KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Debals, O., Van Barel, M., De Lathauwer, L. (2017), **Nonnegative Matrix Factorization using Nonnegative Polynomial Approximations** IEEE Signal Processing Letters, vol. 24, no. 7, p. 948-952.
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	http://dx.doi.org/10.1109/LSP.2017.2697680
Journal homepage	http://signalprocessingsociety.org/publications-resources/ieee-signal-processing-letters/ieee-signal-processing-letters
Author contact	Otto.Debals@esat.kuleuven.be *+ 32 (0)16 3 20364*
IR	/

(article begins on next page)

Nonnegative Matrix Factorization using Nonnegative Polynomial Approximations

Otto Debals, *Student Member, IEEE*, Marc Van Barel, *Member, IEEE*,
and Lieven De Lathauwer, *Fellow, IEEE*

***Abstract*—Nonnegative matrix factorization is a key tool in many data analysis applications such as feature extraction, compression and noise filtering. Many existing algorithms impose additional constraints to take into account prior knowledge and to improve the physical interpretation. This paper proposes a novel algorithm for nonnegative matrix factorization in which the factors are modeled by nonnegative polynomials. Using a parametric representation of finite-interval nonnegative polynomials, we obtain an optimization problem without external nonnegativity constraints which can be solved using conventional quasi-Newton or nonlinear least squares methods. The polynomial model guarantees smooth solutions and may realize a noise reduction. A dedicated orthogonal compression enables a significant reduction of the matrix dimensions, without sacrificing accuracy. The overall approach scales well to large matrices. The approach is illustrated with applications in hyperspectral imaging and chemical shift brain imaging.**

***Index Terms*—Nonnegative matrix factorization, Polynomial approximation, Nonnegative polynomials**

I. Introduction

Matrix factorization techniques compress or analyze a given data matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$ using the bilinear model $\mathbf{X} = \mathbf{W}\mathbf{H}^{\mathrm{T}}$ with $\mathbf{W} \in \mathbb{R}^{I \times R}, \mathbf{H} \in \mathbb{R}^{J \times R}$ and typically $R \ll I, J$. The factor vectors or components in $\mathbf{W}$ and $\mathbf{H}$ can subsequently be used to, e.g., predict missing entries in $\mathbf{X}$ [1] or identify source signals in a blind source separation (BSS) context [2], [3]. Many signals and data are nonnegative by nature such as amplitude spectra, pixel intensities and occurrence counts. Imposing the assumption of non-negativity on $\mathbf{W}$ and $\mathbf{H}$ results in nonnegative matrix factorization (NMF), giving an additive parts-based representation of $\mathbf{X}$ [4]. It is well known that the general problem of NMF does not return unique components [5], [6]. To facilitate the interpretation of the results and to exploit prior knowledge (but not necessarily completely solve the uniqueness problem), many existing NMF techniques impose additional constraints such as sparsity [6]–[8], smoothness [9]–[11] and orthogonality [12].

Multiple NMF algorithms have been developed using two-block coordinate descent approaches (alternatively solving for one factor matrix while keeping the other matrix fixed, resulting in convex subproblems) such as the multiplicative update (MU) [13], alternating least squares (ALS) [14] and alternating nonnegative least squares methods [15]–[17]. Others use a more direct approach by updating both $\mathbf{W}$ and $\mathbf{H}$ in the same step based on projected gradients (PG) [18] or using some parametrization free of non-negativity constraints [19]. It is known that many NMF algorithms are sensitive to initialization and can result in non-smooth solutions [20]–[22]. Second, it can be seen that the number of optimization variables $(I+J)R$ can be large for large dimension sizes I, J.

In [19], the parametrizations $\mathbf{W} = \mathbf{E} * \mathbf{E}$ and $\mathbf{H} = \mathbf{F} * \mathbf{F}$ are used, with $\mathbf{E} \in \mathbb{R}^{I \times R}$, $\mathbf{F} \in \mathbb{R}^{J \times R}$ and $*$ the Hadamard product. The technique yields an optimization problem without non-negativity constraints which can be solved with standard quasi-Newton or non-linear least squares optimization techniques. The approach we propose is also parametrization-based. Rather than parametrizing each variable separately, each entire factor vector (in one or both factor matrices) is represented and approximated by a nonnegative polynomial. This makes sense in various applications, as many (smooth) signals can be well approximated by polynomials with a low degree such as emission or frequency spectra and sinusoidal signals. The approach is related to novel BSS techniques where source and/or mixing vectors are approximated by low-parametric deterministic models such as in [23]–[26].

