Non-negative Matrix factorization

Theory and Methods

Laurberg, Hans

Non-negative Matrix Factorization: Theory and Methods

Ph.D. Thesis

HANS LAURBERG

Multimedia Information and Signal Processing
Aalborg University
9220 Aalborg Ø, Denmark

Laurberg, Hans.
Non-negative Matrix Factorization: Theory and Methods
ISBN XXX
ISSN XXX
First Edition, First Print

Multimedia Information and Signal Processing
Aalborg University
9220 Aalborg Ø
Denmark

Supervisor:
Professor, Ph.D Søren Holdt Jensen, Aalborg University
Assis. Prof., Ph.D Mads G. Christensen, Aalborg University

It sometimes happens that a father has an ugly son with no redeeming grace whatever, yet love will draw a veil over the parental eyes which then behold only cleverness and beauty in place of defects, and in speaking to his friends he will make those defects out to be the signs of comeliness and intellect. I, however, who am but Don Quixote's stepfather, have no desire to go with the current of custom, nor would I, dearest reader, beseech you with tears in my eyes as others do to pardon or overlook the faults you discover in this book; you are neither relative nor friend but may call your soul your own and exercise your free judgment. You are in your own house where you are master as the king is of his taxes, for you are familiar with the saying, "Under my cloak I kill the king". All of which exempts and frees you from any kind of respect or obligation; you may say of this story whatever you choose without fear of being slandered for an ill opinion any more than you will be rewarded for a good one.

— Miguel de Cervantes Saavedra
Don Quixote.

Abstract

The topic for this Thesis is a data analysis method called Non-negative Matrix Factorization (NMF). This method can analyze data with no negative elements e.g. images, spectra and probabilities. The introduction contains a profound review of the NMF literature focusing on the characteristics of the solutions and the underlying cost-functions to minimize for different NMF variations. One often-used method for comparing NMF is Principal Components Analysis (PCA) that is an analysis method for data containing both positive and negative elements. The introduction contains a thorough analysis that explains why PCA rarely finds the wanted solution of non-negative data.

An analysis gives a characterization of data that produces unique NMF i.e. data where NMF gives predictable results. Moreover, we prove that a corruption of data will result in a minor estimation error when the corruption is minor.

There is a description of a novel NMF method that uses Gaussian process priors offers the possibility to specify prior knowledge of the result. It is well known that data with positive offset makes the NMF non-unique. We present an affine NMF method that that jointly finds the offset and makes NMF. When NMF are used for source separation the requirement of single source training data are often assumed essential. We prove that it in many cases it is sufficient to have knowledge about when the sources are inactive.

Resumé

Emnet for denne afhandling er en dataanalysemetode kaldet Non-negative Matrix Factorization (NMF). Analysemetoden kan bruges på data hvor der ikke forekommer negative elementer som f.eks. billeder, spectra og hyppigheder. I introduktionen er der en grundig gennemgang af NMF litteraturen med fokus på det kendskab der er til problemets løsninger samt de kost-funktioner der ønskes minimeret i de forskellige NMF varianter. Den oftest benyttede metode til sammenligning af NMF er Principal Components Analysis (PCA) som er en metode for dataanalyse der både kan arbejde med positive og negative dataelementer. I introduktionen er der en ny grundig analyse af hvorfor PCA generelt ikke giver det ønskede resultat.

I afhandlingen er der en grundig analyse af hvad der skal karakterisere data for at man kan regne med at NMF kan finde de komponenter der blev anvendt da data blev genereret. Ydermere viser analysen at når NMF bruges på data hvor der er additiv støj, så resulterer det i at der bliver lagt støj på de komponenter der findes.

Der er præsenteret en NMF metode hvor det er muligt i en Bayesiansk ramme at specificere den apriori (forhånds) viden man har om NMF analysen. En af de kendte faktorer som kan få NMF til at give uforudsigelige resultater er hvis der er et offset i data. I afhandlingen er der præsenteret en NMF metode der er i stand til at detektere offsettet og dermed gøre det muligt at anvende NMF på data med et offset. Når NMF skal anvendes til separation er det normaltvist antaget at det er nødvendigt at der er optagelser med kilderne enkeltvist. I afhandlingen er der en analyse som dokumenterer at det ikke er nødvendigt at have kilderne enkeltvist da det er tilstrækkeligt at der kendes til tidspunkter hvor kilderne ikke er aktive.

List of Publications

The main body of this thesis consists of the following publications:

[A] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: On the uniqueness of NMF", in *Computational Intelligence and Neuroscience*, 2008

[B] M. N. Schmidt and H. Laurberg, "Non-negative matrix factorization with Gaussian process priors", in *Computational Intelligence and Neuroscience*, 2008.

[C] H. Laurberg, M. N. Schmidt, M. G. Christensen and S. H. Jensen, "Structured Non-negative Matrix Factorization with Sparsity Patterns", Submitted to Proc. *Asilomar*, 2008.

[D] H. Laurberg and L.K. Hansen, "On affine non-negative matrix factorization", in Proc. IEEE *Int. Conf. Acoustics, Speech and Signal Processing*, 2007.

[E] H. Laurberg, "Uniqueness of non-negative matrix factorization", in Proc. IEEE *Statistical Signal Processing Workshop*, 2007.

Publication [A, B, D, E] have been published in peer-reviewed proceedings or journals. Publication [C] has been submitted to a peer-reviewed conference.

Preface

This thesis is submitted to the Faculty of Engineering, Science and Medicine at Aalborg University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The work was carried out during the period August 2005–August 2008 at the Department of Electronic Systems at Aalborg University and it was founded by the Intelligent Sound project, Danish Technical Research Council grant no. 26–02–0092.

I would like to take this opportunity to thank some of the people who have helped and inspired me during the last three years. Firstly, I would like to express thanks to Lars Kai Hansen and the rest of the staff at IMM on DTU that I have been visiting for 6 months. Previous to my visit, I have shortly been investigating NMF using non-unique data that illustrated that NMF does not work (on non-unique data). If it were not for the interesting discussion with you, I would still believe that NMF were not worth investigating.

A special thanks goes to Mikkel N. Schmidt with whom I have written two papers and who has been very helpful with discussions and references throughout the 3 years. Another special appreciation goes to Jesper Højvang Jensen and Morten Mørup for the fruitful discussions of the material presented in the Thesis. Furthermore, I would like to thank Mark D. Plumbley for the discussions and the invitation to visit the Centre for Digital Music on Queen Mary University of London and Dan Ellis for letting me visit the LabROSA on Columbia University New York.

I will also use this opportunity to thank my roommate Jesper Højvang Jensen, Lasse Lohilahti Mølgaard and Morten Højfeldt Rasmussen for the good atmosphere and furthermore I would like to apologize for lowering your productivity. I would also like to recognize the MISP group and the members of the Intelligent Sound project for setting a nice frame of my work.

Finally, I would like to thank my two advisors Søren Holdt Jensen and Mads Græsbøll Christensen for supporting my ideas and forcing me to rethink the results until the format is understandable for other people. In addition, I am grateful for the environment that my advisors are responsible for, that has allowed me to focus on research.

Hans Laurberg
Aalborg, june 2008

Contents

1 Introduction

The amount of data available has increased dramatically over the last 50 years and is now a requirement for automatic unsupervised analysis and classification to use the overwhelming amount of data. There is a large group of data where all the data elements are non-negative, and in this Thesis, we will describe the work with an unsupervised method for non-negative data analysis called Non-negative Matrix Factorisation (NMF).

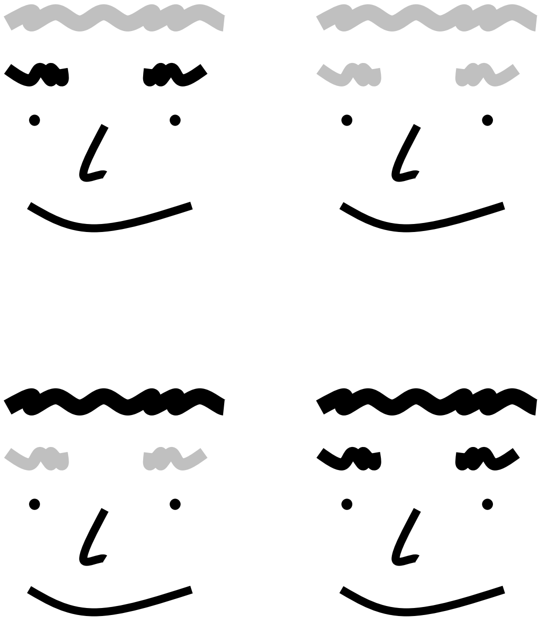


Figure 1: An easy to understand example of why special non-negative methods are needed for image analysis.

Before starting the more technical part of the Thesis, let us take a look at an example of some face drawings in Figure 1 with the intention of illustrating the concept. Let us assume the drawings are given to an unsupervised method for analysis of the problem. The figures are composed of three basis objects namely the hair, the eyebrows and the rest of the face. It can be seen that the hair and the eyebrows can be either colored black or gray. The task is to construct algorithms that are able to find those three basis objects and also are able to determine how intense these are in each drawing. The problem with the usual methods that can operate on both positive and negative data is that they will find solutions where one basis object is found to be both hair and eyebrows together and yet another basis objects that will be the positive hair and negative eyebrows together—which is both meaningless and useless in most applications. In this specific example, a non-negativity constraint of the objects will remove the possibility of making this kind of unwanted solutions and leaving the wanted solution as the only possible solution.

The Non-negative Matrix Factorization (NMF) analyzed in this Thesis can be mathematically described using matrices. The observations are gathered in a matrix $\mathbf{R}$ where each column $\mathbf{R}^i$ represents an observation of for example a picture, a spectrogram or probabilities. The task for the NMF is to find a set of basis objects $\mathbf{w}^i$ (basis picture, basis spectrogram, basis probabilities etc.) such that each observation can seen as a weighted sum of the basis vectors

$$\mathbf{R}^i = \sum_j \mathbf{W}^j \mathbf{H}^i_j. \tag{1}$$

By gathering the basis vectors in a matrix $\mathbf{W}$ and the weights in a matrix $\mathbf{H}$ it is possible to write the problem as

$$\mathbf{R} = \mathbf{W}\mathbf{H}. \tag{2}$$

At this point, the NMF may look like the LU, the QR or any other matrix factorization that are described in all undergrad linear algebra text books, e.g. [163], but this is certainly not the case. Firstly, NMF is despite the name not a factorization[1]. In all practical cases the observations $\mathbf{V} \in \mathbb{R}_+^{n\times m}$ will only be approximated by $\mathbf{W} \in \mathbb{R}_+^{n\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times m}$,

$$\mathbf{V} \approx \mathbf{R} = \mathbf{W}\mathbf{H}, \tag{3}$$

because $r \ll \min(n, m)$ and $r \ll \operatorname{rank}(\mathbf{V})$. Secondly, the NMF differs from the traditional factorizations as there often are additional constraint on $\mathbf{W}$ and $\mathbf{H}$ which makes the model more data specific.

The rest of the introduction to the Thesis is structured as follows. The mathematical fundaments are described in Section 2 and the aspects of NMF cost functions and NMF theory are described in details in Section 3. Section 4 analyses how a method for data that is both positive and negative will perform on non-negative data, and three concrete applications of NMF are described in Section 5. Finally, the Thesis introduction is concluded in Section 6 where the contributions of this PhD Thesis are listed.

2 Fundamentals

Several of the mathematical fundamentals used in this Thesis will be described in this section. The notation used in the Thesis are as follows.

[1] Because NMF is not a factorization some authors has argued for the name non-negative matrix approximation.

Font	Variable Type
x	Scalar
$\mathbf{x}$	Column vector
$\underline{\mathbf{x}}$	Row vector
$\mathbf{X}$	Matrix
$\mathbf{X}_i^j$	i'th colomn and j'th row
$\mathbf{1}$	Vector with 1 enlemnts

Operator	Explanation
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|$	Euclidian norm
$\lvert\cdot\rvert$	Element-wise absolute
$<, \leq, >, \geq$	Element-wise less than etc.
$\mathbf{X}^T$	Transpose of $\mathbf{X}$
$\mathbb{R}_+$	Non-negative real

The rest of the section is structured as follows. Firstly, some well known eigenvalue decomposition and singularvalue decomposition results from linear algebra are described in Section 2.1, followed by a repetition of the Frobenius-Perron Theory in Section 2.2. Finally, a novel generalization of block diagonal matrices and a property of this are given in Section 2.3.

2.1 The Spectral Theorem and Singular Value Decomposition

The Spectral Theorem and the Singular Value Decomposition (SVD) are two of the highlights of linear algebra. In words, the real[2] spectral Theorem states that any symmetric matrix is diagonalizable when the correct orthonormal basis is used.

The proofs of the SVD and the Spectral Theorem take several pages and are therefore omitted in here. The proof is available in several linear algebra textbooks e.g. [13, Theorem 7.13 and 7.46]

Theorem 1 (The Real Spectral Theorem) *For a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ there exist an orthonormal matrix $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda n \end{bmatrix} \in \mathbb{R}^{n\times n}$ such that*

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \tag{4}$$

if and only if $\mathbf{A}$ is symmetric.

When the eigenvalues λ_i are real, they are always written in a decreasing order, $\lambda_1 \geq \cdots \geq \lambda_n$. The decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is referred to as the Eigen-Value Decomposition (EVD). Another useful decomposition is the SVD, which shows that all matrices can be diagonalized.

[2] There is also a spectral Theorem for complex vector spaces. In complex vector spaces, more matrices are diagonalizable with orthonormal basis.

