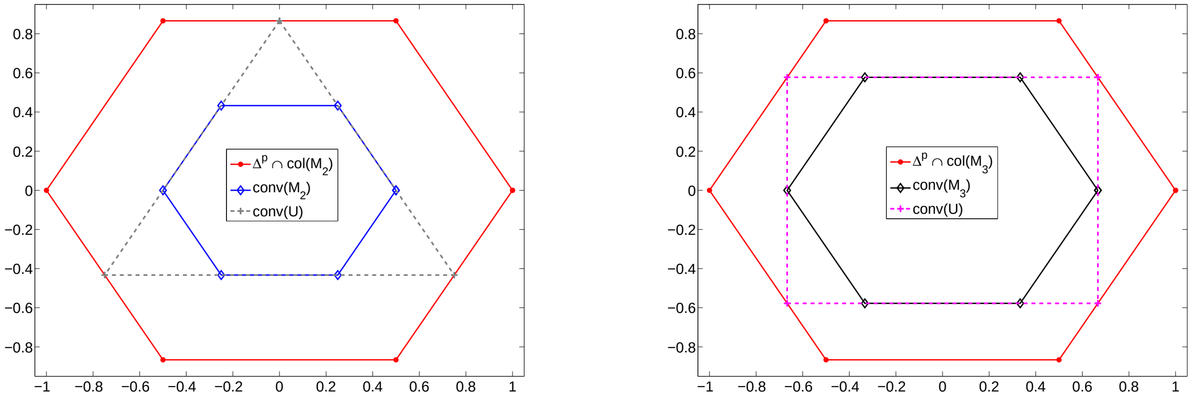

Figure 3: NPP problem corresponding to the exact NMF of the matrix from (4), restricted to the column space of M: (top) the case $a = 2$; (bottom) $a = 3$.

Figure 4, which corresponds to the factorization

$$M = \lim_{a\to+\infty} M_a = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix} = UV = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \tag{5}$$

where $\mathrm{rank}(U) = 4$, and hence $\mathrm{conv}(U)$ has dimension 3.

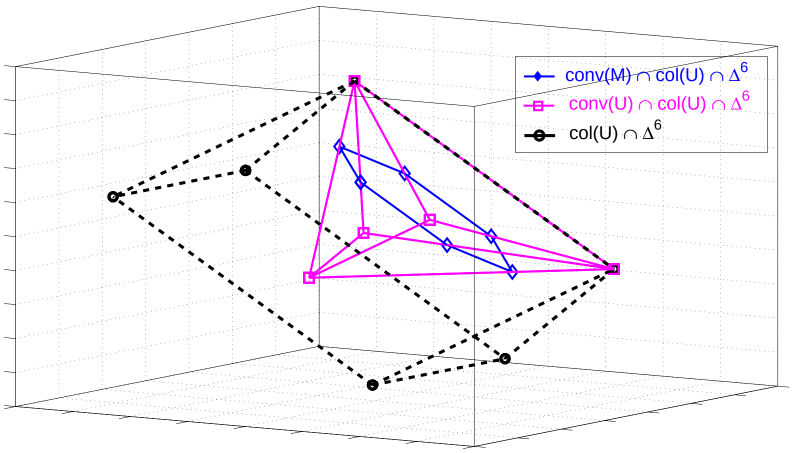

Figure 4: NPP solution corresponding to the exact NMF of the matrix from (5), restricted to the column space of U. It corresponds to the matrix M_a from (4) when $a \to \infty$.

This example illustrates other interesting properties of NMF:

- NMF does not in general have a unique solution (up to scaling and permutation of the rank-one factors). For example, for $a = 2$ (Figure 3, top), four triangles can be fit between the two polytopes (the one shown on the figure, its rotation by 60 degrees, and two triangles whose vertices are three nonadjacent vertices of the outer hexagon). For $1 < a < 2$, this would be even worse since there would be an infinite number of solutions. For this reason, practitioners often add additional constraints to the NMF model to try to identify the most meaningful solution to their problem (such as sparsity, as discussed earlier); see, for example, [63, 40, 54] for more details on the uniqueness of NMF.
- The nonnegative rank can increase only in the neighborhood of a given matrix; that is, the nonnegative rank is upper semicontinuous [10, Th.3.1]: "If P is a nonnegative matrix, without zero columns and with $\text{rank}_+(P) = r$, then there exists a ball $B(P, \epsilon)$ centered at P and of radius $\epsilon > 0$ such that $\text{rank}_+(N) \geq r$ for all $N \in B(P, e)$."