By using a particular parametrization, the complete set of nonnegative polynomials on a finite interval of interest can be modeled. Smooth results are guaranteed without the use of additional smoothing penalization or regularization terms as in [10], [11], and the number of optimization variables is significantly reduced which is beneficial for large-scale NMF. It is tempting in large-scale NMF to reduce the dimensionality of the problem by a singular value decomposition (SVD). However, the nonnegativity is lost using the orthogonal compression. The nonnegative polynomial approach, on the other hand, does admit a dedicated loss-free orthogonal compression. By carefully exploiting the structure in the resulting

Copyright © 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This research is funded by (1) a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT), (2) Research Council KU Leuven: C1 project c16/15/059-nD, (3) F.W.O.: projects G.0830.14N, G.0881.14N and G.0828.14N, (4) the Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO II, Dynamical systems, control and optimization, 2012-2017), (5) EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Advanced Grant: BIOTENSORS (no. 339804). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

O. Debals and L. De Lathauwer are with the Group of Science, Engineering and Technology, KU Leuven Kulak, E. Sabbelaan 53, B-8500 Kortrijk, Belgium and with the Department of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. (e-mail: otto.debals@kuleuven.be, lieven.delathauwer@kuleuven.be).

M. Van Barel is with the Department of Computer Science, KU Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium (e-mail: marc.vanbarel@kuleuven.be).

Manuscript received March 03, 2017; revised April 11, 2017, 2015; accepted April 12, 2017. Date of publication XXX YY, 2017.

problems, efficient algorithms are obtained. These can be used as such or, when polynomials just give a first approximation, to initialize general-purpose NMF algorithms. Note that the technique should not be confused with the fundamentally different polynomial kernel NMF technique from [27].

Notation: Parameters, scalars, vectors and matrices are denoted by upper, lower, bold lower and bold upper case characters, e.g., N, a, $\mathbf{a}$ and $\mathbf{A}$, resp. The Hadamard, Kronecker, column-wise Khatri–Rao and row-wise Khatri–Rao products are denoted with $*$, $\otimes$, $\odot$ and $\odot^{\mathrm{T}}$, resp. The columns of $\mathbf{A}\odot\mathbf{B}$ (resp., rows of $\mathbf{A}\odot^{\mathrm{T}}\mathbf{B}$) are the pairwise Kronecker products of the columns (resp., rows) of $\mathbf{A}$ and $\mathbf{B}$. $[\mathbf{A};\ \mathbf{B}]$ denotes the vertical concatenation of matrices $\mathbf{A}$ and $\mathbf{B}$. The transpose, inverse, pseudoinverse, element-wise square and Frobenius norm are denoted by $\cdot^{\mathrm{T}}$, $\cdot^{-1}$, $\cdot^{\dagger}$, $\cdot^{*2}$ and $||\cdot||$, resp.

II. Nonnegative polynomial-based NMF

Section II-A discusses how to model nonnegative polynomials on infinite and finite intervals. These models are then interpreted from a signal processing point of view in Section II-B and applied to the NMF setting in Section II-C.

A. Modeling nonnegative polynomials on (in)finite intervals

A degree-D polynomial $f(t)$ is obviously nonnegative for $t \in [-\infty, \infty]$ if it can be written as a squared polynomial, i.e., $f(t) = p(t)^2$. While this representation is not sufficient to model the entire set of nonnegative polynomials on the real axis, the sum-of-squares (SOS) representation is [28]:

$$\forall t \in \mathbb{R} : f(t) \geq 0 \ \Leftrightarrow \ f(t) = \sum\nolimits_{k=1}^{K} p_k(t)^2. \qquad (1)$$

It is well known that $K = 2$ is sufficient, and each $p_k(t)$ has degree at most $D/2$. Signals of finite rather than infinite length are typically considered in applications, which makes the representation in (1) too restrictive for our purposes. This can be understood by considering the class of odd-degree polynomials and more specifically the example $f(t) = t + 2$, for which $f(t) \geq 0$ for $t \in [-1, 1]$ but $f(t) < 0$ for $t < -2$.

A fundamental representation exists in the literature for nonnegative polynomials on a finite interval. We limit the analysis to $t \in [-1, 1]$, which can be generalized to $t \in [a, b]$ using the change in variables $\tilde{f}(t) = f(\frac{2t-(a+b)}{b-a})$. Then [28], [29]:

$$\forall t \in [-1, 1] : f(t) \geq 0 \ \Leftrightarrow \ f(t) = f_1(t) + g(t) f_2(t) \qquad (2)$$

with $g(t) = 1 - t^2$ and the degree of $f_1(t)$ and $f_2(t)$ not exceeding $2D$ and $2D - 2$, resp. In [28] it is discussed that $f_1(t)$ and $f_2(t)$ should be SOS as in (1), but judging from the original work [29]–[31] it suffices that $f_1(t)$ and $f_2(t)$ are squared polynomials. Note that the representation is not unique in general, e.g., it is possible that $f_1(t) + g(t)f_2(t) = f_3(t) + g(t)f_4(t)$ for $f_1(t) \neq f_3(t)$ and $f_2(t) \neq f_4(t)$.