Theorem 2 (Singular Value Decomposition) *To any matrix* $\mathbf{A} \in \mathbb{R}^{n \times m}$ *there exist matrix* $\mathbf{\Sigma} \in \mathbb{R}_{+}^{n \times m}$ *on the form:*

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_{\min(n,m)} \end{bmatrix}, \tag{5}$$

where $\sigma_1 \geq \cdots \geq \sigma_{\min(n,m)}$, *such that*

$$\mathbf{A} = \mathbf{U}^T \mathbf{\Sigma} \mathbf{V}, \tag{6}$$

where $\mathbf{U} \in \mathbb{R}^{n \times n}$ *and* $\mathbf{V} \in \mathbb{R}^{m \times m}$ *are orthonormal.*

The following Lemma shows that there are a strong connection between the two decompositions.

Lemma 3 *If* $\mathbf{U}^T \mathbf{\Sigma} \mathbf{V}$ *is the SVD of* $\mathbf{A} \in \mathbb{R}^{n \times m}$ *then* $\mathbf{V}$ *and* $\mathbf{U}$ *consist of the eigenvectors of* $\mathbf{A}^T \mathbf{A}$ *and* $\mathbf{A} \mathbf{A}^T$ *respectively. The non-zero eigenvalues of* $\mathbf{A}^T \mathbf{A}$ *and* $\mathbf{A} \mathbf{A}^T$ *will be the squared singular values of* $\mathbf{A}$.

Proof. The proof is carry out by inserting $\mathbf{A} = \mathbf{U}^T \mathbf{\Sigma} \mathbf{V}$ and verifying that both $\mathbf{\Sigma}^T \mathbf{\Sigma}$ and $\mathbf{\Sigma} \mathbf{\Sigma}^T$ are diagonal. ■

2.2 Frobenius-Perron Theory

The Frobenius-Perron theory deals with the eigenvectors and the eigenvalues of non-negative and positive matrices. The matrices in this section are symmetric to shorten the proofs even though the Theorems are valid without this assumption. For a detailed analysis of the Frobenius-Perron Theory we refer to [122]. The reason for bringing this analysis is two folded. Firstly, the Frobenius-Perron theory will later in this Thesis (Section 4.2) be the starting point for analyzing the results of Principal Components Analysis (PCA) when applied to used on non-negative data. Secondly, dose the theory seams to be unknown for most of the NMF community with [27] as one of the few exceptions.

The following Lemma will be used in the proof of the Frobenius-Perron Theorems.

Lemma 4 *Let* $\mathbf{v} \in \mathbb{R}^n$, $\max(\mathbf{v}) > \mathbf{0}$ *and* $\mathbf{x} \in \mathbb{R}_{+}^{n}$.

a. then $\left|\mathbf{x}^T \mathbf{v}\right| \leq \mathbf{x}^T \left|\mathbf{v}\right|$

b. if $\mathbf{x} > \mathbf{0}$ *then* $\mathbf{v} \geq \mathbf{0}$ *if and only if* $\left|\mathbf{x}^T \mathbf{v}\right| = \mathbf{x}^T \left|\mathbf{v}\right|$

Proof. The statements follows directly from the triangle inequality of the absolute value. ■

Theorem 5 *If* $\mathbf{A} \in \mathbb{R}_+^{n\times n}$ *and* $\mathbf{A}^T = \mathbf{A}$ *then* $\lambda_1 = \max_i |\lambda_i|$ *and* $\mathbf{u}_1 \in \mathbb{R}_+^n$.

Proof. Let $\mathbf{v}'$ be the eigenvector corresponding to the eigenvalue with largest absolute value λ'. Without loss of generality, let us assume that one element of $\mathbf{v}'$ is positive, the norm of $\mathbf{v}'$ is one and therefore

$$\|\mathbf{A}\mathbf{v}'\| = |\lambda'| = \max_i |\lambda_i| = \max_{\|\mathbf{b}\|=1} \|\mathbf{\Lambda}\mathbf{b}\| = \max_{\|\mathbf{b}\|=1} \|\mathbf{A}\mathbf{b}\| \geq \left\|\mathbf{A}\left|\mathbf{v}'\right|\right\|. \tag{7}$$

It is known from Lemma 4.a that $\mathbf{A}|\mathbf{v}'| \geq |\mathbf{A}\mathbf{v}'|$ and in combination with Equation 7 it can be concluded that $|\mathbf{A}\mathbf{v}'| = \mathbf{A}|\mathbf{v}'|$ and moreover

$$|\lambda'|\,|\mathbf{v}'| = |\lambda'\mathbf{v}'| = |\mathbf{A}\mathbf{v}'| = \left|\mathbf{A}\left|\mathbf{v}'\right|\right| = \mathbf{A}\left|\mathbf{v}'\right|. \tag{8}$$

The largest eigenvalue λ_1 therefore equals $|\lambda'|$ and has the nonnegative $|\mathbf{v}'|$ as the corresponding eigenvector. ■

The following Theorem shows that strictly positivity of a matrix is inherit to the first eigenvector and that the positivity also ensures that the first eigenvalue is strictly larger than the other eigenvalues.

Theorem 6 *If* $\mathbf{A} \in \mathbb{R}_+^{n\times n}$ *and* $\mathbf{A}^T = \mathbf{A} > \mathbf{0}$ *then* $\lambda_1 > \max_{i\neq 1} |\lambda_i|$ *and* $\mathbf{u}_1 > 0$.

Proof. From Theorem 5 it is known that $\lambda_1 \geq \max_{i\neq 1} |\lambda_i|$ and that $\mathbf{u}_1 \in \mathbb{R}_+^n$. Since $\lambda_1\mathbf{u}_1 = \mathbf{A}\mathbf{u}_1 > \mathbf{0}$ it can be concluded that $\mathbf{u}_1$ must be strictly positive. All the other eigenvectors are orthogonal to $\mathbf{u}_1$ and can therefore not be non-negative. Lemma 4.b can be used to conclude that for $i \neq 1$

$$\left|\mathbf{A}\left|\mathbf{u}_i\right|\right| > |\mathbf{A}\mathbf{u}_i| \tag{9}$$

$$\lambda_1 = \max_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\| \geq \left\|\mathbf{A}\left|\mathbf{u}_i\right|\right\| > \|\mathbf{A}\mathbf{u}_i\| = |\lambda_i|. \tag{10}$$

It follows that λ_1 must be larger than the absolute value of the other eigenvalues. ■

2.3 Separable Linear Problems

In this section a generalization of block diagonal matrices are analysed. For simplicity,the matrix dimensions are left out. It is implicitly assumed that the block dimensions match up such that $\left[\begin{smallmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{smallmatrix}\right]^T \left[\begin{smallmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{smallmatrix}\right] = \mathbf{v}_1^T\mathbf{b}_1 + \mathbf{v}_2^T\mathbf{b}_2$. When a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is block diagonal, it means that it can be written as

$$\mathbf{A} = \left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]. \tag{11}$$

Many matrix operations like inversion, eigenvalue decomposition and singular value decomposition can be performed block wise, which is useful to lower the computational complexity. A block diagonal matrix can be perceived as a concatenation of two independent matrix problems. Looking at block diagonal matrices this way leads to a generalization of matrices that are a concatenating of two isolated linear operators.

Definition 7 *A matrix $\mathbf{A}$ is called a* ***generalized block diagonal matrix*** *if there exist permutations $\mathbf{P}_1$ and $\mathbf{P}_2$ such that $\mathbf{P}_1\mathbf{A}\mathbf{P}_2$ is block diagonal.*

The following Lemma characterizes a non-negative generalized block diagonal matrices.

Lemma 8 *Let $\mathbf{A} \in \mathbb{R}_{+}^{n\times m}$ be a matrix where there are no zero columns and let $\mathbf{b}$ denote a vector which elements are either 1 or 0 and has at least one 0 and one 1 element. Moreover, let $\mathbf{b}^c$ be a vector whose elements are the complementary of $\mathbf{b}$ meaning that $\mathbf{b}^c = \mathbf{1} - \mathbf{b}$. Then $\mathbf{A}$ is generalized block diagonal if and only if there exist an $\mathbf{b}$ such that $\mathbf{A}\mathbf{b}$ and $\mathbf{A}\mathbf{b}^c$ are orthogonal.*

Proof.
$\Rightarrow$ If $\mathbf{A}$ is generalized block diagonal then $\mathbf{b} = \mathbf{P}_2^T \left[\begin{smallmatrix}\mathbf{1}\\ \mathbf{0}\end{smallmatrix}\right]$ will be a solution because

$$\begin{aligned}(\mathbf{A}\mathbf{b})^T\mathbf{A}\mathbf{b}^c &= \left(\mathbf{P}_1\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]\mathbf{P}_2\mathbf{P}_2^T\left[\begin{smallmatrix}\mathbf{1}\\ \mathbf{0}\end{smallmatrix}\right]\right)^T\mathbf{P}_1\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]\mathbf{P}_2\mathbf{P}_2^T\left[\begin{smallmatrix}\mathbf{0}\\ \mathbf{1}\end{smallmatrix}\right] & (12)\\ &= \left(\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]\left[\begin{smallmatrix}\mathbf{1}\\ \mathbf{0}\end{smallmatrix}\right]\right)^T\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]\left[\begin{smallmatrix}\mathbf{0}\\ \mathbf{1}\end{smallmatrix}\right] & (13)\\ &=\mathbf{1}^T\left[\begin{smallmatrix}\mathbf{A}_1\\ \mathbf{0}\end{smallmatrix}\right]^T\left[\begin{smallmatrix}\mathbf{0}\\ \mathbf{A}_2\end{smallmatrix}\right]\mathbf{1} & (14)\\ &=\mathbf{1}^T\mathbf{0}\mathbf{1} = 0. & (15)\end{aligned}$$

$\Leftarrow$ If there is an $\mathbf{b}$ such that $\mathbf{A}\mathbf{b}$ and $\mathbf{A}\mathbf{b}^c$ are orthogonal then chose $\mathbf{P}_1$ and $\mathbf{P}_2$ such that

$$\mathbf{P}_1\mathbf{A}\mathbf{b} = \left[\begin{smallmatrix}\boldsymbol{\tau}\\ \mathbf{0}\end{smallmatrix}\right] \quad , \quad \mathbf{P}_1\mathbf{A}\mathbf{b}^c = \left[\begin{smallmatrix}\mathbf{0}\\ \boldsymbol{\tau}^c\end{smallmatrix}\right], \mathbf{b} = \mathbf{P}_2\left[\begin{smallmatrix}\mathbf{1}\\ \mathbf{0}\end{smallmatrix}\right], \tag{16}$$

where $\boldsymbol{\tau} > \mathbf{0}$ and $\boldsymbol{\tau}^c > \mathbf{0}$. By inserting we get that $\mathbf{P}_1\mathbf{A}\mathbf{P}_2\left[\begin{smallmatrix}\mathbf{1}\\ \mathbf{0}\end{smallmatrix}\right] = \left[\begin{smallmatrix}\boldsymbol{\tau}\\ \mathbf{0}\end{smallmatrix}\right]$ and $\mathbf{P}_1\mathbf{A}\mathbf{P}_2\left[\begin{smallmatrix}\mathbf{0}\\ \mathbf{1}\end{smallmatrix}\right] = \left[\begin{smallmatrix}\mathbf{0}\\ \boldsymbol{\tau}^c\end{smallmatrix}\right]$ and therefore $\mathbf{P}_1\mathbf{A}\mathbf{P}_2$ must be on the form $\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]$. ■

In Section 4.1 it will be of interest when $\mathbf{A}^T\mathbf{A}$ are generalized block diagonal. It turns out that there is a simple characterization that determine when $\mathbf{A}^T\mathbf{A}$ are generalized block diagonal.

Lemma 9 *An non-negative matrix $\mathbf{A}$ with no zero columns is an generalized block diagonal if and only if $\mathbf{A}^T\mathbf{A}$ is an generalized block diagonal matrix.*

Proof. $\Rightarrow$ If $\mathbf{P}_1\mathbf{A}\mathbf{P}_2 = \left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]$ then

$$\mathbf{P}_2^T\mathbf{A}^T\mathbf{A}\mathbf{P}_2 = \mathbf{P}_2^T\mathbf{A}^T\mathbf{P}_1^T\mathbf{P}_1\mathbf{A}\mathbf{P}_2 = \left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right]^T\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{smallmatrix}\right] = \left[\begin{smallmatrix}\mathbf{A}_1^T\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2^T\mathbf{A}_2\end{smallmatrix}\right] \tag{17}$$

$\Leftarrow$ From Lemma 8 it is known that any vector $\mathbf{b}$ with 1 and 0 elements only, will make

$\mathbf{Ab}$ and $\mathbf{Ab}^c$ non-orthogonal. When $\mathbf{Ab}$ and $\mathbf{Ab}^c$ are non-orthogonal they must have a common positive element and this common positive element will be maintained when multiplying by $\mathbf{A}^T$. Therefore, $\mathbf{A}^T\mathbf{Ab}$ and $\mathbf{A}^T\mathbf{Ab}^c$ will not be orthogonal and by Lemma 8 it can be concluded that $\mathbf{A}^T\mathbf{A}$ is not a generalized block diagonal matrix. ■

3 NMF Fundamentals

For a century, Principal Component Analysis (PCA) [136] (see Section 4) has been used to make rank reduction of matrices. Less than 2 decades ago a suboptimal PCA-like algorithm was proposed in [130] where it was possible to inpose a non-negativity restriction on the components. This was done in [129] under the name Positive Matrix Factorization (PMF) within the area of environmental science. The PMF was applied within this application area, but was not used in other areas before the renaming of the method to Nonnegative Matrix Factorisation (NMF) in [104] that proposed a new "easy-to-understand" algorithm and applied it in two very different areas namely face images and semantic text. In this Section, the different variants of NMF are explained together with some general properties of NMF. For a shorter overview of NMF methods and algorithms we refer to [18, 133]. Examples of NMF application will be given in Section 5.