5 Complexity

Given a nonnegative matrix M, checking whether $\text{rank}(M) = \text{rank}_+(M) = r$ is NP hard: unless $P = NP$, there is no polynomial time algorithm in m, n and r for this problem [90]. If r is fixed, however, there is a polynomial time algorithm running in $O\big((pn)^{r^2}\big)$ [5, 75]. The argument is based on quantifier elimination theory (in particular the fact that checking whether a system of ℓ equations in n variables up to degree d can be solved in time polynomial in ℓ and d). Unfortunately, as far as we know, this cannot be used in practice, even for small matrices (e.g., checking whether a 4-by-4 matrix has nonnegative rank 3 seems already impractical with current solvers). Developing an effective code for exact NMF for small matrices is an important direction for further research. Note that we have developed a code based on heuristics that allows solving exact NMF for matrices up to a few dozen rows and columns (although our code comes with no guarantee) [88].

More recently, Shitov [83] and independently Chistikov et al. [21] answered an important open problem showing that the nonnegative rank over the reals might be different from the nonnegative rank over the rationals, implying that the nonnegative rank computation is not in NP since the size of the output is not bounded by the size of the input.

6 Algorithms

In this section, we briefly describe the two main classes of NMF algorithms. As mentioned in the introduction, there does not exist, to the best of our knowledge, a successful convexification approach for NMF, as opposed to other low-rank models. Note, however, that there does exist a convexification approach to compute lower bounds for the nonnegative rank [35]. An explanation is that we cannot work directly with the low-rank approximation $X = UV$ and use the nuclear norm of X, because even if we were given the best nonnegative approximation X of nonnegative rank r for M, in general recovering the exact NMF (U, V) of X would be difficult. Writing directly a convexification in variables (U, V) seems difficult (for rank higher than one[2]) because of the symmetry of the problem (permuting

[2] Note that the rank-one NMF problem is equivalent to the rank-one unconstrained problem since for any rank-one solution uv^T, one can easily check that $|u||v|^T$ is a solution with lower objective function value. This also follows from the Perron-Frobenius and Eckart-Young theorems.

columns of U and rows of V accordingly provides an equivalent solution). Breaking this symmetry seems nontrivial; see [39, pp. 146-148] for a discussion and a tentative SDP formulation. This is an interesting direction for further research.

6.1 Standard nonlinear optimization schemes

As for CLRMA problems, most NMF algorithms use a two-block coordinate descent scheme:

0. Initialize $(U, V) \geq 0$.
1. $U \leftarrow X$, where X solves exactly or approximately $\min_{X \geq 0} \|M - XV\|_F$.
2. $V \leftarrow Y$, where Y solves exactly or approximately $\operatorname{argmin}_{Y \geq 0} \|M - UY\|_F$.

Note that the subproblems to be solved are so-called nonnegative least squares (NNLS). Because NMF is NP hard, these algorithms can only guarantee convergence (usually to a first-order stationary point).

The most well-known algorithm for NMF is the multiplicative updates, namely,

$$U \leftarrow U \circ \frac{[MV^T]}{[UVV^T]}, \quad V \leftarrow V \circ \frac{[U^T M]}{[U^T UV]},$$

where $\circ$ (resp. $\frac{[\,]}{[\,]}$) is the componentwise product (resp. division) between two matrices. It is extremely popular because of its simplicity and because it was proposed in the paper of Lee and Seung [64] that launched the research on NMF. However, it converges slowly; it cannot modify zero entries; and it is not guaranteed to converge to a stationary point. Note that it can be interpreted as a rescaled gradient descent; see, for example, [42].

Methods that try to solve the subproblems exactly are referred to as alternating nonnegative least squares; among these, active set methods seem to be the most efficient, and dedicated codes have been implemented by Haesun Park and collaborators; see [60] and the references therein.

In practice, a method that seems to work extremely well is to apply a few steps of coordinate descent on the NNLS subproblems: the subblocks are the columns of U and the rows of V [24, 45]—the reason is that the subproblems can be solved in closed form. In fact, the optimal kth column of U (all other variables being fixed) is given by

$$\arg \min_{U(:,k) \geq 0} \|R_k - U(:,k)V(k,:)\|_F^2 = \max\left(0, \frac{R_k V(k,:)^T}{\|V(k,:)\|_2^2}\right),$$

for $R_k = M - \sum_{j \neq k} U(:,j)V(j,:)$, and similarly by symmetry for the kth row of V.