B. Connection with discrete-time signals

Let us first consider the discrete-time signal $\mathbf{p} \in \mathbb{R}^I$ with $p_i = p(t_i)$ for $i = 1, \ldots, I$ and with $p(t)$ a degree-D polynomial. The points t_i do not need to be equidistant. It is clear that one can write $\mathbf{p} = \mathbf{V}\mathbf{c}$ for some evaluated polynomial basis $\mathbf{V} \in \mathbb{R}^{I\times(D+1)}$ and corresponding coefficients $\mathbf{c} \in \mathbb{R}^{D+1}$. For the standard monomial basis, $\mathbf{V}$ has Vandermonde structure. Using a Chebyshev or Legendre basis instead can avoid ill-conditioned situations. An evaluated squared polynomial $\mathbf{f} = \mathbf{p}^{*2}$ then results in $\mathbf{f} = (\mathbf{V}\mathbf{c})^{*2} = (\mathbf{V}\odot^{\mathrm{T}}\mathbf{V})(\mathbf{c}\odot\mathbf{c})$. Furthermore, from the representation in (2), an evaluated nonnegative polynomial on a finite interval can be represented by

$$\mathbf{f} = (\mathbf{V}\odot^{\mathrm{T}}\mathbf{V})(\mathbf{a}\odot\mathbf{a}) + (\mathbf{U}\odot^{\mathrm{T}}\mathbf{U})(\mathbf{b}\odot\mathbf{b}), \qquad (3)$$

with matrices $\mathbf{V} \in \mathbb{R}^{I\times(D+1)}, \mathbf{U} \in \mathbb{R}^{I\times D}$, and with coefficient vectors $\mathbf{a} \in \mathbb{R}^{D+1}, \mathbf{b} \in \mathbb{R}^{D}$. The weighting function $g(t)$ is absorbed in $\mathbf{U}$, i.e., $\forall i, d : u_{id} = \sqrt{1 - t_i^2}\, p_{id}$ with $\mathbf{P} \in \mathbb{R}^{I\times D}$ some evaluated polynomial basis of degree $D-1$.

C. NMF using nonnegative polynomials

The NMF problem consists of finding the factorization $\mathbf{X} = \mathbf{W}\mathbf{H}^{\mathrm{T}}$ such that $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$ with element-wise nonnegativity. Assuming that each factor vector in $\mathbf{W}$ can be well approximated by a finite-interval nonnegative polynomial, the model from (3) can be used. With $\mathbf{A} \in \mathbb{R}^{(D+1)\times R}$ and $\mathbf{B} \in \mathbb{R}^{D\times R}$, this results in the following single-sided version:

$$\begin{aligned}\mathbf{X} &= \left[(\mathbf{V}\odot^{\mathrm{T}}\mathbf{V})(\mathbf{A}\odot\mathbf{A}) + (\mathbf{U}\odot^{\mathrm{T}}\mathbf{U})(\mathbf{B}\odot\mathbf{B})\right]\mathbf{H}^{\mathrm{T}} \\ &= \mathbf{Q}\begin{bmatrix}\mathbf{A}\odot\mathbf{A}\\ \mathbf{B}\odot\mathbf{B}\end{bmatrix}\mathbf{H}^{\mathrm{T}} \ \text{ with } \mathbf{Q} = \begin{bmatrix}\mathbf{V}\odot^{\mathrm{T}}\mathbf{V} & \mathbf{U}\odot^{\mathrm{T}}\mathbf{U}\end{bmatrix}, \qquad (4)\end{aligned}$$

and $\mathbf{H} \geq 0$. The number of variables in $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{H}$ is reduced from $(I+J)R$ to $(2D+1+J)R$. Furthermore, using the model in (3) also for the factor vectors in $\mathbf{H}$, one obtains the following double-sided polynomial-based NMF model:

$$\mathbf{X} = \mathbf{Q}^{(w)}\begin{bmatrix}\mathbf{A}\odot\mathbf{A}\\ \mathbf{B}\odot\mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{C}\odot\mathbf{C}\\ \mathbf{D}\odot\mathbf{D}\end{bmatrix}^{\mathrm{T}}\mathbf{Q}^{(h)^{\mathrm{T}}}, \qquad (5)$$

with $\mathbf{a}_r, \mathbf{b}_r$ and $\mathbf{c}_r, \mathbf{d}_r$ the coefficients related to each factor vector $\mathbf{w}_r$ and $\mathbf{h}_r$, resp., and $\mathbf{Q}^{(w)}, \mathbf{Q}^{(h)}$ similar matrices as in (4). The number of variables is reduced to $(2D^{(w)} + 2D^{(h)} + 2)R$, with $D^{(w)}$ (resp., $D^{(h)}$) the degrees of the polynomials used to approximate each factor vector $\mathbf{w}_r$ (resp., $\mathbf{h}_r$).

There exist NMF techniques that make use of a linear model $\mathbf{W} = \mathbf{V}\mathbf{A}$ to represent one or both factor matrices, with $\mathbf{V}$ containing some basis functions such as radial basis functions and with $\mathbf{A}$ containing the corresponding coefficients [9]. While the latter methods require explicit non-negativity constraints on $\mathbf{V}\mathbf{A}$ and/or $\mathbf{A}$, the constraints are implicitly imposed in (5). Second, results of these methods highly depend on the function parameters in $\mathbf{V}$, while the proposed approach only requires predefined degrees $D^{(w)}, D^{(h)}$.