3.1 Traditional NMF

NMF algorithms factorize a non-negative matrix $\mathbf{V} \in \mathbb{R}_+^{n\times m}$ into two non-negative matrices $\mathbf{W} \in \mathbb{R}_+^{n\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times m}$. Often it is only possible to decompose a matrix $\mathbf{R}$ that is an approximation to $\mathbf{V}$

$$\mathbf{V} \approx \mathbf{R} = \mathbf{WH}. \tag{18}$$

Later in this Thesis there will be discussions about how close an estimated $\mathbf{W}'$ and $\mathbf{H}'$ pair is to the generating $\mathbf{W}$ and $\mathbf{H}$ matrices. In this discussion, it is fruitful to use the following viewpoint

$$\mathbf{R} = \mathbf{WH} = \begin{bmatrix} \mathbf{w}_1, \cdots, \mathbf{w}_r \end{bmatrix} \begin{bmatrix} \underline{\mathbf{h}}_1 \\ \vdots \\ \underline{\mathbf{h}}_r \end{bmatrix} = \sum_{d=1}^{r} \mathbf{w}_d \underline{\mathbf{h}}_d, \tag{19}$$

where $\mathbf{R}$ is seen as the sum of r outer products. The outer products will sometimes be referred to as the components. There has been derived numerous of NMF algorithms for different distance measures between $\mathbf{V}$ and $\mathbf{R}$. Most of these distance measures are element-wise measures i.e.

$$D(\mathbf{R}\|\mathbf{V}) = \sum_{n'=1}^{n} \sum_{m'=1}^{m} d(\mathbf{R}_{n'}^{m'} \| \mathbf{V}_{n'}^{m'}), \tag{20}$$

with [72] being one of the few exceptions. Most NMF algorithms uses the two error functions investigated in [105] (Euclidian and Kullback-Leibler). An overview of the papers that deals with these functions can be seen in the following table.

Name	$d(x\|y)$	Applied	Algorithms	Property
Euclidian	$(x-y)^2$	[8, 10–12, 19, 41, 57, 88, 91, 104, 109, 115, 124, 125, 131, 132, 142, 147, 158, 160, 172, 175, 176, 180, 183, 185, 186, 192]	[21, 26, 33–35, 37, 66, 90, 105, 113, 114, 129, 148, 177, 181]	[43, 47]
Kullback-Leibler	$x \log \frac{x}{y} - x + y$	[6, 9, 20, 22, 24, 31, 32, 59, 67–71, 77–80, 86, 132, 135, 144, 155, 158, 172, 187, 191]	[36, 37, 54, 55, 105, 108, 113, 127, 149, 181, 189]	[45, 46, 61, 147, 150]

The Euclidian distance minimization can be seen as a maximum likelihood estimator where the difference is due to additive Gaussian noise. The generalized version of Kullback-Leibler divergence[3] can be shown to be equivalent to the EM algorithm [45, 61] and maximum likelihood for Poisson processes [147].

Some papers deals with groups of divergence like Bergman divergence [42, 162] Csiszar's φ divergence [39, 162], alpha divergence [106, 189], Young's divergence [162], and the interpolating cost function [99].

3.2 Sparse NMF

From the very first NMF paper [129] the possibility of seeking special NMF solutions were mentioned. The most popular special type of solutions are the sparse NMF solutions i.e. NMF where one or both of $\mathbf{W}$ and $\mathbf{H}$ has many zero entries. One of the advantages of NMF that was reported in [104] is that NMF tends to make part based and sparse solutions. Before looking into cost functions for sparse NMF, some of the arguments of sparse models are given below:

Occam's razor Sparse models are in some ways more simple model, and it turns out that simple models often outperform the more complex models.

Prior knowledge In many applications it is prior knowledge that the underlying data is sparse and it is therefore natural to make NMF look for such models—see Section 5 for examples.

[3] Also called I-divergence

Data fitting Sparse NMF becomes more vector quantization like. If $\mathbf{H}$ is maximum sparse it only has one non-zero element in each column and such NMF is a vector quantifier [62]. By making NMF more like a vector quantizer the column vectors of $\mathbf{W}$ get closer to the data and therefore, hopefully, describes data better.

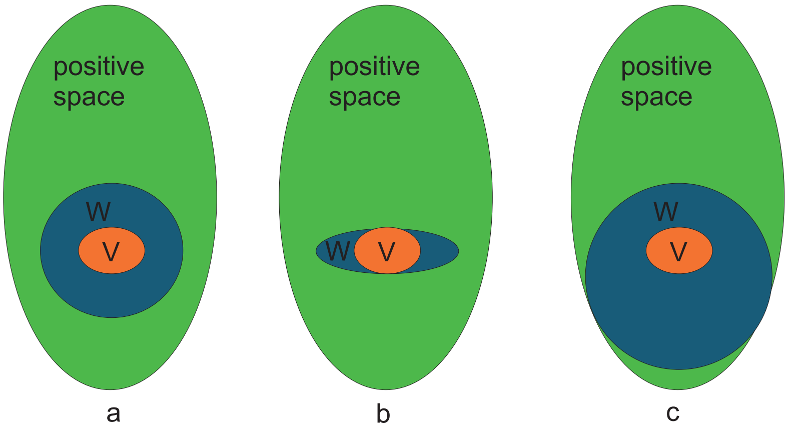


Figure 2: Illustration of why there is no sparse solution if the underlying NMF not sparse. The green area represent $\mathbb{R}^n_+$ with the outer border boundary being vectors with at least one zero element. The blue $\mathbf{W}$ area is all weighted averaged of the column vectors in $\mathbf{W}$ with the outer boundary being averaged where one of the weight is zero. The Orange area is the column vectors in $\mathbf{V}$. In **(a)** an non-sparse problem is shown. Because $\mathbf{W}$ are not sparse the W-space does not intersect with the boundary of the positive space, and because $\mathbf{H}$ is not sparse the W-space does not intersect with the V-space. In **(b)** $\mathbf{H}$ is sparse and the W-space does therefore intersects with the V-space, but at the same time the distance between the W-space and the boundary of the positive space increased. In **(c)** $\mathbf{W}$ is sparse and the W-space do therefore intersect with the boundary of the positive space, but at the same time the distance between the border of the W-space and the V-space increased.

Now after arguing for the sparse NMF it is worth mentioning that there also is one big counter argument to sparse NMF. If $\mathbf{R}$ is generated by non-sparse matrices $\mathbf{W} > 0$ and $\mathbf{H} > 0$ then there are no sparse $\mathbf{W}'$ and $\mathbf{H}'$ such that $\mathbf{R} = \mathbf{W}'\mathbf{H}'$. The argument for this can be seen in Figure 2. As explained in the figure caption it is either a matter of a sparse $\mathbf{W}$ (Figure 2.C) or a sparse $\mathbf{H}$ (Figure 2.B). In many applications it is not clear why the authors chose $\mathbf{W}$ to be sparse instead of $\mathbf{H}$ and vice versa. A counter argument to the analysis above is to look what happens when a not sparse NMF is used—with traditional NMF it is random which of the possible solution that is found, whereas the sparse NMF always gives the same output.

Most sparse NMF algorithms are formed by adding a penalty term to get an error function on the form

$$E(\mathbf{W}, \mathbf{H}) = D(\overline{\mathbf{W}}\mathbf{H} \| \mathbf{V}) + \lambda C(\mathbf{H}), \tag{21}$$

where $C(\cdot)$ is the penalty function and $\overline{\mathbf{W}}$ is a normalized version of $\mathbf{W}$. In some studies e.g. [83] $\mathbf{H}$ is both normalized and used in the penalty function. In sparse NMF most penalty functions are element wise,

$$C(\mathbf{H}) = \sum_{i,j} c(\mathbf{H}_i^j). \tag{22}$$

An overview of penalty functions are shown in the following table.

Name	$c(x)$	Reference
L_0 norm	$1(x \neq 0)$	[5, 19]
L_1 norm	$\|x\|$	[14, 16, 51, 72, 81, 82, 85, 92, 95, 112, 113, 117, 146, 151, 154]
L_2 norm	x^2	[7, 17, 60, 87, 90, 92, 134, 156, 158, 168, 190]
L_α norm	$\|x\|^\alpha$	[16, 156]

Another way of making sparse NMF is by a Lagrange multiplier approach where the level of sparsity fixed and the NMF is minimized with this level of sparsity [7, 72, 76, 83, 164]. At first it look, it seams very different to add a penalty term in the cost function and select the level of sparsity (the value of the penalty term), but it turns out that the solutions are the same. The argument is as follows. Let for a given λ

$$[\mathbf{W}', \mathbf{H}'] = \arg\min_{\mathbf{W},\mathbf{H}} \big(E(\mathbf{W}, \mathbf{H}) \big)$$

and thereby resulting in the rate $\gamma = C(\mathbf{H}')$. Now its easy to se that

$$[\mathbf{W}', \mathbf{H}'] = \arg\min_{\substack{\mathbf{W},\mathbf{H} \\ C(\mathbf{H})=\gamma}} \big(D(\overline{\mathbf{W}}\mathbf{H} \| \mathbf{V}) \big)$$

and therefore can the choice of λ be seen as a choice of rate γ.

A flavor of NMF referred to as Local Non-negative Matrix Factorization (LNMF) [53] has the penalty term

$$C(\mathbf{W}, \mathbf{H}) = \|\mathbf{W}\mathbf{1}\|^2 - \lambda' \|\mathbf{H}^T \mathbf{1}\|^2 \tag{23}$$

and is therefore an example of sparse NMF with not-elementwise penalty term [23, 49, 53, 110, 133, 178, 188, 188].

There is also a method for obtaining NMF with small estimation error and sparse $\mathbf{W}$ and $\mathbf{H}$ by adding a constant smoothing matrix $\mathbf{S}$. The smoothing matrix can thereby contain the non-sparseness of both $\mathbf{W}$ and $\mathbf{H}$:

$$\mathbf{V} \approx \mathbf{R} = \mathbf{W}\mathbf{S}\mathbf{H}. \tag{24}$$

This method presented in [133] is called NonSmooth Nonnegative Matrix Factorization (NSNMF) [92, 96, 97, 133]. An generalisation of NSNMF where there are different smoothing matrix for each component are described in [52] under the name Transformation-invariant NMF.

3.3 Structured NMF

In some applications, it is known that the solution has other characteristics or structure than general sparseness. Some of these applications and NMF algorithms are described in the following.

Affine NMF

One such application is the Swimmer Database (see Section 5.1) introduced in [48] where it was prove that traditional NMF cannot find the "correct" decomposition because all the column vectors in $\mathbf{V}$ has a constant part. To deal with NMF problems having an offset, an nmf algorithm called Affine NMF was introduced [101] where an extra term is added

$$\mathbf{V} \approx \mathbf{R} = \mathbf{W}\mathbf{H} + \mathbf{w}_0\mathbf{1}^T. \tag{25}$$

In Affine NMF $\mathbf{W}$ and $\mathbf{H}$ are updated using a sparse NMF method [51] and $\mathbf{w}_0$ is updated using a traditional NMF method [104]. Further details about Affine NMF can be found in paper D.

Smooth NMF

In some NMF problems it is known that the rows in $\mathbf{H}$ are smooth. An example of this is the spectrum of music instruments (see Section 5.3) where it is known that the spectra changes slowly over time [155]. In the very first NMF paper from 1994 [129] the possibility of making NMF algorithms that give smooth row vectors of $\mathbf{H}$ was mentioned. From that paper it has taken almost one decade before the first smooth NMF algorithm [170] was proposed in 2003. In this paper, a penaly term of the form

$$C(\mathbf{H}) = \sum_{i,j} \left|\mathbf{H}_j^i - \mathbf{H}_{j-1}^i\right|, \tag{26}$$

is used. Later the penalty function

$$C(\mathbf{H}) = \|(\mathbf{I} - \mathbf{T})\mathbf{H}\|_F^2, \tag{27}$$

where $\mathbf{T}$ is a matrix, that calculates a weighted averaged, was proposed to make smooth NMF [29]. Smooth NMF algorithms have superior performance to other NMF algorithms in several application [18, 29, 30, 152, 170, 172, 173]. When looking for smooth NMF algorithms it is worth noticing that in some articles, like for example [87, 134], the sparse NMF with L_2 regularization are called smooth.

Harmonic NMF
Another structured NMF algorithm is the Harmonic NMF [50, 143, 167] for note transcription. The Harmonic NMF uses the prior knowledge that tones have a known harmonic structure by forcing most elements in $\mathbf{W}$ to zero.

General structured NMF framework
In the last years, there has been a growing interest in a probabilistic interpretation of NMF [3, 45, 46, 54, 55, 61, 69, 111, 125, 131, 132, 147, 148, 174, 179, 182, 183]. A general framework for specifying the structure of a structured NMF is presented in [152] (found in Appendix B) by which it is possible to perform NMF with a chosen marginal distribution $p(\mathbf{H}_{i,j})$ and a chosen correlation between a re-parameterized version of the elements in $\mathbf{W}$ and $\mathbf{H}$.

3.4 NMF Extensions

Plenty of work has been done to generalize the NMF framework. Examples of this is the Non-negative Tensor Factorization, convolutive NMF and semi-NMF.

Non-negative Tensor Factorization
The tensor version of NMF called Non-negative Tensor Factorization (NTF) was started whilst NMF was still called PMF [128]. Later, there has been made a tensor product version of almost all NMF cost functions [38, 40, 56, 74, 75, 96–98, 107, 157, 184].

Convolutive NMF
Another generalization of NMF is the convolutive NMF where the column of $\mathbf{W}$ are exchanged with matrices. By doing this $\mathbf{W}$ consist of basis matrices instead of basis vectors [15, 93, 94, 153, 159, 161, 171].