Many other approaches can be applied to the NNLS subproblems (e.g., projected gradient method [67], fast/accelerated gradient method (Nesterov's method) [53], and Newton-like method [23]).

6.2 Separable NMF

Although they usually provide satisfactory results in practice, the methods described in the preceding section do not come with any guarantee. In their paper on the complexity of NMF, Arora et al. [5] also identify a subclass of matrices for which the NMF problem is much easier. These are the so-called separable matrices defined as follows.

Definition 1. *A matrix M is separable if there exists a subset $\mathcal{K}$ of r of its columns with $r = \operatorname{rank}_+(M)$ and a nonnegative matrix V such that $M = M(:,\mathcal{K})V$.*

This requires each column of the basis matrix U in an NMF decomposition to be present in the input matrix M. Equivalently, this requires the matrix V in an NMF decomposition to contain the identity matrix as a submatrix. The separable NMF problem is the problem to identify the subset $\mathcal{K}$ (in the noisy case, this subset should be such that $\min_{V \geq 0} \|M - M(:, \mathcal{K})V\|$ is minimized).

Although this condition is strong, it makes sense in several applications, for example the following.

- Document classification: for each topic, there is a "pure" word used only by that topic (an "anchor" word) [4].
- Time-resolved Raman spectra analysis: each substance has a peak in its spectrum while the other spectra are (close to) zero [68].
- Blind hyperspectral unmixing: for each endmember, there exists a pixel that contains only that endmember. This is the so-called pure-pixel assumption that has been used since the 1990s in that community.

Other applications include video summarization [31] and foreground-background separation [62].

Geometrically, in the exact case and after normalization of the columns of X and U, the separability assumption is equivalent to having $\text{conv}(U) = \text{conv}(M)$. Therefore, the so-called separable NMF problem reduces to identify the vertices of the convex hull of the columns of M. This is a relatively easy geometric problem. It becomes tricky when noise is added to the separable matrix, and many recent works have tried to quantify the level of noise that one can tolerate and still be able to recover the vertices, up to some error.

6.2.1 Geometric algorithms

Most algorithms for separable NMF are based on the geometric interpretation, many being developed within the blind HU community (sometimes referred to as pure-pixel search algorithms). Only recently, however, has robustness to noise of these algorithms been analyzed.

One of the simplest algorithm, often referred to as the successive projection algorithm, is closely related to the modified Gram-Schmidt algorithm with column pivoting and has been discovered several times [3, 81, 18]; see the discussion in [70]. Over a polytope, a strongly convex function (such as the ℓ_2 norm) is always maximized at a vertex: this can be used to identify a vertex, that is, a column of U (recall that we assume that the columns of M are normalized so that $\text{conv}(U) = \text{conv}(M)$ under the separability assumption). Once a column of U has been identified, one can project all columns of M onto the orthogonal complement of that column (so that this particular column projects onto 0): this amounts to applying a linear transformation to the polytope. If U is full rank (meaning the polytope is a simplex, which is the case usually in practice), then the other vertices do not project onto 0, and one can use these two steps recursively. This approach is a greedy method to identify a subset of the columns with maximum volume [17, 18]. This algorithm was proved to be robust to noise [49] and can be made more robust to noise by using strategies such as

- applying dimensionality reduction, such as PCA, to the columns of M in order to filter the noise [77];
- using a preconditioning based on minimum-volume ellipsoid [43, 73];
- going over the identified vertices (once r vertices have been identified) to check whether they still maximize the strongly convex function once projected onto the orthogonal complement of the other vertices (otherwise, they are replaced, increasing the volume of the identified vertices) [4];

- taking into account the nonnegativity constraints in the projection step [41].

We refer the reader to [8, 70] for surveys on these approaches. Most geometric approaches for separable NMF are computationally cheap. Usually, however, they are sensitive to outliers.

6.2.2 Convex models

If M is separable, there exist an index set $\mathcal{K}$ of size r and a nonnegative matrix V such that $M = M(:,\mathcal{K})V$. Equivalently, there exists an n-by-n nonnegative matrix X with r nonzero rows such that $M = MX$ with $X(\mathcal{K},:) = V$. Solving separable NMF can therefore be formulated as

$$\min_{X\geq 0} \|X\|_{\text{row},0} \quad \text{such that } M = MX,$$

where $\|X\|_{\text{row},0}$ counts the number of nonzero rows of X. A standard convexification approach is to use the ℓ_1 norm, replacing $\|X\|_{\text{row},0}$ with $\sum_{i=1}^{n} \|X(i,:)\|_k$ for some k; for example, [31] uses $k = \infty$ and [32] uses $k = 2$.