III. Algorithms and computational aspects

A. Quasi-Newton and nonlinear least squares algorithms

Let us consider the single-sided nonnegative polynomial-based NMF (NP-NMF) model in (4), with $\mathbf{H} = \mathbf{F}^{*2}$ with $\mathbf{F} \in \mathbb{R}^{J\times R}$ to impose the non-negativity constraint [19]. Given data matrix $\mathbf{X}$, one can find matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{F}$ by minimizing the following least-squares objective function:

$$J = \tfrac{1}{2}\,||\mathbf{X} - \mathbf{W}\mathbf{H}^{\mathrm{T}}||^2 = \tfrac{1}{2}\,||\mathbf{R}||^2, \qquad (6)$$

with residual matrix $\mathbf{R} = \mathbf{Q}\left[(\mathbf{A}\odot\mathbf{A});(\mathbf{B}\odot\mathbf{B})\right]\left(\mathbf{F}^{*2}\right)^{\mathrm{T}} - \mathbf{X}$.

While NMF methods are typically based on two-block coordinate descent or alternating least squares (ALS, linear convergence) approaches, we consider more direct techniques to minimize (6) by updating the estimates in the same iteration. These so-called all-at-once algorithms such as quasi-Newton (QN, superlinear convergence) and nonlinear least squares (NLS, close to quadratic convergence) have been shown to be more robust to overfactoring and to outperform ALS-based algorithms for ill-conditioned cases, although requiring a higher computational load per iteration [32], [33]. For a further discussion, we refer the reader to more detailed literature [34].

QN methods require the expression of the gradient $\mathbf{g}$:

$$\frac{\partial J}{\partial \mathbf{a}_r} = ((\mathbf{I}_{D+1}\otimes\mathbf{a}_r) + (\mathbf{a}_r\otimes\mathbf{I}_{D+1}))^{\mathrm{T}}(\mathbf{V}\odot^{\mathrm{T}}\mathbf{V})^{\mathrm{T}}\mathbf{R}\mathbf{f}_r^{*2} \quad (7)$$

$$= 2\mathbf{V}^{\mathrm{T}}\mathrm{diag}(\mathbf{V}\mathbf{a}_r)\mathbf{R}\mathbf{f}_r^{*2}, \qquad r = 1,\ldots,R, \quad (8)$$

$$\frac{\partial J}{\partial \mathbf{b}_r} = ((\mathbf{I}_D\otimes\mathbf{b}_r) + (\mathbf{b}_r\otimes\mathbf{I}_D))^{\mathrm{T}}(\mathbf{U}\odot^{\mathrm{T}}\mathbf{U})^{\mathrm{T}}\mathbf{R}\mathbf{f}_r^{*2}$$

$$= 2\mathbf{U}^{\mathrm{T}}\mathrm{diag}(\mathbf{U}\mathbf{b}_r)\mathbf{R}\mathbf{f}_r^{*2}, \qquad r = 1,\ldots,R, \quad (9)$$

$$\frac{\partial J}{\partial \mathbf{F}} = 2\mathbf{F} * \left(\mathbf{R}^{\mathrm{T}}\mathbf{Q}\left[(\mathbf{A}\odot\mathbf{A});(\mathbf{B}\odot\mathbf{B})\right]\right),$$

in which $\mathrm{diag}(\mathbf{e})$ returns a $I\times I$ diagonal matrix with diagonal $\mathbf{e}\in\mathbb{R}^I$, and in which $\mathbf{I}_I$ is the $I\times I$ identity matrix. To approximate the Hessian by its Gramian $\mathbf{G} = \mathbf{J}^{\mathrm{T}}\mathbf{J}$, NLS methods additionally require the Jacobian matrix $\mathbf{J}$:

$$\partial\,\mathrm{vec}\,(\mathbf{R})/\partial\,\mathbf{a}_r^{\mathrm{T}} = (\mathbf{f}_r^{*2})\otimes(\mathrm{diag}(\mathbf{V}\mathbf{a}_r)\mathbf{V}),$$

$$\partial\,\mathrm{vec}\,(\mathbf{R})/\partial\,\mathbf{b}_r^{\mathrm{T}} = (\mathbf{f}_r^{*2})\otimes(\mathrm{diag}(\mathbf{W}\mathbf{b}_r)\mathbf{W}),$$

$$\partial\,\mathrm{vec}\,(\mathbf{R})/\partial\,\mathbf{f}_r^{\mathrm{T}} = \mathbf{I}_N\odot\left(\mathbf{Q}\left[(\mathbf{a}_r\odot\mathbf{a}_r);(\mathbf{b}_r\odot\mathbf{b}_r)\right]\mathbf{f}_r^{\mathrm{T}}\right).$$

These expressions can then be plugged into a numerical optimization solver. Evaluating the objective function and calculating the gradient both cost $\mathcal{O}(I(J+D)R)$ flop, which reduces to $\mathcal{O}(IJR)$ flop as typically $J \gg D$. Rather than constructing and inverting the possibly large Gramian $\mathbf{G}$, fast matrix-vector products $\mathbf{J}^{\mathrm{T}}\mathbf{J}\mathbf{p}$ are used in inexact NLS approaches to iteratively solve the system $\mathbf{G}\mathbf{p} = -\mathbf{g}$. These matrix-vector products and corresponding (block-)diagonal preconditioner can be computed in $\mathcal{O}(IJR)$ flop as well. This order of complexity is equal to that of many standard NMF approaches, e.g., MU or accelerated hierarchical ALS [35]–[37]; recall also the convergence properties mentioned above. In Section III-B, we will further reduce the cost per iteration.