Semi-NMF
Another way of changing the NMF is by discarding the non-negativity constraint. There are several NMF algorithms that work on negative $\mathbf{V}$ [147, 148, 152] and there are semi-NMF algorithms where the non-negativity constraint is only restricted to $\mathbf{H}$ [44, 111]. The semi-NMF also relates to non-negative PCA [126, 141] where $\mathbf{H}$ is non-negative and where $\mathbf{W}$ has orthonormal columns and also relates to the non-negative ICA [138–140] where $\mathbf{H}$ is non-negative and where $\mathbf{W}$ is orthonormal.

3.5 Is the NMF Problem Solvable?

The NMF theory is as old as the factorization it self. In 1972, a paper investigated when the LU-factorization of a non-negative matrix is also non-negative [119]. In 1974, it was investigated when an nmf exist with $r = \mathrm{rank}(V)$ [165] and later an analysis of the minimum possible r for which there exist an nmf was given in 1999 [166].

The first article to analyze the uniqueness of NMF was [48] in 2003 followed by [123] in 2005, [25] in 2007 and this author's papers [100, 102] found in paper A and paper E. The remaining of this section will contain a short overview of the uniqueness results of NMF.

When talking about a unique NMF it is assumed that there exist a ground truth $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{R} = \mathbf{WH}$ is the decomposition of interest. All other decompositions are denoted $\mathbf{R} = \tilde{\mathbf{W}}\tilde{\mathbf{H}}$. The only type of matrices that are non-negative and have a non-negative inverse are matrices on the form $\mathbf{PD}$ where $\mathbf{P}$ is a permutation and $\mathbf{D}$ is a diagonal matrix [122, Lemma 1.1]. This naturally leads to the definition of uniqueness that is directly or indirectly used in the NMF literature.

Definition 10 *A matrix $\mathbf{R} = \mathbf{WH}$ has a **unique NMF** if any solution $\mathbf{R} = \tilde{\mathbf{W}}\tilde{\mathbf{H}}$ fulfills that $\tilde{\mathbf{W}} = \mathbf{WPD}$ and $\tilde{\mathbf{H}} = \mathbf{D}^{-1}\mathbf{P}^{-1}\mathbf{H}$ where $\mathbf{P}$ is a permutation and $\mathbf{D}$ is a diagonal matrix.*

The permutation and scaling ambiguity with regard to the NMF uniqueness definition are also known from other areas such as Blind Source Separation (BSS). It is trivial to verify that unique NMF problems exists—an example is $\mathbf{I} = \mathbf{R} = \mathbf{WH} = \mathbf{II}$. When $\mathrm{rank}(\mathbf{R}) = r$ the non-uniqueness can be described by an invertible matrix $\mathbf{Q}$ where $\tilde{\mathbf{W}} = \mathbf{WQ}$ and $\tilde{\mathbf{H}} = \mathbf{Q}^{-1}\mathbf{H}$, but if $\mathrm{rank}(\mathbf{R}) \neq r$ this might is not be the case [25, 100, 165].

To verify that an nmf is unique it is necessary to jointly investigate the conditions on $\mathbf{W}$ and $\mathbf{H}$ [100], but there are a results of a condition fore the matrices individually as described in the following.

Definition 11 *A matrix $\mathbf{W}$ is called **boundary close** if for all $j \neq i$ there exist a k such that*

$$0 = \mathbf{W}_k^i \tag{28}$$

$$0 \neq \mathbf{W}_k^j \tag{29}$$

Theorem 12 *If the NMF of $\mathbf{R} = \mathbf{WH}$ is unique then both $\mathbf{W}$ and $\mathbf{H}^T$ are **boundary close**.*

The proof of Theorem 12 follows directly from the proof of Theorem A.10 on page A7. The Theorem is also presented in [123, Theorem 2] where one needs to be aware of a minor error in the formulation. The most loose condition which is sufficient for unique NMF is given by the following two definitions.

Definition 13 *A matrix $\mathbf{W}$ is called **sufficiently spread** if for all j there exist a k such that for all $i \neq j$*

$$0 = \mathbf{W}_k^i \tag{30}$$

$$0 \neq \mathbf{W}_k^j \tag{31}$$

Definition 14 *A matrix* $\mathbf{W}$ *is called* ***strongly boundary close*** *if* $\mathbf{W}$ *is* ***boundary close*** *and there exist a permutation* $\mathbf{P}$ *such that* $\widehat{\mathbf{W}} = \mathbf{WP}$*, for which all* $i < r$ *there exist a set* $\{k_1, \cdots, k_{r-i}\}$ *fulfilling*

1. $\widehat{\mathbf{W}}^i_{k_j} = 0$ *for all* $j \leq r - i$
2. *and the matrix* $\begin{bmatrix} \widehat{\mathbf{W}}^{i+1}_{k_1} & \cdots & \widehat{\mathbf{W}}^{r}_{k_1} \\ \vdots & \ddots & \vdots \\ \widehat{\mathbf{W}}^{i+1}_{k_{r-i}} & \cdots & \widehat{\mathbf{W}}^{r}_{k_{r-i}} \end{bmatrix}$ *is invertible.*

Theorem 15 *If* $\mathbf{W}$ *is* ***sufficiently spread*** *and* $\mathbf{H}^T$ *is* ***strongly boundary close*** *then the NMF of* $\mathbf{R} = \mathbf{WH}$ *is unique.*

Theorem 15 is the same as Theorem A.15 on page A11 where the proof can be found[4]. In many examples, the **strongly boundary close** condition will be a sufficient condition for both matrices, but in order to constructs such examples it is necessary to evaluate the "condition number" [65, p. 81] of the matrices in the item 2 of Definition 14. Example 3 on page A9 illustrates the connection between the condition number and uniqueness of **strongly boundary close** $\mathbf{W}$ and $\mathbf{H}$. The final theoretical comment in this section is that when $\mathbf{R}$ is unique then the non-uniqueness of $\mathbf{W}$ and $\mathbf{H}$ can be bounded by bounding the difference between $\mathbf{R}$ and $\mathbf{V}$ as given in Theorem A.16 on page A12 also in [100]. Using the wordings from the "Inverse Problems" area one can say that Theorem A.16 shows that the NMF problem is "well-posed" as defined in [73].

4 Principal Component Analysis and Non-negative Data

There has been published several papers where NMF outperforms PCA [9, 23, 67–72, 74, 85, 104, 115, 179]. In this section, Frobenius-Perron theory will be used to analyze the outcome of Principal Component Analysis (PCA) when the observation is non-negative. This analysis shows that PCA will output only one purely positive component and the remaining components will contain both positive and negative elements.

In PCA a set of vectors $\mathbf{v}_1, \cdots, \mathbf{v}_m \in \mathbb{R}^n$ is projected to a r-dimensional space such that most variance is obtained. In other words PCA finds a matrix $\mathbf{P}_{PCA} \in \mathbb{R}^{r\times n}$ with orthonormal row vectors that fulfils

$$\mathbf{P}_{PCA} = \underset{\substack{\mathbf{P}\in\mathbb{R}^{r\times n} \\ \mathbf{PP}^T=\mathbf{I}}}{\arg\max} \|\mathbf{PV}\|_F^2 \,. \tag{32}$$

[4] The definition in Appendix A looks different from the ones given in this section because those definition are given for sets and therefore are more general than the definitions here that are matrix specific.

Note, there are many solution to the maximization problem. Therefore, $\arg\max$ means that $\mathbf{P}_{PCA}$ is just one of the optimal matrices. The following Theorem shows that it is easy to find one $\mathbf{P}_{PCA}$ using the Singular Value Decomposition (SVD).

Theorem 16 *Let* $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$, $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_{\min(n,m)} \end{bmatrix} \in \mathbb{R}^{n\times m}$ *and* $\mathbf{K} = \begin{bmatrix} \underline{\mathbf{k}}_1 \\ \vdots \\ \underline{\mathbf{k}}_m \end{bmatrix} \in \mathbb{R}^{m\times m}$ *be the SVD of* $\mathbf{V}$*,*

$$\mathbf{V} = \mathbf{U\Sigma K}. \tag{33}$$

Then

$$\mathbf{P}_{PCA} = \mathbf{U}_r^T = [\mathbf{u}_1, \cdots, \mathbf{u}_r]^T = \underset{\substack{\mathbf{P}\in\mathbb{R}^{r\times m}\\ \mathbf{PP}^T=\mathbf{I}}}{\arg\max} \|\mathbf{PV}\|_F^2 . \tag{34}$$

Proof.

Since a rotation do not change the Frobenius norm (Equation 36) and $\mathbf{PU}$ also has orthonormal row vectors (Equation 37) we get

$$\max_{\substack{\mathbf{P}\in\mathbb{R}^{r\times m}\\ \mathbf{PP}^T=\mathbf{I}}} \|\mathbf{PV}\|_F^2 = \max_{\substack{\mathbf{P}\in\mathbb{R}^{r\times n}\\ \mathbf{PP}^T=\mathbf{I}}} \|\mathbf{PU\Sigma K}\|_F^2 \tag{35}$$

$$= \max_{\substack{\mathbf{P}\in\mathbb{R}^{r\times n}\\ \mathbf{PP}^T=\mathbf{I}}} \|\mathbf{PU\Sigma}\|_F^2 \tag{36}$$

$$= \max_{\substack{\mathbf{P}\in\mathbb{R}^{r\times n}\\ \mathbf{PP}^T=\mathbf{I}}} \|\mathbf{P\Sigma}\|_F^2 \tag{37}$$

$$= \sum_{i=1}^{r} \sigma_i^2 \tag{38}$$

The proof is concluded by testing $\mathbf{P}_{PCA}$

$$\|\mathbf{P}_{PCA}\mathbf{V}\|_F^2 = \left\|[\mathbf{u}_1, \cdots, \mathbf{u}_r]^T \mathbf{U\Sigma K}\right\|_F^2 \tag{39}$$

$$= \left\|[\mathbf{u}_1, \cdots, \mathbf{u}_r]^T \mathbf{U\Sigma}\right\|_F^2 \tag{40}$$

$$= \left\|\begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix} \mathbf{\Sigma}\right\|_F^2 \tag{41}$$

$$= \sum_{i=1}^{r} \sigma_i^2 \tag{42}$$

■

This closed form solution of PCA [5] makes it one of the most used algorithms for rank reduction. By using the SVD it is easy to calculate the result of the r-dimension representation, since

$$\begin{aligned}
\mathbf{P}_{PCA}\mathbf{V} &= \mathbf{U}_r^T\mathbf{U}\boldsymbol{\Sigma}\mathbf{K} && (43)\\
&= \begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}\boldsymbol{\Sigma}\mathbf{K} && (44)\\
&= \begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{bmatrix}\begin{bmatrix}\underline{\mathbf{k}}_1\\ \vdots \\ \underline{\mathbf{k}}_r\end{bmatrix} = \boldsymbol{\Sigma}_r\mathbf{K}_r, && (45)
\end{aligned}$$

the r-dimension representation consist of scaled versions of the singular vectors in $\mathbf{K}$. Another characterization of PCA is its ability to find the best rank r approximation for $\mathbf{V}$

$$\begin{aligned}
\underset{\text{rank}(\hat{\mathbf{V}})\le r}{\arg\max}\|\hat{\mathbf{V}}-\mathbf{V}\|_F &= \underset{\text{rank}(\hat{\mathbf{V}})\le r}{\arg\max}\|\hat{\mathbf{V}}-\mathbf{U}\boldsymbol{\Sigma}\mathbf{K}\|_F && (46)\\
&= \underset{\text{rank}(\hat{\mathbf{V}})\le r}{\arg\max}\|\mathbf{U}(\mathbf{U}^T\hat{\mathbf{V}}\mathbf{K}^T-\boldsymbol{\Sigma})\mathbf{K}\|_F && (47)\\
&= \underset{\text{rank}(\hat{\mathbf{V}})\le r}{\arg\max}\|\mathbf{U}^T\hat{\mathbf{V}}\mathbf{K}^T-\boldsymbol{\Sigma}\|_F && (48)\\
&= \mathbf{U}\left(\underset{\text{rank}(\hat{\mathbf{V}})\le r}{\arg\max}\left\|\hat{\mathbf{V}}-\boldsymbol{\Sigma}\right\|_F\right)\mathbf{K} && (49)\\
&= \mathbf{U}\begin{bmatrix}\boldsymbol{\Sigma}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{K} && (50)\\
&= \mathbf{U}_r\boldsymbol{\Sigma}_r\mathbf{K}_r. && (51)
\end{aligned}$$

This formulation of PCA has the same form as the NMF. Therefore, one could argue that PCA is an NMF without the non-negative constraint of $\mathbf{U}_r$ and $\boldsymbol{\Sigma}_r\mathbf{K}_r$. Before introducing more advanced algorithms to ensure the non-negativeness of $\mathbf{U}_r$ and $\boldsymbol{\Sigma}_r\mathbf{K}_r$ it is interesting to investigate if there are non-negative data matrices that naturally will lead to non-negative principal components. The analysis in the following section shows that this is not the case.

4.1 SVD of Non-negative Matrices

In this section, a Frobenius-Perron like analysis is applied to the SVD. From Lemma 3 it is known that the singular vectors of $\mathbf{A}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$. In the following, we will firstly investigate the Frobenius-Perron further for symmetric matrices and secondly apply the conclusions to the SVD.

[5] There is only a closed for solution of PCA from the programmers viewpoint. There is no closed form solution to find the SVD and therefore no closed form solution of PCA from an algorithmic point of view.