If the columns of M are normalized, the entries of V are bounded above by one (since the columns of U are vertices), and another formulation for separable NMF is obtained:

$$\min_{X\geq 0} \|\operatorname{diag}(X)\|_0 \;\text{ such that }\; M = MX \text{ and } X(i,j) \leq X(i,i) \leq 1\, i,j.$$

Because on each row the diagonal entry has to be the largest and because the goal is to minimize the number of nonzero entries of the diagonal of X, the optimal solution will contain r nonzero diagonal entries and hence r nonzero rows. (Note that requiring the diagonal entries of X to be binary would allow one to model this problem exactly by using mixed-integer linear programming.) Using the ℓ_1 norm, we get another convex model (proposed in [9] and improved in [47]):

$$\min_{X\geq 0} \operatorname{tr}(X) \text{ such that} M = MX \;\text{ and }\; X(i,j) \leq X(i,i) \leq 1\, \forall i,j,$$

where $\operatorname{tr}(X)$ is equal to $\|\operatorname{diag}(X)\|_1$ since X is nonnegative. In practice, when noise is present, the equality term $M = MX$ is replaced with $\|M - MX\| \leq \epsilon$ for some appropriate norm (typically the ℓ_1, ℓ_2, or Frobenius norm) or is added in the objective function as a penalty.

The two models presented above turn out to be essentially equivalent [48]. The main drawback is the computational cost, since these models have n^2 variables. For example, in hyperspectral imaging, n is the number of pixels and is typically on the order of millions; hence, solving these problems is challenging (if not impractical). A natural approach is therefore to first select a subset of good candidates among the columns of M (e.g., using geometric algorithms) and then optimize only over this subset of the rows of X [32, 48]. The main advantage of this approach is that the resulting models are provably the most robust for separable NMF [47]. Intuitively, the reason is not only that the model focuses in identifying, for example, a subset of columns with large volume but also that it requires all the data points to be well approximated with the selected vertices (since $\|M - MX\|$ should be small). For this reason, they are also much less sensitive to outliers than are most geometric approaches.

7 Nonnegative Rank and Extended Formulations

We now describe the link between extended formulations of polyhedra and NMF. This is closely related to the geometric interpretation of NMF described earlier. Let $\mathcal{P}$ be a polytope

$$\mathcal{P} = \{x \in \mathbb{R}^k \mid b_i - A(i,:)x \geq 0 \text{ for } 1 \leq i \leq p\},$$

and let $(w_1, \cdots, w_n)$ be its vertices. Let $S_{\mathcal{P}}$ be the p-by-n slack matrix of $\mathcal{P}$ defined as follows:

$$S_{\mathcal{P}}(i,j) = b_i - A(i,:)w_j \qquad 1 \leq i \leq p, 1 \leq j \leq n.$$

An extended formulation of $\mathcal{P}$ is a higher-dimensional polyhedron $Q \subseteq \mathbb{R}^{k+p}$ that (linearly) projects onto P. The minimum number of facets (that is, inequalities) of such a polytope is called the extension complexity, $\text{xp}(\mathcal{P})$, of $\mathcal{P}$.

Theorem 1. *(Yannakakis, [92]). Let $S_{\mathcal{P}}$ be the slack matrix of the polytope $\mathcal{P}$. Then,* $\text{rank}_+(S_{\mathcal{P}}) =$ *$xp(\mathcal{P})$.*

Let us just show that $\text{xp}(\mathcal{P}) \leq \text{rank}_+(S_{\mathcal{P}})$, because it is elegant and straightforward. Given $\mathcal{P} = \{x \in \mathbb{R}^k \mid b - Ax \geq 0\}$, any exact NMF of $S_{\mathcal{P}} = UV$ with $U \geq 0$ and $V \geq 0$ provides an explicit extended formulation (with some redundant equalities) of $\mathcal{P}$:

$$Q = \{(x, y) \mid b - Ax = Uy \text{ and } y \geq 0\}.$$

In fact, let us show that $Q_x = \{x | \exists y \text{ s.t. } (x, y) \in Q\} = \mathcal{P}$. We have $Q_x \subseteq \mathcal{P}$ since $U \geq 0$ and $y \geq 0$; hence $b - Ax = Uy \geq 0$ for all $(x, y) \in Q$. We have $\mathcal{P} \subseteq Q_x$ because all vertices of $\mathcal{P}$ belong to Q_x: by construction, $(w_j, V(:, j)) \in Q$ since $S_{\mathcal{P}}(:, j) = b - Aw_j = UV(:, j)$ and $V(:, j) \geq 0$.