The gradient vector and Jacobian matrix corresponding to the objective function in (6) given the double-sided model from (5) can be derived in a similar way.

B. Preprocessing with orthogonal compression

Consider the double-sided NP-NMF model in (5). As the typically tall matrices $\mathbf{Q}^{(w)} \in \mathbb{R}^{I\times(2D^{(w)^2}+2D^{(w)}+1)}$ and $\mathbf{Q}^{(h)}\in\mathbb{R}^{J\times(2D^{(h)^2}+2D^{(h)}+1)}$ are known beforehand, a significant compression is possible. Assuming that $I \geq 2D^{(w)}+1$ and $J\geq 2D^{(h)}+1$, it can be shown that $\mathbf{Q}^{(w)}$ and $\mathbf{Q}^{(h)}$ are rank-deficient and have rank $2D^{(w)}+1$ and $2D^{(h)}+1$, resp., independent of the polynomial basis type. Let us compute the corresponding reduced SVDs $\mathbf{Q}^{(w)} = \mathbf{Y}^{(w)}\mathbf{\Sigma}^{(w)}\mathbf{Z}^{(w)^{\mathrm{T}}}$ and $\mathbf{Q}^{(h)} = \mathbf{Y}^{(h)}\mathbf{\Sigma}^{(h)}\mathbf{Z}^{(h)^{\mathrm{T}}}$. Since $\mathbf{Y}^{(w)}$ and $\mathbf{Y}^{(h)}$ are column-wise orthonormal, the double-sided version of (6) can without loss of accuracy be expressed as:

$$J = \frac{1}{2}\left\|\tilde{\mathbf{X}} - \left(\mathbf{\Sigma}^{(w)}\mathbf{Z}^{(w)}\right)\begin{bmatrix}\mathbf{A}\odot\mathbf{A}\\ \mathbf{B}\odot\mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{C}\odot\mathbf{C}\\ \mathbf{D}\odot\mathbf{D}\end{bmatrix}^{\mathrm{T}}\left(\mathbf{\Sigma}^{(h)}\mathbf{Z}^{(h)}\right)^{\mathrm{T}}\right\|^2,$$

with $\tilde{\mathbf{X}} = \mathbf{Y}^{(w)^{\mathrm{T}}}\mathbf{X}\mathbf{Y}^{(h)^{\mathrm{T}}} \in \mathbb{R}^{(2D^{(w)}+1)\times(2D^{(h)}+1)}$ the compressed data matrix which needs to be computed only once. While the dimensions are significantly reduced, the Khatri–Rao structure in $\mathbf{Z}^{(w)}, \mathbf{Z}^{(h)}$ is lost. Consequently, we cannot use the simplifications in, e.g., (8) and (9). Instead, $\partial J/\partial\,\mathbf{a}_r$ is then computed using an expression as in (7) but with $\mathbf{V}\odot^{\mathrm{T}}\mathbf{V}$ replaced by the first $(D^{(w)}+1)^2$ columns of $\mathbf{\Sigma}^{(w)}\mathbf{Z}^{(w)}$; like-wise for $\partial J/\partial\,\mathbf{b}_r$, $\partial\,\mathrm{vec}\,(\mathbf{R})/\partial\,\mathbf{a}_r^{\mathrm{T}}$ and $\partial\,\mathrm{vec}\,(\mathbf{R})/\partial\,\mathbf{b}_r^{\mathrm{T}}$. Nevertheless, the complexity is reduced from $\mathcal{O}(IJR)$ to $\mathcal{O}((2D^{(w)}+1)(2D^{(h)}+1)R)$ flop per iteration. This is significant if the number of samples is large or if the components can be well approximated by polynomials of low degree.

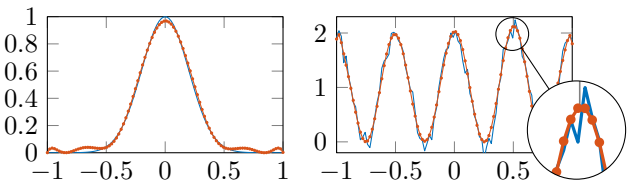


Fig. 1: A Gaussian function $f(t) = e^{-\frac{1}{2}(5t)^2}$ (left) and a sinusoidal function $f(t) = 1+\cos(4\pi t)$ with 20 dB SNR (right). Both signals are sampled uniformly 100 times in $[-1,1]$ (——) and approximated by finite-interval nonnegative polynomials (——) of degrees 6 (left) and 10 (right).