Theorem 17 *If $\mathbf{A} \in \mathbb{R}_+^{n \times n} \geq \mathbf{0}$, symmetric, and is not an generalized block diagonal, then $\lambda_1 > \max_{i \neq 1} |\lambda_i|$ and $\mathbf{u}_1$ is positive.*

Proof. Let $\mathbf{A}^T = \mathbf{A}$ not be a an generalized block diagonal matrix and let $\mathbf{u}$ be one of the vectors with the largest absolute eigenvalue. Because of the largest absolute eigenvalue we know that $|\mathbf{Au}| \geq \mathbf{A}\,|\mathbf{u}|$ and from Lemma 4.a we know that $|\mathbf{Au}| \leq \mathbf{A}\,|\mathbf{u}|$ and therefore $|\mathbf{Au}| = \mathbf{A}\,|\mathbf{u}|$. With no loss of generality it is assumed that the zero elements in $\mathbf{u}$ (if there are such) are grouped at the end, $\mathbf{u} = \left[\begin{smallmatrix}\hat{\mathbf{u}}\\ \mathbf{0}\end{smallmatrix}\right]$ where $\hat{\mathbf{u}}$ has only non-zero elements.

An analysis of the sub-blocks $\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{A}_2^T\\ \mathbf{A}_2 & \mathbf{A}_3\end{smallmatrix}\right]$ of $\mathbf{A}$ shows that $\mathbf{A}$ is block diagonal if there are zeros in $\mathbf{u}$

$$\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{A}_2^T\\ \mathbf{A}_2 & \mathbf{A}_3\end{smallmatrix}\right]\left[\begin{smallmatrix}\hat{\mathbf{u}}\\ \mathbf{0}\end{smallmatrix}\right] = \lambda \left[\begin{smallmatrix}\hat{\mathbf{u}}\\ \mathbf{0}\end{smallmatrix}\right] \tag{52}$$
$$\Downarrow$$
$$\mathbf{A}_2|\hat{\mathbf{u}}| = |\mathbf{A}_2\hat{\mathbf{u}}| = \mathbf{0} \tag{53}$$
$$\Downarrow$$
$$\mathbf{A}_2 = \mathbf{0}. \tag{54}$$

There cannot be any zero elements in $\mathbf{u}$ because a zero element in $\mathbf{u}$ will lead to a generalized block diagonal $\mathbf{A}$—which contradict with our starting assumption.

Now that it is known that $\mathbf{u}$ has no zero elements, lets split it in the positive elements and the negative elements, $\mathbf{u} = \left[\begin{smallmatrix}\mathbf{u}_p\\ -\mathbf{u}_n\end{smallmatrix}\right]$, where both $\mathbf{u}_p$ and $\mathbf{u}_n$ are strictly positive. An analysis of the sub blocks of $\mathbf{A}$ shows that

$$\left[\begin{smallmatrix}\mathbf{A}_1 & \mathbf{A}_2\\ \mathbf{A}_2^T & \mathbf{A}_3\end{smallmatrix}\right]\left[\begin{smallmatrix}\mathbf{u}_p\\ -\mathbf{u}_n\end{smallmatrix}\right] = \lambda \left[\begin{smallmatrix}\mathbf{u}_p\\ -\mathbf{u}_n\end{smallmatrix}\right] \tag{55}$$
$$\Downarrow$$
$$|\mathbf{A}_1\mathbf{u}_p - \mathbf{A}_2\mathbf{u}_n| = \mathbf{A}_1|\mathbf{u}_p| + \mathbf{A}_2|-\mathbf{u}_n| = \mathbf{A}_1\mathbf{u}_p + \mathbf{A}_2\mathbf{u}_n \tag{56}$$
$$\Downarrow$$
$$\mathbf{A}_1 = \mathbf{0} \text{ or } \mathbf{A}_2 = \mathbf{0}. \tag{57}$$

Repeating the steps above it can be shown that $\mathbf{A}_3 = \mathbf{0}$ when $\mathbf{A}_2 \neq \mathbf{0}$. When $\mathbf{u}$ has both negative and positive elements, $\mathbf{A}$ is therefore a generalized block diagonal matrix. ■

Theorem 18 *Let $\mathbf{V}$ be a non-negative matrix without zero rows or zero columns and not generalized block diagonal, then the singular vectors corresponding to the largest singular value will have positive elements.*

Proof. Lemma 3 states that the singular vectors of $\mathbf{V}$ are eigenvectors of $\mathbf{VV}^T$ and $\mathbf{V}^T\mathbf{V}$, and Lemma 9 states that when $\mathbf{V}$ is not generalized block diagonal then neither are $\mathbf{VV}^T$ nor $\mathbf{V}^T\mathbf{V}$. The use of Theorem 17 on those matrices concludes the proof. ■

Corollary 19 *Let $\mathbf{V}$ be a non-negative matrix without zero rows or zero columns and not generalized block diagonal then $\mathbf{U}$ and $\mathbf{K}$ from the SVD $\mathbf{V} = \mathbf{U}^T\mathbf{\Sigma}\mathbf{K}$ will both only have one non-negative column.*

Proof. Theorem 18 state that $\mathbf{U}$ and $\mathbf{K}$ have a strictly positive column vector and because the matrices are orthonormal, the other columns must have negative elements—and are therefore not non-negative. ■

4.2 PCA on Non-negative Matrices

Theorem 18 and Corollary 19 show directly that there is a problem with using PCA for non-negative matrices—namely, that only the first component is non-negative. If data is constructed by a sum of non-negative components as shown in Equation 18 then PCA will not find any components that are close to the generating components. The first element will be an averaged element and the other elements will be both positive and negative. The following example shows that PCA gathers all the energy in the first component when data is constructed using Equation 18.

Example 1 *Generate $\mathbf{W}$ and $\mathbf{H}$ from a uniform IID and an exponential IID and analyse how PCA performs on $\mathbf{WH}$.*

Different matrix sizes have been evaluated, but the result has been oserved to be consistent for all sizes. The first PCA component almost describes $\mathbf{R} = \mathbf{WH}$ by itself and the following components accounts for less than 2% of the total energy each. This is in contrast to the constructing components where the energy is even distributed over the component.

The average component energy over 1000 realizations with exponentially generated matrices where $n = m = 100$ and $r = 20$ are shown in Figure 3. The components are normalized so the total component energy sums to one. The blue dots are the energy of the normalized PCA component and the red squares are the energy in the sorted normalized constructing components. From Figure 3 it can be seen that the PCA components do not have the same energy distribution as the constructing components. This shows how different the two solutions are.

5 NMF Applications

5.1 Swimmer Database

The Swimmer Database was introduces in [48] as an example of a non-unique NMF problem. The database consist of 256 (32×32 pixel) black-and-white pictures of a ‘stick-man’ with 4 limbs that can be in one of 4 positions and a ‘torso’ as shown in Figure 4. Each of the 256 picture is a column vector in the data matrix such that $\mathbf{R} \in$

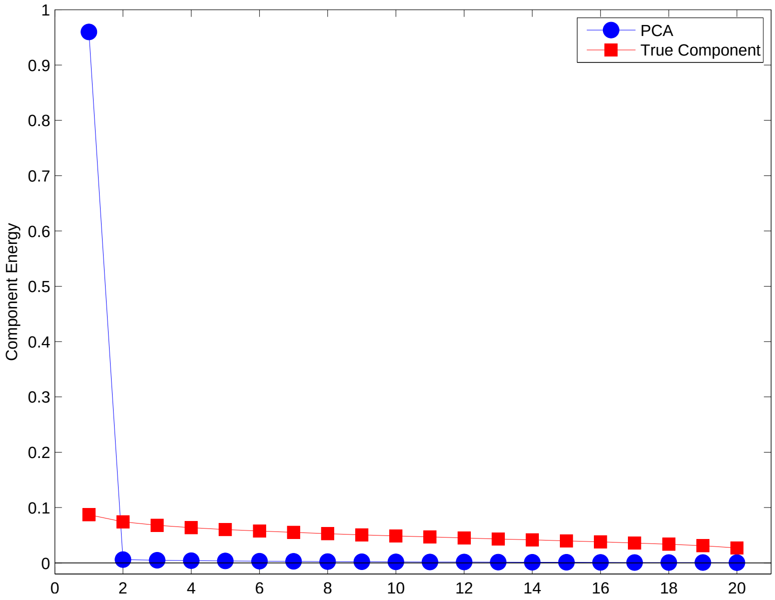


Figure 3: The average of normalized singular values of $\mathbf{WH}$ from 1000 realizations. Both $\mathbf{W} \in \mathbb{R}_+^{100\times 20}$ and $\mathbf{H} \in \mathbb{R}_+^{20\times 100}$ are generated with exponential IID.

$\mathbb{R}_+^{1024\times 256}$. In the paper that introduce the Swimmer database the model order r is set to 16 [48]. It is possible to decompose $\mathbf{R}$ using this model order by letting each limb in each positions be a basis vector and put on the torso to one of the limbs in all positions, but these basis will not be a good description of the pictures. In the later use of the Swimmer database is the model order set to 17 [33, 64, 74, 101, 133, 157, 193] in the hope that the algorithms can find the 16 limbs and the torso as basis vectors.

It has been shown that this NMF problem is very difficult to solve and many NMF algorithms fails to find the correct 17 basis vectors [33, 48, 74, 133, 157]. Two different strategies have been able to solve the problem. One strategy is to look for non-overlapping basis vectors and because the 17 basis vectors are non-overlapping, this gives the correct result as reported in [64, 193]. Another successful strategy is to use the Affine NMF [101] method that was described briefly in Section 3.3 and detailed in Appendix D.

5.2 Face Images

Faces images are one of the most used applications for NMF and was one of the examples that was given in the first paper named NMF [104]. The face images are in most cases passport like images that are cropped and scaled such that the eyes, mouth etc. are

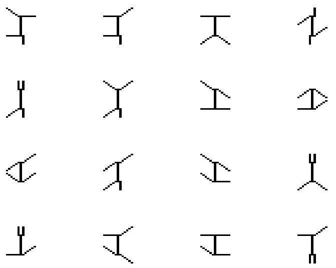


Figure 4: Samples of the Swimmer database

in the same position. Each image is a column vector in $\mathbf{V}$ and the columns in thereby becomes basis faces. There are several different application when using face images, e.g. face recognition, classification and illustrating purpose [37, 104, 190].

An overview of the databases and the use of NMF on the database can be seen in the following table.

Face database	Reference
AR [120]	[67, 68]
Cambridge ORL [2]	[53, 66, 84, 91, 95–97, 110, 113, 114, 178]
CBCL [1]	[21, 66, 74, 75, 84, 95, 99, 99, 113–117, 133, 157]
Cohn-Kanade [89]	[23]
FERET [137]	[178, 191]
JAFFE [118]	[23]
XM2VTS [121]	[145, 188]
Other and unknown	[4, 5, 9, 28, 37, 58, 76, 90, 104, 149, 190]

5.3 Music

In the last decade there has been a growing interest in music information retrieval and since 2000 there has been the ISMIR (The International Conferences on Music Information Retrieval and Related Activities) conference that has build a community for music research with a yearly competition MIREX (Music Information Retrieval Evaluation eXchange).

One of the tasks within the music area is to analyze frequency spectra of music signals and perform instrument separation or note transcription. Plenty of papers describe how to use NMF on music frequency spectra starting with [158]. When using NMF on music $\mathbf{V}_{ij}$ is the absolute value of the spectra at the i'th frequency bin at the j'th time index. The NMF of $\mathbf{V}$ results in a $\mathbf{W}$ where the columns are basis vectors for one instrument playing one note and a $\mathbf{H}$ that indicate when the notes are active. The music NMF papers can be split in to two major groups. The first group use NMF to make source separation of instruments or notes [9–12, 56, 63, 77, 93, 94, 103, 153, 168, 169, 171–173, 175, 176] by using Equation 19 and the second group does music transcription [3, 31, 32, 50, 63, 143, 158–160]. When performing transcription, the column vectors in $\mathbf{W}$ need to be classified as a note and the elements in $\mathbf{H}$ need to be classified as active or not active.

It is known that the additive NMF model is only an approximation for both amplitude and power spectra, but the more sparse the constructing components are the better will the approximation be. It is also known that independent power spectra in ensemble are additive, which is an argument for using this type of spectra. A counter argument is that it has also been reported that the amplitude spectra performs better than the power spectra [151]. The additive model error is the inspiration for [131, 132] that incorporate this error in the cost function.

6 Contributions

The NMF is a relative new factorization for analyzing large quantity of data. There has been published an overwhelming number of papers in the last decade but only a couple of the papers investigate when the factorization will produce a reliable result. Paper A and E investigate what data that gives a reliable NMF. Often when people use NMF, it is known that the solution has a certain structure. In Paper B, C and D NMF methods that looks for solutions with special structure are proposed. Next, the contribution of the individual papers are listed.

Paper A This paper investigates the uniqueness of NMF. Several new Theorems show that it is possible to characterize the conditions under which an nmf is unique. In most NMF applications either $\mathbf{R}$ or $(\mathbf{W}, \mathbf{H})$ is corrupted with noise and a Theorem shows that the estimation error is small when the noise is small. Finally this papers looks into stochastic properties of generating $\mathbf{W}$ and $\mathbf{H}$ that will make the NMF unique.

Paper B In this paper is a general method for making NMF that seeks $\mathbf{W}$ and $\mathbf{H}$ with special characteristic. The method makes it possible to determine both the marginal distribution of the source elements and the correlation between the elements. The method can be combined with any cost function with a probabilistic interpretation. The theoretical fundation of the method is Gaussian processes.

Paper C There are plenty of papers that use NMF to perform blind source separation under the assumption that it is possible to train the models of the sources individually. In this paper a method for training source models when several sources are presented. The proposed method only needs to information about when the sources are inactive which makes it possible to perform instrument separation in for example modern music. A Theorem shows that the method works under mild assumptions, and the theoretical findings is backed up by simulations on music data.

Paper D In this paper NMF is generalize to incorporate an explicit offset. The proposed method is able to find the correct decomposition on simulated data and when the method is used on face images where it constructs a more part-based decomposition than the reference NMF methods. The affine NMF method has the same computational complexity as the reference methods.