Example. The extension complexity of the regular n-polygons is $O(\log_2(n))$ [37]. This result can be used to approximate a second-order cone program with a linear program [6]. In particular, we have seen that the extension complexity of the regular hexagon is 5; see Equation (5) and Figure 4.

Recent results. Several recent important results for understanding the limits of linear programming for solving combinatorial problems are based on Theorem 1 and on constructing lower bounds for the nonnegative rank, usually based on the sparsity pattern of the slack matrix [36]; see [58] for a survey. In particular, Rothvoß showed recently that the prefect matching problem cannot be written with polynomially many constraints [82].

These ideas can be generalized in two ways:

- To characterize the size of approximate extended formulations (for a given precision) [13].
- To any convex cone [51], which leads to other CLRMA problems. For example, for the cone of positive semidefinite (PSD) matrices, the rows of U and the columns of V are required to be vectorized PSD matrices. The smallest PSD extension of a given set (e.g., a polyhedron) is equal to the so-called PSD rank of its slack matrix; see the recent survey [34]. (Note that for non-polyhedral sets, the slack matrix is infinite since the number of extreme points and facets is not finite.)

These ideas, for example, recently allowed Hamza Fawzi to prove that the PSD code cannot be represented using the second-order cone [33]; the proof relies on the fact that the second-order cone rank of the cone of 3-by-3 PSD matrices is infinite.

8 Conclusion

In this paper, we have introduced the NMF problem and discussed several of its aspects. The opportunity for meaningful interpretations is the main reason why NMF became so popular and has been used in many applications. NMF is tightly connected with difficult geometric problems; hence developing fast and reliable algorithms is a challenge. Although important challenges remain to be tackled (e.g., developing exact algorithms for small-scale problems), even more challenges exist in generalizations of NMF. In particular, we mentioned cone factorizations (such as the PSD factorization and its symmetric variant [52]), which are more recent problems and have not been explored to their full extent.

Acknowledgments. The author would like to thank the editors of the SIAM Activity Group on Optimization's Views and News (see `http://wiki.siam.org/siag-op/index.php/View_and_News`), Stefan Wild and Jennifer Erway, for providing insightful feedback that helped improve the presentation of the paper. This paper appeared in SIAG/OPT Views and News 25 (1), pp. 7-16. The author also acknowledges the support by the F.R.S.-FNRS (incentive grant for scientific research n^o F.4501.16) and by the ERC (starting grant n^o 679515).

References

[1] Special issue on source separation and application, IEEE Signal Process. Mag. (2014). `http://online.qmags.com/SIPR0514`

[2] Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P., Tandon, R.: Learning sparsely used overcomplete dictionaries. In: COLT, pp. 123–137 (2014)

[3] Araújo, U., Saldanha, B., Galvão, R., Yoneyama, T., Chame, H., Visani, V.: The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr. Intell. Lab. **57**(2), 65–73 (2001)

[4] Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. In: Int. Conf. on Machine Learning (ICML '13), vol. 28, pp. 280–288 (2013)

[5] Arora, S., Ge, R., Kannan, R., Moitra, A.: Computing a nonnegative matrix factorization – provably. In: Proc. of the 44th Symp. on Theory of Computing (STOC '12), pp. 145–162 (2012)

[6] Ben-Tal, A., Nemirovski, A.: On polyhedral approximations of the second-order cone. Math. Oper. Res. **26**(2), 193–205 (2001)

[7] Bhojanapalli, S., Neyshabur, B., Srebro, N.: Global optimality of local search for low rank matrix recovery. arXiv:1605.07221 (2016)

[8] Bioucas-Dias, J., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., Chanussot, J.: Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens. **5**(2), 354–379 (2012)

[9] Bittorf, V., Recht, B., Ré, E., Tropp, J.: Factoring nonnegative matrices with linear programs. In: Adv. Neur. In. (NIPS), pp. 1223–1231 (2012)

[10] Bocci, C., Carlini, E., Rapallo, F.: Perturbation of matrices and nonnegative rank with a view toward statistical models. SIAM J. Matrix Anal. Appl. **32**(4), 1500–1512 (2011)