IV. Simulations and Applications

In this section, we present some results illustrating the modeling power of nonnegative polynomials for synthetic data and for two real-life applications, together with a timing comparison of the proposed NMF techniques. Unless stated otherwise, the NLS version of NP-NMF with compression step, Chebyshev basis, diagonal preconditioner, random initial coefficients and 50 iterations is used. The polynomial degrees are chosen by visual inspection and with some trial-and-error. Altering the degrees changes the possible number of local extrema and alters the smoothing effect. The complex optimization toolbox [33], [38] is used as numerical optimization solver. Additive Gaussian noise is used. We denote NMF without polynomial structure as 'general NMF'. Unless stated otherwise, it is solved using the accelerated hierarchical ALS (aHALS) technique [36], [39].

A first experiment in Fig. 1 considers the approximation of two basic functions. A relative fitting error of 0.047 and 0.092 is obtained, illustrating that nonnegative polynomials can model a variety of shapes and also have a denoising effect.

In a second experiment, we apply NP-NMF to the spectral unmixing problem for non-resolved space object characterization [40]–[42]. The spectral signatures (wavelength-dependent absorption features) of space object materials (endmembers) such as aluminum, mylar, paint and silicon from solar cells mix together in the spectral reflectance measurements of the entire object. Given a hyperspectral data matrix $\mathbf{X}$ with I spectral bands and J spectral measurements, the goal is

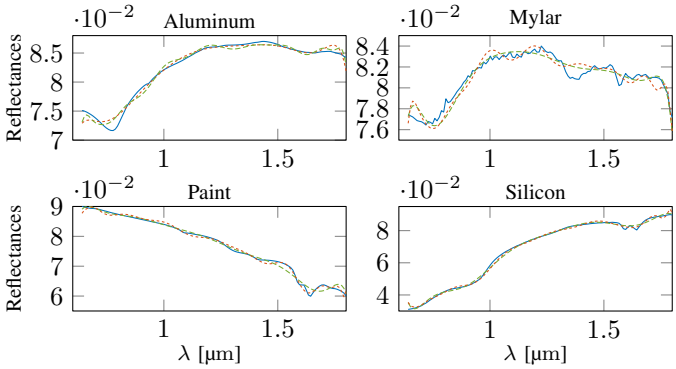


Fig. 2: The endmembers signatures in the second experiment (——) and aligned NP-NMF estimates using degrees 6 (------) and degrees 12 (·····).

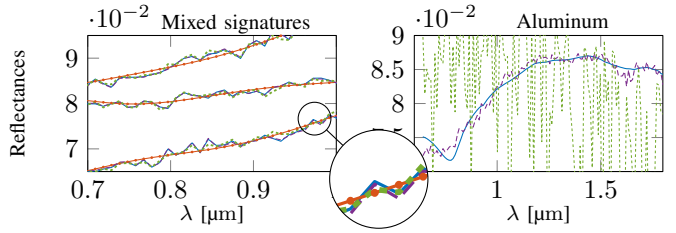


Fig. 3: (Left) Zoom-in on the mixed signatures in $\mathbf{X}$, with the original signatures (——), fitted NP-NMF signatures with $D^{(w)} = 12$ (——) and fitted aHALS signatures. For the latter, two initializations for $\mathbf{W}$ are used: the same initial $\mathbf{W}$ as in the NP-NMF experiment, i.e., a degree-12 nonnegative polynomial with random coefficients (------), and a random nonnegative $\mathbf{W}$ without polynomial structure (·····). (Right) Aluminum signatures estimated by aHALS, to be compared with the NP-NMF estimates in Fig. 2 (top, left).

to identify the different endmembers in $\mathbf{W}$ and fractional abundances in $\mathbf{H}$ using the NMF model $\mathbf{X} = \mathbf{W}\mathbf{H}^{\mathrm{T}}$. Let us consider the real-life spectral signatures of $R = 4$ endmembers in the 0.6-1.8 μm range with $I = 120$ spectral bands in Fig. 2 [41]. The spectral measurements in $\mathbf{X}$ are constructed by using a matrix $\mathbf{H} \in \mathbb{R}^{4\times 4}$ with entries drawn uniformly from [0,1]. Noise is added to $\mathbf{X}$ in the simulations with a signal-to-noise ratio (SNR) of 40 dB [43]. The median relative fitting error ϵ_f is reported on the noiseless $\mathbf{W}\mathbf{H}^{\mathrm{T}}$ across 10 runs with different initializations. The single-sided NP-NMF method is applied with degrees 6 and 12, each with 30 iterations ($\epsilon_f = 0.0062$ and 0.0049, resp.). General NMF is applied with two types of initializations and 300 iterations ($\epsilon_f = 0.0088$ and 0.0079, resp.). Fig. 3 (left) shows that NP-NMF yields smoother estimates than general NMF and is less prone to noise fitting.