Paper E This paper presents some of the preliminary work that resulted in Paper A. The scope of the paper is to illustrate the novel definitions that are used in the NMF uniqueness Theorems. When the performance of NMF are simulated the elements in $\mathbf{W}$ and $\mathbf{H}$ are often generated as a stochastic process. The article argues for looking at real NMF data as be generated the same way and looks for the statistical properties of the stochastic process that makes NMF unique.

Thesis In the Thesis itself an overview of most of the NMF literature is given. Several papers have used PCA as a reference method when they present NMF. In the Thesis it is analysed how PCA perform on non-negative data. The reason for this is twofold. Firstly, it is possible to show why PCA rarely perform well, when it is used on non-negative data, and secondly we want to introduce the Frobenius-Perron Theory (the theory about eigenvalues and eigenvectors of non-negative matrices) for the NMF community where it is appears to be unknown. We strongly believe that the Frobenius-Perron Theory can increase the general knowledge of the NMF theory and methods.

References

[1] CBCL face database #1, MIT center for biological and computation learning. [Online]. Available: http://www.ai.mit.edu/projects/cbcl

[2] ORL face image database. [Online]. Available: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[3] S. A. Abdallah and M. D. Plumbley, “Polyphonic transcription by non-negative sparse coding of power spectra,” in *Music Information Retrieval, International Conference on (ISMIR)*, Oct 2004, pp. 318–325.

[4] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, November 2006.

[5] M. Aharon, M. Elad, and A. M. Bruckstein, “K-SVD and its non-negative variant for dictionary design,” in *Proc. of the SPIE conference wavelets*, vol. 5914, July 2005.

[6] J.-H. Ahn, S.-K. Kim, J.-H. Oh, and S. Choi, “Multiple nonnegative-matrix factorization of dynamic pet images,” in *Computer Vision, Asian Conference on*, Jan 2004, pp. 1009–1013.

[7] R. Albright, J. Cox, D. Duling, A. Langville, and C. Meyer, “Algorithms, initializations, and convergence for the nonnegative matrix factorization,” North Carolina State University, Tech. Rep. 81706, 2006.

[8] P. Anttila, P. Paatero, U. Tapper, and O. Järvinen, “Source identification of bulk wet deposition in finland by positive matrix factorization,” *Atmospheric Environment*, vol. 29, no. 14, pp. 1705–1718, 1995.

[9] H. Asari, “Non-negative matrix factorization: A possible way to learn sound dictionaries,” Tony Zador Lab, Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Tech. Rep., Aug 2005.

[10] ——, “Auditory system characterization,” Ph.D. dissertation, Watson School of Biological Sciences, Jul 2007.

[11] H. Asari, R. Olsson, B. Pearlmutter, and A. Zador, “Sparsification for monaural source separation,” in *Blind Speech Separation*, Springer, Ed. Springer, Sep 2007, ch. 14, pp. 387–410.

[12] H. Asari, B. A. Pearlmutter, and A. M. Zador, “Sparse representations for the cocktail party problem,” *The Journal of Neuroscience*, vol. 26(28), pp. 7477–7490, July 2006.

[13] S. J. Axler, *Linear algebra done right*, 2nd ed. New York, NY: Springer, 1997.

[14] L. Badea, “Clustering and metaclustering with nonnegative matrix decompositions,” in *Machine Learning, European Conference on (ECML), Lecture Notes in Computer Science (LNCS)*, vol. 3720. Springer, Nov 2005, pp. 10–22.

[15] S. Behnke, “Discovering hierarchical speech features using convolutional nonnegative matrix factorization,” in *Neural Networks (IJCNN), Proceeding of the International Joint Conference on*, vol. 4, Jul 2003, pp. 2758–2763.

[16] L. Benaroya, L. M. Donagh, F. Bimbot, and R. Gribonval, “Non negative sparse representation for wiener based source separation with a single sensor,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 6, Apr 2003, pp. 613–616.

[17] M. W. Berry and M. Brown, “Email surveillance using nonnegative matrix factorization,” *Computational and Mathematical Organization Theory*, vol. 11, pp. 249–264, Feb 2005.

[18] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, “Algorithms and applications for approximate nonnegative matrix factorization,” *Computational Statistics and Data Analysis*, 2006.

[19] N. Bertin, R. Badeau, and G. Richard, “Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark,” *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, pp. 65–68, 15-20 April 2007.

[20] A. Bertrand, K. Demuynck, V. Stouten, and H. Van hamme, “Unsupervised learning of auditory filter banks using non-negative matrix factorisation,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4713–4716, 2008.

[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] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 101, no. 12, pp. 4164–4169, Mar 2004.

[23] I. Buciu and I. Pitas, “Application of non-negative and local non negative matrix factorization to facial expression recognition,” in *Pattern Recognition, International Conference on (ICPR)*, vol. 1, Aug 2004, pp. 288–291.

[24] N. Cahill and R. Lawlor, “A novel approach to mixed phase room impulse response inversion for speech dereverberation,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4593–4596, 2008.

[25] B. Cao, D. Shen, J.-T. Sun, X. Wang, Q. Yang, and Z. Chen, “Detect and track latent factors with online nonnegative matrix factorization,” in *Artificial Intelligence, International Joint Conference on (IJCAI)*, 2007, pp. 2689–2694.

[26] M. Catral, L. Han, M. Neumann, and R. J. Plemmons, “On reduced rank nonnegative matrix factorization for symmetric nonnegative matrices,” *Linear Algebra and its Applications*, vol. 393, pp. 107–126, Dec 2004.

[27] D. Chen and R. J. Plemmons, “Nonnegativity constraints in numerical analysis,” in *The birth of numerical analysis, Symposium on*, 2007.

[28] X. Chen, L. Gu, S. Li, and H.-J. Zhang, “Learning representative local features for face detection,” *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. 1126–1131, 2001.

[29] Z. Chen and A. Cichocki, “Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints,” Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep., 2005.

[30] Z. Chen, A. Cichocki, and T. M. Rutkowski, “Constrained non-negative matrix factorization method for eeg analysis in early detection of alzheimer disease,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, May 2006.

[31] Y.-C. Cho, S. Choi, and S.-Y. Bang, “Non-negative component parts of sound for classification,” in *Signal Processing and Information Technology, IEEE International Symposium on (ISSPIT)*, Dec 2003, pp. 633–636.

[32] Y.-C. Cho and S. Choi, “Learning nonnegative features of spectro-temporal sounds for classification,” *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1327–1336, Jul 2005.

[33] M. T. Chu and M. M. Lin, “Low dimensional polytope approximation and its applications to nonnegative matrix factorization,” *SIAM Journal on Scientific Computing*, pp. 1131–1155, Mar 2008.

[34] M. Chu, F. Diele, R. J. Plemmons, and S. Ragni, “Optimality, computation, and interpretations of nonnegative matrix factorizations,” Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, Tech. Rep., 2004.

[35] M. Chu and R. Plemmons, “Nonnegative matrix factorization and applications,” *IMAGE, Bulletin of the International Linear Algebra Society*, vol. 34, pp. 2–7, Jul 2005.

[36] A. Cichocki and R. Zdunek, “Multilayer nonnegative matrix factorization,” *Electronic Letters*, vol. 42, no. 16, pp. 947–958, 2006.

[37] ——, “Multilayer nonnegative matrix factorization using projected gradient approaches,” in *Neural Information Processing, International Conference on (ICONIP)*, Oct 2006.

[38] ——, “Regularized alternating least squares algorithms for non-negative matrix/tensor factorization,” in *Neural Networks, Advances in (ISSN), Lecture Notes in Computer Science*, vol. 4493, 2007, pp. 793–802.

[39] A. Cichocki, R. Zdunek, and S. ichi Amari, “Csiszár’s divergences for non-negative matrix factorization: Family of new algorithms,” in *Lecture Notes in Computer Science (LNCS)*, vol. 3889. Springer, 2006, pp. 32–39.

[40] A. Cichocki, R. Zdunek, R. plemmons, and S. ichi Amari, “Non-negative tensor factorization using alpha and beta divergences,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, Apr 2007, pp. 1393–1396.

[41] M. Cooper and J. Foote, “Summarizing video using non-negative similarity matrix factorization,” in *Multimedia Signal Processing, IEEE Workshop on*, Dec 2002, pp. 22–28.

[42] I. S. Dhillon and S. Sra, “Generalized nonnegative matrix approximations with bregman divergences,” University of Texas at Austin, Department of Computer Sciences, Tech. Rep., 2005.

[43] C. Ding, X. He, and H. D. Simon, “On the equivalence of nonnegative matrix factorization and spectral clustering,” in *Data Mining, Proceedings of SIAM International Conference on*, 2005, pp. 606–610.

[44] C. Ding, T. Li, and M. I. Jordan, “Convex and semi-nonnegative matrix factorization,” Lawrence Berkeley National Laboratory, University of California, Berkeley, Tech. Rep. 60428, Nov 2006.

[45] C. Ding, T. Li, and W. Peng, “Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method,” in *Artificial Intelligence, AAAI National Conference on*, Jul 2006.

[46] ——, “On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing,” *Comput. Stat. Data Anal.*, vol. 52, no. 8, pp. 3913–3927, 2008.

[47] C. Ding, T. Li, W. Peng, and H. Park, “Orthogonal nonnegative matrix tri-factorizations for clustering,” in *Knowledge Discovery and Data Mining, International Conference on*, 2006, pp. 126–135.

[48] D. Donoho and V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts?” in *Neural Information Processing Systems, Advances in (NIPS)*, 2003.

[49] J. Duchateau, T. Leroy, K. Demuynck, and H. Van hamme, “Fast speaker adaptation using non-negative matrix factorization,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4269–4272, 2008.

[50] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 169–172, 2008.

[51] J. Eggert and E. Körner, "Sparse coding and NMF," in *Neural Networks, IEEE International Conference on*, vol. 4, 2004, pp. 2529–2533.

[52] J. Eggert, H. Wersing, and E. Körner, "Transformation-invariant representation and nmf," in *Neural Networks (IJCNN), Proceeding of the International Joint Conference on*, 2004, pp. 2535–2539.

[53] T. Feng, S. Z. Li, H.-Y. Shum, and H. Zhang, "Local non-negative matrix factorization as a visual representation," in *Development and Learning, International Conference on*, 2002, pp. 178–183.

[54] L. Finesso and P. Spreij, "Approximate nonnegative matrix factorization via alternating minimization," *ArXiv Mathematics e-prints*, 2004.

[55] ——, "Nonnegative matrix factorization and I-divergence alternating minimization," *Linear Algebra and Its Applications*, vol. 416, no. 2-3, pp. 270–287, 2006.

[56] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted 2d non-negative tensor factorization," in *Statistics in Signal Processing, IEEE Conference on*, Jul 2005.

[57] P. Fogel, S. S. Young, D. M. Hawkins, and N. Ledirac, "Inferential, robust non-negative matrix factorization analysis of microarray data," *Bioinformatics*, vol. 23, no. 1, pp. 44–49, 2007.

[58] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, February 2007.

[59] H.-T. Gao, T.-H. Li, K. Chen, W.-G. Li, and X. Bi, "Overlapping spectra resolution using non-negative matrix factorization," *Talanta*, vol. 66, no. 1, pp. 65–73, Mar 22005.

[60] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, pp. 3970–3975, 2005.

[61] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implication," in *Research and Development in Information Retrieval, Proceedings of the International SIGIR Conference on*, 2005, pp. 601–602.

[62] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[63] O. Gillet and G. Richard, “Transcription and separation of drum signals from polyphonic music,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 3, pp. 529–540, Mar 2008.

[64] N. Gillis and F. Glineur, “Nonnegative matrix factorization and underapproximation,” in *22nd national conference of the Belgian Operations Research Society*, Jan. 2008, p. 51.

[65] G. H. Golub and C. F. V. Loan, *Matrix computations*, 3rd ed. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[66] E. F. Gonzalez and Y. Zhang, “Accelerating the lee-seung algorithm for nonnegative matrix factorization,” Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, Tech. Rep. TR05-02, 2005.

[67] D. Guillamet and J. Vitrià, “Classifying faces with non-negative matrix factorization,” in *Artificial Intelligence, Catalonian Conference on (CCIA)*, 2002, pp. 24–31.

[68] ——, “Non-negative matrix factorization for face recognition,” in *Topics in Artificial Intelligence*, ser. Lecture Notes in Computer Science (LNCS). Springer, 2002, vol. 2504, pp. 336–344.

[69] D. Guillamet, J. Vitrià, and B. Schiele, “Introducing a weighted non-negative matrix factorization for image classification,” *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2447–2454, Oct 2003.

[70] D. Guillamet and J. Vitria, “Determining a suitable metric when using non-negative matrix factorization,” in *Pattern Recognition, International Conference on (ICPR)*, vol. 2, 2002, pp. 128–131.

[71] ——, “Evaluation of distance metrics for recognition based on non-negative matrix factorization,” *Pattern Recognition Letters*, vol. 24, no. 9, pp. 159–1605, Jun 2003.

[72] A. Hamza and D. Brady, “Reconstruction of reflectance spectra using robust non-negative matrix factorization,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 9, pp. 3637–3642, 2006.

[73] P. C. Hansen, *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. Philadelphia, USA: Society for Industrial and Applied Mathematics, 1998.

[74] T. Hazan, S. Polak, and A. Shashua, “Sparse image coding using a 3d non-negative tensor factorization,” in *Computer Vision, IEEE International Conference on*, vol. 1, Oct 2005, pp. 50–57.

[75] M. Heiler and C. Schnörr, “Controlling sparseness in nonnegative tensor factorization,” in *Computer Vision, European Conference on (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), vol. 3951. Springer, 2006, pp. 56–67.

[76] ——, “Learning sparse representations by non-negative matrix factorization and sequential cone programming,” *Journal of Machine Learning Research*, vol. 7, pp. 1385–1407, Jul 2006.