[11] Boumal, N., Absil, P.A.: RTRMC: A Riemannian trust-region method for low-rank matrix completion. In: Adv. Neur. In. (NIPS), pp. 406–414 (2011)

[12] Boumal, N., Voroninski, V., Bandeira, A.: The non-convex Burer-Monteiro approach works on smooth semidefinite programs. arXiv:1606.04970 (2016)

[13] Braun, G., Fiorini, S., Pokutta, S., Steurer, D.: Approximation limits of linear programs (beyond hierarchies). Math. Oper. Res. **40**(3), 756–772 (2015)

[14] Burer, S., Monteiro, R.: A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Math. Program. **95**(2), 329–357 (2003)

[15] Candès, E., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), 11 (2011)

[16] Candès, E., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)

[17] Çivril, A., Magdon-Ismail, M.: On selecting a maximum volume sub-matrix of a matrix and related problems. Theor. Comput. Sci. **410**(47–49), 4801–4811 (2009)

[18] Chan, T.H., Ma, W.K., Ambikapathi, A., Chi, C.Y.: A simplex volume maximization framework for hyperspectral endmember extraction. IEEE Trans. Geosci. Remote Sens. **49**(11), 4177–4193 (2011)

[19] Chandrasekaran, V., Sanghavi, S., Parrilo, P., Willsky, A.: Rank-sparsity incoherence for matrix decomposition. SIAM J. Optim. **21**(2), 572–596 (2011)

[20] Chistikov, D., Kiefer, S., Marušić, I., Shirmohammadi, M., Worrell, J.: On restricted nonnegative matrix factorization. In: Proceedings of the 43rd International Colloquium on Automata Languages and Programming (ICALP) (2016)

[21] Chistikov, D., Kiefer, S., Marušić, I., Shirmohammadi, M., Worrell, J.: On rationality of nonnegative matrix factorization. In: Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (2017)

[22] Chung, F.: Spectral Graph Theory, vol. 92. American Mathematical Soc. (1997)

[23] Cichocki, A., Zdunek, R., Amari, S.I.: Non-negative matrix factorization with quasi-Newton optimization. In: Lecture Notes in Artificial Intelligence, Springer, vol. 4029, pp. 870–879 (2006)

[24] Cichocki, A., Zdunek, R., Amari, S.I.: Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In: Lecture Notes in Computer Science, Springer, vol. 4666, pp. 169–176 (2007)

[25] Cichocki, A., Zdunek, R., Phan, A., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. John Wiley & Sons (2009)

[26] Clarkson, K., Woodruff, D.: Input sparsity and hardness for robust subspace approximation. In: 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2015) (2015)

[27] Das, G., Joseph, D.: The complexity of minimum convex nested polyhedra. In: Proceedings of the 2nd Canadian Conference on Computational Geometry, pp. 296–301 (1990)

[28] d'Aspremont, A., El Ghaoui, L., Jordan, M., Lanckriet, G.: A direct formulation for sparse PCA using semidefinite programming. SIAM Rev. **49**(3), 434–448 (2007)

[29] Doan, X., Vavasis, S.: Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm. SIAM J. Optim. **26**(1), 274–312 (2016)

[30] Eldén, L.: Matrix Methods in Data Mining and Pattern Recognition, vol. 4. SIAM (2007)

[31] Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: Sparse modeling for finding representative objects. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '12) (2012)

[32] Esser, E., Moller, M., Osher, S., Sapiro, G., Xin, J.: A convex model for nonnegative matrix factorization and dimensionality reduction on physical space. IEEE Trans. Image Process. **21**(7), 3239–3252 (2012)

[33] Fawzi, H.: On representing the positive semidefinite cone using the second-order cone. arXiv:1610.04901 (2016)

[34] Fawzi, H., Gouveia, J., Parrilo, P., Robinson, R., Thomas, R.: Positive semidefinite rank. Math. Program. **153**(1), 133–177 (2015)

[35] Fawzi, H., Parrilo, P.: Lower bounds on nonnegative rank via nonnegative nuclear norms. Math. Program. **153**(1), 41–66 (2015)

[36] Fiorini, S., Kaibel, V., Pashkovich, K., Theis, D.: Combinatorial bounds on nonnegative rank and extended formulations. Discrete Math. **313**(1), 67–83 (2013)

[37] Fiorini, S., Rothvoss, T., Tiwary, H.: Extended formulations for polygons. Discrete Comput. Geom. **48**(3), 658–668 (2012)