Without additional assumptions or constraints, the NMF components cannot be uniquely determined [5], [6]. To be able to compare the estimates, we determine the matrix $\mathbf{Q} \in \mathbb{R}^{R\times R}$ such that $\tilde{\mathbf{W}}\mathbf{Q}$ approximates $\mathbf{W}$ in optimal least-square sense, with $\tilde{\mathbf{W}}$ the estimate of $\mathbf{W}$. The aligned estimates $\tilde{\mathbf{W}}\mathbf{Q}$ obtained from NP-NMF are given in Fig. 2; for comparison, the aligned estimated aluminum signature obtained from general NMF is given in Fig. 3. The signatures are well approximated by low-degree nonnegative polynomials, i.e., the latter constitute a good class of models. Second, although general NMF yields small fitting errors ϵ_f, the component estimates can be highly non-smooth and difficult to interpret.

A third experiment considers ^{31}P chemical shift imaging data of the human brain [43]–[45]. The data set consists of $J = 512$ spectra measured on a $8 \times 8 \times 8$ grid in the brain,

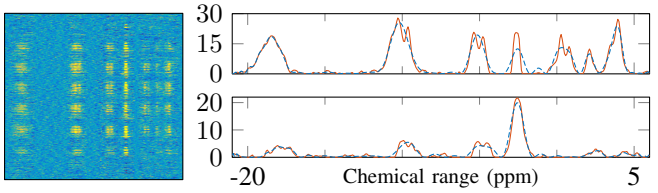


Fig. 4: (Left) Visualization of $\mathbf{X}^{\mathrm{T}}$ from the third experiment. (Right) Aligned estimated components using NP-NMF (------) and aHALS (——).

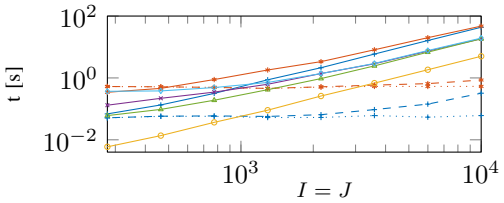


Fig. 5: Median timing results across 10 simulations for the fourth experiment, for aHALS (——), PG [18] (——), CG NP-NMF (——) and NLS NP-NMF (——). The solid and dashed lines for NP-NMF show the uncompressed and compressed versions, respectively. The dotted lines correspond to only the optimization part of the compressed versions, without the compression itself. The element-wise parametrization $\mathbf{W} = \mathbf{E} * \mathbf{E}$ and $\mathbf{H} = \mathbf{F} * \mathbf{F}$ is also considered [19], with CG (——) and NLS (——) versions implemented using the structured data fusion framework from Tensorlab [46], [47].

each containing $I = 369$ resonance bands between approx. -20 and 5 ppm. Both the single-sided NP-NMF ($D^{(w)} = 34$) and general NMF are applied with $R = 2$ components (brain and muscle). Fig. 4 shows similar estimated components, with NP-NMF yielding smooth approximations of the general NMF components. As discussed in the literature, additional constraints (e.g., local sparsity) are needed to incorporate prior knowledge and separate the brain and muscle components.

A fourth experiment compares the timing characteristics of the different methods in the double-sided setting, each with 20 iterations. The columns of $\mathbf{W}$ and $\mathbf{H}$ contain degree-20 nonnegative polynomials. The SNR is 10 dB, $R = 2$ and $D^{(w)}, D^{(h)} = 20$. Fig. 5 shows timing results for varying dimensions of $\mathbf{X}$. It can be seen that the compression-based QN and NLS NP-NMF methods are faster than aHALS and PG [18] for large dimensions, with the cost of the optimization steps independent of I, J. We note that, in general, NLS needs fewer iterations than CG or ALS to obtain the same accuracy; Fig. 5 essentially compares the cost per iteration.

V. Conclusion and Discussion

This paper proposes novel NMF methods by modeling the components using finite-interval nonnegative polynomials. The parametric representation avoids explicit non-negativity constraints and guarantees smooth solutions. Single- and double-sided variants have been discussed, and quasi-Newton and nonlinear least squares algorithms have been derived. We have indicated that nonnegative polynomials can model a variety of shapes, and the models have successfully been applied on two real-life datasets. A particular orthogonal compression has been proposed that can significantly speed up the NP-NMF methods and can outperform state-of-the-art methods. As future work, the algorithms can be adapted in a spline-like manner to obtain approximating polynomials of lower degrees in different regions of the components.

REFERENCES

[1] Y. Koren, R. Bell, C. Volinsky *et al.*, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[2] P. Comon and C. Jutten, *Handbook of blind source separation: Independent component analysis and applications*. Academic Press, 2010.

[3] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[4] N. Gillis, "The why and how of nonnegative matrix factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines*, ser. Machine Learning and Pattern Recognition. Chapman & Hall / CRC, 2014, ch. 12, pp. 257–291.

[5] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems*. MIT Press, 2004, pp. 1141–1148.

[6] N. Gillis, "Sparse and unique nonnegative matrix factorization through data preprocessing," *Journal of Machine Learning Research*, vol. 13, pp. 3349–3386, 2012.

[7] J. Eggert and E. Korner, "Sparse coding and NMF," *IEEE Proceedings of International Joint Conference on Neural Networks*, vol. 4, pp. 2529–2533, 2004.