[77] M. Helén and T. Virtanen, “Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine,” in *European Signal Processing Conference, Proceedings of (EUSIPCO)*, Sep 2005.

[78] T. Hofmann, “Probabilistic latent semantic analysis,” in *Uncertainty in Artificial Intelligence (UAI)*, 1999, pp. 289–296.

[79] A. Holzapfel and Y. Stylianou, “A statistical approach to musical genre classification using non-negative matrix factorization,” *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2, pp. 693–696, 15-20 April 2007.

[80] ——, “Musical genre classification using nonnegative matrix factorization-based features,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 424–434, Feb. 2008.

[81] P. O. Hoyer, “Non-negative sparse coding,” in *IEEE Proc. Neural Networks for Signal Processing*, 2002, pp. 557–565.

[82] ——, “Modeling receptive fields with non-negative sparse coding,” *Neurocomputing*, vol. 52–54, pp. 547–552, Jun 2003.

[83] ——, “Non-negative matrix factorization with sparseness constraints,” *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, Nov 2004.

[84] ——, “Non-negative matrix factorization with sparseness constraints,” *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, Nov 2004.

[85] C. Hu, B. Zhang, S. Yan, Q. Yang, J. Yan, Z. Chen, and W.-Y. Ma, “Mining ratio rules via principal sparse non-negative matrix factorization,” in *Data Mining, IEEE/WIC/ACM International Conference on (ICDM)*, Nov 2004, pp. 407–410.

[86] S. Huang, K. Rahn, and R. Arimoto, “Testing and optimizing two factor-analysis techniques on aerosol at narragansett, rhode island,” *Atmospheric Environment*, vol. 33, pp. 2169–2185, June 1999.

[87] j. Piper, V. P. Pauca, R. Plemmons, and M. Giffin, “Object characterization from spectral data using nonnegative factorization and information theory,” in *Proceedings of Amos Technical Confefence*, Sep. 2004.

[88] M. Juvela, K. Lehtinen, and P. Paatero, “The use of positive matrix factorization in the analysis of molecular line spectra,” *Royal Astronomical Society, Monthly Notices of the*, vol. 280, no. 2, pp. 616–626, 1996.

[89] T. Kanade, J. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 46–53, 2000.

[90] D. Kim, S. Sra, and I. S. Dhillon, “Fast newton-type methods for the least squares nonnegative matrix approximation problem,” in *Data Mining, Proceedings of SIAM Conference on*, 2007.

[91] H. Kim and H. Park, “Discriminant analysis using nonnegative matrix factorization for nonparametric multiclass classification,” in *Granular Computer, IEEE International Conference on*, May 2006, pp. 182–187.

[92] ——, “Sparse non-negative matrix factorization via alternating non-negativity-constrained least squares for microarray data analysis,” *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[93] M. Kim and S. Choi, “On spectral basis selection for single channel polyphonic music separation,” in *Artificial Neural Networks, International Conference on (ICANN)*, ser. Lecture Notes in Computer Science (LNCS), vol. 3697. Springer, Sep 2005, pp. 157–162.

[94] ——, “Monaural music source separation: Nonnegativity, sparseness, and shift-invariance,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science (LNCS), vol. 3889. Springer, Apr 2006, pp. 617–624.

[95] Y.-D. Kim and S. Choi, “A method of initialization for nonnegative matrix factorization,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 2, Apr 2007, pp. 537–540.

[96] ——, “Nonnegative tucker decomposition,” in *Computer Vision and Pattern Recognition, IEEE Conference on (CVPR)*, Jun 2007, pp. 1–8.

[97] Y.-D. Kim, A. Cichocki, and S. Choi, “Nonnegative tucker decomposition with alpha-divergence,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, Mar 2008.

[98] ——, “Nonnegative tucker decomposition with alpha-divergence,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 1829–1832, 2008.

[99] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2007.

[100] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, “Theorems on positive data: On the uniqueness of nmf,” *Computational Intelligence and Neuroscience*, 2008.

[101] H. Laurberg and L. Hansen, “On affine non-negative matrix factorization,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. II, 2007, pp. 653–656.

[102] H. Laurberg, “Uniqueness of non-negative matrix factorization,” in *Statistical Signal Processing, IEEE Workshop on*, Aug 2007, pp. 44–48.

[103] H. Laurberg, M. N. Schmidt, M. G. Christensen, and S. H. Jensen, “Structured non-negative matrix factorization with sparsity patterns,” in *Signals, Systems, and Computers, Asilomar Conference on*, 2008, submitted to.

[104] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization.” *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[105] ——, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2000, pp. 556–562.

[106] H. Lee, A. Cichocki, and S. Choi, “Nonnegative matrix factorization for motor imagery eeg classification,” in *Artificial Neural Networks, International Conference on (ICANN)*, ser. Lecture Notes in Computer Science (LNCS), vol. 4132. Springer, Sep 2006, pp. 250–259.

[107] H. Lee, Y.-D. Kim, A. Cichocki, and S. Choi, “Nonnegative tensor factorization for continuous eeg classification,” *Neural Systems, International Journal of*, 2007.

[108] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee, “Application of non-negative matrix factorization to dynamic positron emission tomography,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, Dec 2001, pp. 629–632.

[109] J. S. Lee, D. Lee, S. Choi, K. S. Park, and D. S. Lee, “Non-negative matrix factorization of dynamic images in nuclear medicine,” in *Nuclear Science Symposium Conference Record*, vol. 4, 2001, pp. 2027–2030.

[110] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. Cheng, “Learning spatially localized, parts-based representation,” in *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, vol. 1, Dec 2001, pp. 207–212.

[111] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in *Data Mining, IEEE/WIC/ACM International Conference on (ICDM)*, Dec 2006, pp. 362–371.

[112] Y. Li and A. Cichocki, "Non-negative matrix factorization and its application in blind sparse source separation with less sensors than sources," in *Theoretical Electrical Engineering, International Symposium on (ISTET)*, 2003, pp. 285–288.

[113] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *Neural Networks, IEEE Transactions on*, vol. 18, pp. 1589–1596, 2007.

[114] ——, "Projected gradient methods for non-negative matrix factorization," *Neural Computation*, vol. 19, pp. 2756–2779, 2007.

[115] W. Liu and N. Zheng, "Non-negative matrix factorization based methods for object recognition," *Pattern Recognition Letters*, vol. 25, no. 8, pp. 893–897, Jun 2004.

[116] W. Liu, N. Zheng, and X. Li, "Relative gradient speeding up additive updates for nonnegative matrix factorization," in *New Aspects in Neurocomputing: European Symposium on Artificial Neural Networks*, ser. Neurocomputing, vol. 57. Elsevier, Mar 2004, pp. 493–499.

[117] W. Liu, N. Zheng, and X. Lu, "Non-negative matrix factorization for visual coding," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 3, Apr 2003, pp. 293–296.

[118] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 200–205, Apr 1998.

[119] T. L. Markham, "Factorizations of nonnegative matrices," *Proceedings of the American Mathematical Society*, vol. 32, no. 1, Mar. 1972.

[120] A. Martinez and R. Benavente, "The AR face database," CVC Tech. Report #24, 1998.

[121] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99)*, March 1999, pp. 72–77.

[122] H. Minc, *Nonnegative Matrices*, 1st ed. John Wiley and Sons, 1988.

[123] S. Moussaoui, D. Brie, and J. Idier, “Non-negative source separation: range of admissible solutions and conditions for the uniqueness of the solution,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 5, Mar 2005, pp. 289–292.

[124] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, “Separation of non-negative mixture of non-negative sources using a bayesian approach and MCMC sampling,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4133–4145, Nov 2006.

[125] S. Moussaoui, H. Hauksdottir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, and J. A. Benediktsson, “On the decomposition of mars hyperspectral data by ICA and bayesian positive source separation,” *Neurocomputing*, 2008.

[126] E. Oja and M. Plumbley, “Blind separation of positive sources using non-negative PCA,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, 2003, pp. 11–16.

[127] O. Okun and H. Priisalu, “Fast nonnegative matrix factorization and its application for protein fold recognition,” *EURASIP J. Appl. Signal Process.*, pp. 62–62, 2006.

[128] P. Paatero, “A weighted non-negative least squares algorithm for three-way parafac factor analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 223–242, 1997.

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

[130] ——, “Analysis of different modes of factor analysis as least squares fit problems,” *Chemometrics and Intelligent Laboratory Systems*, vol. 18, pp. 183–194, February 1993.

[131] R. M. Parry and I. Essa, “Incorporating phase information for source separation via spectrogram factorization,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 2, 2007, pp. 664–664.

[132] ——, “Phase-aware non-negative spectrogram factorization,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science (LNCS), vol. 4666. Springer, Sep 2007, pp. 536–543.

[133] A. Pascual-Montano, J. Carazo, K. Kochi, D. Lehmann, and R. Pascual-Marqui, “Nonsmooth nonnegative matrix factorization (nsNMF),” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 403–415, Mar 2006.

[134] V. P. Pauca, J. Pipery, and R. J. Plemmons, “Nonnegative matrix factorization for spectral data analysis,” *Linear Algebra and its Applications*, vol. 416, no. 1, pp. 29–47, Jul 2006.

[135] J. Paulus and T. Virtanen, “Drum transcription with non-negative spectrogram factorisation,” in *Proc. of the 13th European Signal Processing Conference*, Antalya, Turkey, Sep. 2005.

[136] K. Pearson, “On lines and planes of closest fit to systems of points in space,” *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[137] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, “The FERET database and evaluation procedure for face-recognition algorithms,” *Image and Vision Computing*, vol. 16, pp. 295–306, April 1998.

[138] M. Plumbley, “Conditions for nonnegative independent component analysis,” *Signal Processing Letters, IEEE*, vol. 9, no. 6, pp. 177–80, Jun 2002.

[139] M. D. Plumbley, “Conditions for nonnegative independent component analysis,” *Signal Processing Letters, IEEE*, vol. 9, no. 6, pp. 177–180, Jun 2002.

[140] ——, “Algorithms for nonnegative independent component analysis,” *Neural Networks, IEEE Transactions on*, vol. 14, no. 3, pp. 534–543, May 2003.

[141] M. D. Plumbley and E. Oja, “A “nonnegative PCA” algorithm for independent component analysis,” *Neural Networks, IEEE Transactions on*, vol. 15, no. 1, pp. 66–76, Jan 2004.

[142] A. V. POLISSAR, P. K. HOPKE, W. C. MALM, and S. J. F., “The ratio of aerosol optical absorption coefficients to sulfur concentrations, as an indicator of smoke from forest fires when sampling in polar regions,” *Atmospheric environment*, vol. 30, no. 7, pp. 1147–1157, 1996.

[143] S. A. Raczyński, N. Ono, and S. Sagayama, “Multipitch analysis with harmonic nonnegative matrix approximation,” in *Music Information Retrieval, International Conference on (ISMIR)*, Sep 2007.

[144] B. Raj, R. Singh, and P. Smaragdis, “Recognizing speech from simultaneous speakers,” in *Speech Communication and Technology, European Conference on (EUROSPEECH)*, 2005, pp. 3317–3320.

[145] M. Rajapakse and L. Wyse, “Nmf vs ica for face recognition,” *Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the 3rd International Symposium on*, vol. 2, pp. 605–610, Sept. 2003.

[146] S. Rebhan, J. Eggert, H.-M. Groß, and E. Körner, "Sparse and transformation-invariant hierarchical NMF," in *Artificial Neural Networks, International Conference on (ICANN)*, ser. Lecture Notes in Computer Science (LNCS), vol. 4668. Springer, Sep 2007, pp. 894–903.

[147] P. Sajda, S. Du, and L. Parra, "Recovery of constituent spectra using non-negative matrix factorization." in *Wavelets: Applications in Signal and Image Processing, Proceedings of SPIE*, vol. 5207, 2003, pp. 321–331.

[148] 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," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 12, pp. 1453–1465, Dec 2004.

[149] R. Salakhutdinov and S. Roweis, "Adaptive overrelaxed bound optimization methods," in *Proceedings of the International Conference on Machine Learning*, vol. 20, 2003, pp. 664–671.

[150] R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani, "On the convergence of bound optimization algorithms," in *Uncertainty in Artificial Intelligence (UAI)*, 2003, pp. 509–516.

[151] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Machine Learning for Signal Processing, IEEE International Workshop on*, Aug 2007, pp. 431–436.

[152] M. N. Schmidt and H. Laurberg, "Nonnegative matrix factorization with gaussian process priors," *Computational Intelligence and Neuroscience*, 2008.

[153] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-d deconvolution for blind single channel source separation," in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science (LNCS). Springer, Apr 2006, vol. 3889, pp. 700–707.

[154] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Spoken Language Proceeсing, ISCA International Conference on (INTERSPEECH)*, 2006.

[155] F. Sha and L. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Neural Information Processing Systems, Advances in (NIPS)*, 2005.

[156] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing and Management*, vol. 42, no. 2, pp. 373–386, 2006.

[157] A. Shashua and T. Hazan, “Non-negative tensor factorization with applications to statistics and computer vision,” in *Machine Learning, International Conference on (ICML)*, Aug 2005, pp. 793–800.

[158] P. Smaragdis and J. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*, Oct. 2003, pp. 177–180.

[159] P. Smaragdis, “Discovering auditory objects through non-negativity constraints,” in *Statistical and Perceptual Audio Processing, Workshop on (SAPA)*, Oct 2004.

[160] ——, “Non-negative matrix factor deconvolution; extraction of multiple sound sourses from monophonic inputs,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science (LNCS), vol. 3195. Springer, Sep 2004, pp. 494–499.

[161] ——, “Convolutive speech bases and their application to supervised speech separation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 1–12, Jan 2007.

[162] S. Sra and I. S. Dhillon, “Nonnegative matrix approximation: Algorithms and applications,” University of Texas at Austin, Tech. Rep., Jun 2006.

[163] G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 1993.