[38] Ge, R., Lee, J., Ma, T.: Matrix completion has no spurious local minimum. arXiv:1605.07272 (2016)

[39] Gillis, N.: Nonnegative matrix factorization: Complexity, algorithms and applications. Ph.D. thesis, Université Catholique de Louvain (2011). `https://sites.google.com/site/nicolasgillis/`

[40] Gillis, N.: Sparse and unique nonnegative matrix factorization through data preprocessing. Journal of Machine Learning Research **13**(Nov), 3349–3386 (2012)

[41] Gillis, N.: Successive nonnegative projection algorithm for robust nonnegative blind source separation. SIAM J. Imaging Sci. **7**(2), 1420–1450 (2014)

[42] Gillis, N.: The why and how of nonnegative matrix factorization. In: J. Suykens, M. Signoretto, A. Argyriou (eds.) Regularization, Optimization, Kernels, and Support Vector Machines, pp. 257–291. Chapman & Hall/CRC, Machine Learning and Pattern Recognition Series (2014)

[43] Gillis, N., A. Vavasis, S.: Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. SIAM J. Optim. **25**(1), 677–698 (2015)

[44] Gillis, N., Glineur, F.: Low-rank matrix approximation with weights or missing data is NP-hard. SIAM J. Matrix Anal. Appl. **32**(4), 1149–1165 (2011)

[45] Gillis, N., Glineur, F.: Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. Neural Comput. **24**(4), 1085–1105 (2012)

[46] Gillis, N., Glineur, F.: On the geometric interpretation of the nonnegative rank. Linear Algebra Appl. **437**(11), 2685–2712 (2012)

[47] Gillis, N., Luce, R.: Robust near-separable nonnegative matrix factorization using linear optimization. J. Mach. Learn. Res. **15**(1), 1249–1280 (2014)

[48] Gillis, N., Luce, R.: A fast gradient method for nonnegative sparse regression with self dictionary. arXiv:1610.01349 (2016)

[49] Gillis, N., Vavasis, S.: Fast and robust recursive algorithms for separable nonnegative matrix factorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**(4) (2014)

[50] Gillis, N., Vavasis, S.: On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. arXiv:1509.09236 (2015)

[51] Gouveia, J., Parrilo, P.A., Thomas, R.R.: Lifts of convex sets and cone factorizations. Math. Oper. Res. **38**(2), 248–264 (2013)

[52] Gribling, S., de Laat, D., Laurent, M.: Matrices with high completely positive semidefinite rank. Linear Algebra Appl. **513**, 122–148 (2017)

[53] Guan, N., Tao, D., Luo, Z., Yuan, B.: NeNMF: an optimal gradient method for nonnegative matrix factorization. IEEE Trans. Signal Process. **60**(6), 2882–2898 (2012)

[54] Huang, K., Sidiropoulos, N., Swami, A.: Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Trans. Signal Process. **62**(1), 211–224 (2014)

[55] Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. In: Proceedings of the 45th annual ACM Symposium on Theory of Computing, pp. 665–674. ACM (2013)

[56] Jolliffe, I.: Principal Component Analysis, second edn. Springer (2002)

[57] Journée, M., Nesterov, Y., Richtárik, P., Sepulchre, R.: Generalized power method for sparse principal component analysis. J. Mach. Learn. Res. **11**(Feb.), 517–553 (2010)

[58] Kaibel, V.: Extended formulations in combinatorial optimization. Optima **85**, 2–7 (2011)

[59] Keshavan, R., Oh, S., Montanari, A.: Matrix completion from a few entries. In: 2009 IEEE International Symposium on Information Theory, pp. 324–328. IEEE (2009)

[60] Kim, J., He, Y., Park, H.: Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. Journal of Global Optimization **58**(2), 285–319 (2014)

[61] Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Computer **42**(8), 30–37 (2009)

[62] Kumar, A., Sindhwani, V.: Near-separable non-negative matrix factorization with l1 and Bregman loss functions. In: Proceedings of the 2015 SIAM Int. Conf. on Data Mining, pp. 343–351 (2015)

[63] Laurberg, H., Christensen, M., Plumbley, M., Hansen, L., Jensen, S.: Theorems on positive data: On the uniqueness of NMF. Computational Intelligence and Neuroscience **2008** (2008)

[64] Lee, D., Seung, H.: Learning the parts of objects by nonnegative matrix factorization. Nature **401**, 788–791 (1999)

[65] Lemon, A., So, A.M.C., Ye, Y.: Low-rank semidefinite programming: Theory and applications. Foundations and Trends in Optimization **2**(1-2), 1–156 (2016)