[8] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[9] T. Yokota, R. Zdunek, A. Cichocki, and Y. Yamashita, "Smooth nonnegative matrix and tensor factorizations for robust multi-way data analysis," *Signal Processing*, vol. 113, pp. 234–249, 2015.

[10] Z. Chen, A. Cichocki, and T. M. Rutkowski, "Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer disease," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2006, pp. 896–896.

[11] Q. Dong and L. Li, "Smooth incomplete matrix factorization and its applications in image/video denoising," *Neurocomputing*, vol. 122, pp. 458–469, 2013.

[12] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *IEEE International Joint Conference on Neural Networks*, 2008, pp. 1828–1832.

[13] D. Lee, H. Seung *et al.*, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[14] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[15] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1995, vol. 15.

[16] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 713–730, 2008.

[17] M. H. Van Benthem and M. R. Keenan, "Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems," *Journal of chemometrics*, vol. 18, no. 10, pp. 441–450, 2004.

[18] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[19] M. Chu, F. Diele, R. Plemmons, and S. Ragni, "Optimality, computation and interpretation of nonnegative matrix factorizations," *SIAM Journal on Matrix Analysis*, pp. 4–8030, 2004.

[20] S. Wild, J. Curry, and A. Dougherty, "Improving non-negative matrix factorizations through structured initialization," *Pattern Recognition*, vol. 37, no. 11, pp. 2217–2232, 2004.

[21] C. Boutsidis and E. Gallopoulos, "SVD-based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[22] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, "Initializations for the nonnegative matrix factorization," in *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*. Citeseer, 2006, pp. 23–26.

[23] L. De Lathauwer, "Blind separation of exponential polynomials and the decomposition of a tensor in rank-$(L_r, L_r, 1)$ terms," *SIAM Journal on Matrix Analysis and Applications*, vol. 32, no. 4, pp. 1451–1474, 2011.

[24] O. Debals, M. Van Barel, and L. De Lathauwer, "Löwner-based blind signal separation of rational functions with applications," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1909–1918, April 2016.

[25] M. Boussé, O. Debals, and L. De Lathauwer, "A tensor-based method for large-scale blind source separation using segmentation," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 346–358, Jan 2017.

[26] O. Debals and L. De Lathauwer, "Stochastic and deterministic tensorization for blind signal separation," in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, vol. 9237. Springer Berlin / Heidelberg, 2015, pp. 3–13.

[27] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Transactions on Neural Networks*, vol. 19, no. 6, pp. 1090–1100, June 2008.

[28] J.-B. Lasserre, *Moments, positive polynomials and their applications*. World Scientific, 2009, vol. 1.

[29] M. Fekete, "Proof of three propositions of Paley," *Bulletin of the American Mathematical Society*, vol. 41, no. 2, pp. 138–144, 1935.

[30] G. Pólya and G. Szegö, *Problems and Theorems in Analysis II: Theory of Functions*. Springer Science & Business Media, 1997.

[31] V. Powers and B. Reznick, "Polynomials that are positive on an interval," *Transactions of the American Mathematical Society*, vol. 352, no. 10, pp. 4677–4692, 2000.

[32] E. Acar, D. Dunlavy, T. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41 – 56, 2011.

[33] L. Sorber, M. Van Barel, and L. De Lathauwer, "Unconstrained optimization of real functions in complex variables," *SIAM Journal on Optimization*, vol. 22, no. 3, pp. 879–898, 2012.

[34] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed. Springer, 2006.

[35] N. Gillis, "Nonnegative matrix factorization: Complexity, algorithms and applications," Ph.D. dissertation, UCL, 2011.

[36] N. Gillis and F. Glineur, "Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization," *Neural Computation*, vol. 24, no. 4, pp. 1085–1105, 2012.

[37] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2009.

[38] L. Sorber, M. Van Barel, and L. De Lathauwer. "Complex optimization toolbox v1.0." Available online. http://esat.kuleuven.be/stadius/cot/.

[39] A. Cichocki, R. Zdunek, and S.-I. Amari, "Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 169–176.

[40] C.-I. Chang, "An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1927–1932, Aug 2000.

[41] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 650–663, March 2004.

[42] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 155 – 173, 2007.

[43] M. N. Schmidt and H. Laurberg, "Nonnegative matrix factorization with Gaussian process priors," *Computational intelligence and neuroscience*, vol. 2008, p. 3, 2008.

[44] M. F. Ochs, R. Stoyanova, F. Arias-Mendoza, and T. Brown, "A new method for spectral decomposition using a bilinear Bayesian approach," *Journal of Magnetic Resonance*, vol. 137, no. 1, pp. 161–176, 1999.

[45] P. Sajda, S. Du, T. R. Brown, R. Stoyanova, D. C. Shungu, X. Mao, and L. C. Parra, "Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain," *IEEE transactions on medical imaging*, vol. 23, no. 12, pp. 1453–1465, 2004.

[46] L. Sorber, M. Van Barel, and L. De Lathauwer, "Structured data fusion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 586–600, June 2015.

[47] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer. "Tensorlab 3.0." [Online]. Available: http://www.tensorlab.net