[66] Li, Q., Tang, G.: The nonconvex geometry of low-rank matrix optimizations with general objective functions. arXiv:1611.03060 (2016)

[67] Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**(10), 2756–2779 (2007)

[68] Luce, R., Hildebrandt, P., Kuhlmann, U., Liesen, J.: Using separable non-negative matrix factorization techniques for the analysis of time-resolved Raman spectra. Appl. Spectrosc. **70**(9), 1464–1475 (2016)

[69] Luss, R., Teboulle, M.: Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. SIAM Rev. **55**(1), 65–98 (2013)

[70] Ma, W.K., Bioucas-Dias, J., Chan, T.H., Gillis, N., Gader, P., Plaza, A., Ambikapathi, A., Chi, C.Y.: A signal processing perspective on hyperspectral unmixing. IEEE Signal Process. Mag. **31**(1), 67–81 (2014)

[71] Mahoney, M.W.: Randomized algorithms for matrices and data. Foundations and Trends in Machine Learning **3**(2), 123–224 (2011)

[72] Markovsky, I.: Low Rank Approximation: Algorithms, Implementation, Applications. Springer Science & Business Media (2011)

[73] Mizutani, T.: Robustness analysis of preconditioned successive projection algorithm for general form of separable NMF problem. Linear Algebra Appl. **497**, 1–22 (2016)

[74] Moghaddam, B., Weiss, Y., Avidan, S.: Generalized spectral bounds for sparse LDA. In: Proceedings of the 23rd international conference on Machine learning, pp. 641–648. ACM (2006)

[75] Moitra, A.: An almost optimal algorithm for computing nonnegative rank. In: Proc. of the 24th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA '13), pp. 1454–1464 (2013)

[76] Mond, D., Smith, J., Van Straten, D.: Stochastic factorizations, sandwiched simplices and the topology of the space of explanations. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 459, pp. 2821–2845. The Royal Society (2003)

[77] Nascimento, J., Bioucas-Dias, J.: Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Trans. Geosci. Remote Sens. **43**(4), 898–910 (2005)

[78] Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A.: Non-convex robust PCA. In: Adv. Neur. In. (NIPS), pp. 2080–2088 (2014)

[79] Razenshteyn, I., Song, Z., Woodruff, D.P.: Weighted low rank approximations with provable guarantees. In: Proc. of the 48th Symp. on Theory of Computing (STOC '16), pp. 250–263 (2016)

[80] Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)

[81] Ren, H., Chang, C.I.: Automatic spectral target recognition in hyperspectral imagery. IEEE Trans. Aerosp. Electron. Syst. **39**(4), 1232–1249 (2003)

[82] Rothvoß, T.: The matching polytope has exponential extension complexity. In: Proceedings of the 46th annual ACM Symposium on Theory of Computing, pp. 263–272. ACM (2014)

[83] Shitov, Y.: Nonnegative rank depends on the field II. arXiv:1605.07173 (2016)

[84] Smith, R.: Introduction to hyperspectral imaging (2006). Microimages, `http://www.microimages.com/documentation/Tutorials/hyprspec.pdf`

[85] Song, Z., Woodruff, D.P., Zhong, P.: Low rank approximation with entrywise ℓ_1-norm error. arXiv:1611.00898 (2016)

[86] Srebro, N., Jaakkola, T.: Weighted low-rank approximations. In: ICML, vol. 3, pp. 720–727 (2003)

[87] Udell, M., Horn, C., Zadeh, R., Boyd, S.: Generalized low rank models. Foundations and Trends in Machine Learning **9**(1), 1–118 (2016)

[88] Vandaele, A., Gillis, N., Glineur, F., Tuyttens, D.: Heuristics for exact nonnegative matrix factorization. J. of Global Optim. **65**(2), 369–400 (2016)

[89] Vandereycken, B.: Low-rank matrix completion by Riemannian optimization. SIAM J. Optim. **23**(2), 1214–1236 (2013)

[90] Vavasis, S.: On the complexity of nonnegative matrix factorization. SIAM J. Optim. **20**(3), 1364–1377 (2010)

[91] Woodruf, D.: Sketching as a tool for numerical linear algebra. Foundations and Trends in Machine Learning **10**(1-2), 1–157 (2014)

[92] Yannakakis, M.: Expressing combinatorial optimization problems by linear programs. J. Comput. Syst. Sci. **43**(3), 441–466 (1991)