[164] F. Theis, K. Stadlthanner, and T. Tanaka, “First results on uniqueness of sparse non-negative matrix factorization,” in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2005.

[165] L. Thomas, “Solution to problem 73-14, rank factorizations of nonnegative matrices,” *SIAM*, vol. 16, pp. 393–Ű394, 1974.

[166] J. M. van den Hof and J. H. van Schuppen, “Positive matrix factorization via extremal polyhedral cones,” *Linear Algebra and its Applications*, vol. 293, pp. 171–186, 15 May 1999.

[167] E. Vincent, N. Bertin, and R. Badeau, “Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 109–112, 2008.

[168] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, “Model-based audio source separation,” Queen Mary University of London, Tech. Rep., 2006.

[169] E. Vincent and M. D. Plumbley, “Single-channel mixture decomposition using bayesian harmonic models,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science, vol. 3889, Feb 2006, pp. 722–730.

[170] T. Virtanen, “Sound source separation using sparse coding with temporal continuity objective,” in *International Computer Music Conference, ICMC*, 2003.

[171] ——, “Separation of sound sources by convolutive sparse coding,” in *Statistical and Perceptual Audio Processing, Workshop on (SAPA)*, Oct 2004.

[172] ——, “Sound source separation in monaural music signals,” Ph.D. dissertation, Tampere University of Technology, 2006.

[173] ——, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, Mar 2007.

[174] T. Virtanen, A. Taylan Cemgil, and S. Godsill, “Bayesian extensions to non-negative matrix factorisation for audio signal modelling,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 1825–1828, 2008.

[175] B. Wang and M. D. Plumbley, “Musical audio stream separation by non-negative matrix factorization,” in *DMRN Summer Conference, Glasgow, Proceedings of the*, Jul 2005.

[176] ——, “Investigating single-channel audio source separation methods based on non-negative matrix factorization,” in *ICA Research Network International Workshop, Proceedings of*, Sep 2006, pp. 17–20.

[177] G. Wang, A. V. Kossenkov, and M. F. Ochs, “LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates,” *BMC Bioinformatics*, vol. 7, no. 175, Mar 2006.

[178] Y. Wang, Y. Jia, C. Hu, and M. Turk, “Fisher non-negative matrix factorization for learning local features,” in *Computer Vision, Asian Conference on*, Jan 2004.

[179] L. Weixiang, Z. Nanning, and Y. Qubo, “Nonnegative matrix factorization and its applications in pattern recognition,” *Chinese Science Bulletin*, vol. 51, no. 1, pp. 7–18, Jan 2006.

[180] S. Wild, J. Curry, and A. Dougherty, “Motivating non-negative matrix factorizations,” in *Applied Linear Algebra, SIAM Conference on*, Jul 2003.

[181] ——, “Improving non-negative matrix factorizations through structured initialization,” *Pattern Recognition*, vol. 37, no. 11, pp. 2217–2232, Nov 2004.

[182] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4029–4032, 2008.

[183] O. Winther and K. B. Petersen, “Bayesian independent component analysis: Variational methods and non-negative decompositions,” *Digital Signal Processing*, vol. 17, no. 5, 2007.

[184] Y.-L. Xie, P. K. Hopke, and P. Paatero, “Positive matrix factorization applied to a curve resolution problem,” *Chemometrics, Journal of*, vol. 12, no. 6, pp. 357–364, 1998.

[185] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Research and Development in Information Retrieval, Proceedings of the International SIGIR Conference on*, 2003, pp. 267–273.

[186] J. Xue, G. Wichern, H. Thornburg, and A. Spanias, “Fast query by example of environmental sounds via robust and efficient cluster-based indexing,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 5–8, 2008.

[187] F. Yin, J. Wang, C. Guo, W. Liu, N. Zheng, and X. Li, “Nonnegative matrix factorization for eeg signal classification,” in *Advances in Neural Networks (ISSN)*, ser. Lecture Notes in Computer Science (LNCS), vol. 3174. Springer, 2001, pp. 470–475.

[188] S. Zafeiriou and A. Tefas, “Exploit discriminant information in nonnegative matrix factorization with application to frontal face verification,” *Neural Networks, IEEE Transactions on*, vol. 17, no. 3, pp. 683–695, May 2006.

[189] R. Zdunek and A. Cichocki, “Non-negative matrix factorization with quasi-newton optimization,” in *Artificial Intelligence and Soft Computing, International Conference on (ICAISC)*, vol. 4029, Jun 2006, pp. 870–879.

[190] ——, “Nonnegative matrix factorization with quadratic programming,” *Neurocomputing*, 2007.

[191] D. Zhang, Z.-H. Zhou, and S. Chen, “Non-negative matrix factorization on kernels,” in *Artificial Intelligence, Pacific Rim International Conference on*, ser. Lecture Notes in Artificial Intelligence (LNAI), vol. 4099. Springer, Aug 2006, pp. 404–412.

[192] J. Zhang, L. Wei, Q. Miao, and Y. Wang, “Image fusion based on nonnegative matrix factorization,” in *Image Processing, IEEE International Conference on (ICIP)*, vol. 2, Oct 2004, pp. 973–976.

[193] W.-S. Zheng, S. Li, J. Lai, and S. Liao, “On constrained sparse matrix factorization,” in *Computer Vision, IEEE International Conference on (ICCV)*, Oct. 2007, pp. 1–8.

Publication A

Theorems on Positive Data: on the uniqueness of NMF

Hans Laurberg, Mads Græsbøll Christensen, Mark D. Plumbley, Lars Kai Hansen, and Søren Holdt Jensen

The paper has been published in
Computational Intelligence and Neuroscience, 2008.

The layout has been revised.

Abstract

We investigate the conditions for which non-negative matrix factorization (NMF) is unique and introduce several theorems which can determine whether the decomposition is in fact unique or not. The theorems are illustrated by several examples showing the use of the theorems and theire limitations. We have shown that if a data matrix is a unique NMF matrix corrupted by additive noise this leads to a noisy estimation of the noise free unique solution. Finally, we use a stochastic view of NMF to analyse which characterization of the underlying model will result in a NMF with small estimation errors.

1 Introduction

Large quantities of positive data occur in research areas such as music analysis, text analysis, image analysis and probability theory. Before deductive science is applied to large quantities of data it is often appropriate to reduce data by preprocessing, e.g. by matrix rank reduction or by feature extraction. Principal component analysis is an example of such preprocessing. When the original data is non-negative, it is often desirable to preserve this property in the preprocessing. For example, elements in a power spectrogram, probabilities and pixel intensities should still be non-negative after the processing to be meaningful. This has led to the construction of algorithms for rank reduction of matrices and feature extraction generating non-negative output. Many of the algorithms are related to the non-negative matrix factorization (NMF) algorithm proposed by Lee and Seung [2, 3]. NMF algorithms factorize a non-negative matrix $\mathbf{V} \in \mathbb{R}_+^{n \times m}$ into two non-negative matrices $\mathbf{W} \in \mathbb{R}_+^{n \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times m}$:

$$\mathbf{V} \approx \mathbf{R} = \mathbf{W}\mathbf{H}. \tag{1}$$

There are no closed-form solutions to the problem of finding $\mathbf{W}$ and $\mathbf{H}$ given a $\mathbf{V}$, but Lee and Seung [2, 3] proposed two computationally efficient algorithms for minimizing the difference between $\mathbf{V}$ and $\mathbf{WH}$ for two different error functions. Later, numerous other algorithms have been proposed (see Berry et al. [4]).

An interesting question is whether the NMF of a particular matrix is unique. The importance of this question depends on the particular application of NMF. There can be two different viewpoints when using a model like NMF—either, one can believe that the model describes nature and that the variables $\mathbf{W}$ and $\mathbf{H}$ have a physical meaning or one can believe that the model can capture the part of interest even though there is not a one-to-one mapping between the parameters and the model, and the physical system. When using NMF, one can wonder whether $\mathbf{V}$ is a disturbed version of some underlying $\mathbf{WH}$ or whether the data is constructed by another model. Or, in other words, does a ground truth $\mathbf{W}$ and $\mathbf{H}$ exist. These questions are important in evaluating whether or not it is a problem that there is another NMF solution, $\mathbf{W}'\mathbf{H}'$, to the same data, i.e.

$$\mathbf{V} \approx \mathbf{R} = \mathbf{W}\mathbf{H} = \mathbf{W}'\mathbf{H}'. \tag{2}$$

If NMF is used even though the data is not assumed to be generated by (1), it may not be a problem that there are several other solutions. On the other hand, if one assumes that there exists a ground truth, it may be a problem if that model is not detectable, i.e. if it is not possible to find $\mathbf{W}$ and $\mathbf{H}$ from the data matrix $\mathbf{V}$.

The first article on the subject was the correspondence between Berman and Thomas. Berman [5] asked the question which in NMF terminology would be: Find a simple characterization of the class of non-negative matrices $\mathbf{R}$ for which an NMF exists. As we shall see, the answer by Thomas [6] can be transferred into a NMF uniqueness theorem.

The first article investigating the uniqueness of NMF is Donoho and Stodden [7]. They use convex duality to conclude that in some situations where the column vectors of $\mathbf{W}$ "describe parts", and for that reason are non-overlapping and thereby orthogonal, then the NMF solution is unique.

Simultaneously with the development of NMF, Plumbley [8] worked with non-negative independent component analysis, where one of the problems is to estimate a rotation matrix $\mathbf{Q}$ from observations $\mathbf{Q}\mathbf{s}$, where $\mathbf{s}$ is a non-negative vector. In this setup Plumbley investigates a property for a non-negative Independent and Identically Distributed (IID) vector $\mathbf{s}$ such that $\mathbf{Q}$ can be estimated. He shows that if the elements in $\mathbf{s}$ are grounded and a sufficiently large set of observation is used, then $\mathbf{Q}$ can be estimated. The uniqueness constraint in [8] is a statistical condition of $\mathbf{s}$.

The result in [8] is highly relevant for the NMF uniqueness due to the fact that in most cases new NMF solutions will have the form $\mathbf{W}\mathbf{Q}$ and $\mathbf{Q}^{-1}\mathbf{H}$ as described in Section 3. By using Plumbley's result twice, a restricted uniqueness theorem for NMF can be constructed.

In this paper, we investigate the circumstances under which NMF of an observed non-negative matrix is unique. We present novel necessary and sufficient conditions for the uniqueness. Several examples illustrating these conditions and their interpretations are given. Additionally, we show that NMF is robust to additive noise. More specifically, we show that it is possible to obtain accurate estimates of $\mathbf{W}$ and $\mathbf{H}$ from noisy data when the generating NMF is unique. Lastly, we consider the generating NMF as a stochastic process and show that particular classes of such processes almost surely result in unique NMFs.

This paper is structured as follows. Section 2 introduces the notation, some definitions, and basic results. A precise definition and two characterisations of a unique NMF are given in Section 3. The minimum constraints of $\mathbf{W}$ and $\mathbf{H}$ for a unique NMF are investigated in Section 4. Conditions and examples of a unique NMF are given in Section 5. In Section 6, it is shown that in situations where noise is added to a data matrix with a unique NMF it is possible to bound the error of the estimates of $\mathbf{W}$ and

H. A probabilistic view on the uniqueness is taken in Section 7. The implication of the theorems is discussed in Section 8, and Section 9 concludes the paper.

2 Fundamentals

We will here introduce convex duality that will be the framework of the paper, but first we shall define the notation to be used. Non-negative real numbers are denoted as $\mathbb{R}_+$, $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathrm{span}(\cdot)$ is the space spanned by the columns of a matrix. Each type of variables has its own font. For instance, a scalar is denoted x, a column vector is denoted $\mathbf{x}$, a row vector is denoted $\underline{\mathbf{x}}$, a matrix is denoted $\mathbf{X}$, a set is denoted $\mathcal{X}$, and a random variable is denoted $\mathscr{X}$. Moreover, $\mathbf{x}_i^j$ is the i'th index of the vector $\mathbf{x}^j$. When a condition for sets is used to describe a matrix, it is referring to the set of column vectors in the matrix. The NMF is symmetric in $\mathbf{W}^T$ and $\mathbf{H}$ so the theorems for one of the matrices may also be used for the other matrix.

In the paper, we make a geometric interpretation the NMF similar to that used in both [6] and [7]. For that, we need the following definitions.

Definition A.1 *The* ***positive span*** *is given by* $\mathrm{span}^+(\mathbf{b}^1, \cdots, \mathbf{b}^d) = \{\mathbf{v} = \sum_i \mathbf{b}^i \mathbf{a}_i \mid \mathbf{a} \in \mathbb{R}_+^d\}$.

In some literature, the positive span is called the conical hull.

Definition A.2 *A set $\mathcal{A}$ is called a* ***simplicial cone*** *if there is a set $\mathcal{B}$ such that $\mathcal{A} = \mathrm{span}^+(\mathcal{B})$. The* ***order*** *of a simplicial cone $\mathcal{A}$ is the minimum number of elements in $\mathcal{B}$.*

Definition A.3 *The* ***dual*** *to a set $\mathcal{A}$, denoted $\mathcal{A}^*$, is given by $\mathcal{A}^* = \{\mathbf{v} | \mathbf{v}^T\mathbf{a} \geq 0$ for all $\mathbf{a} \in \mathcal{A}\}$.*

The following lemma is easy to prove and will be used subsequently. For a more general introduction to convex duality, see [9].

Lemma A.4

a. *If $\mathcal{X} = \mathrm{span}^+(\mathbf{b}^1, \cdots, \mathbf{b}^d)$ then $\mathbf{y} \in \mathcal{X}^*$ if and only if $\mathbf{y}^T\mathbf{b}^n \geq 0$ for all n.*

b. *If $\mathcal{X} = \mathrm{span}^+(\mathbf{B}^T)$ and $\mathbf{B}^T = [\mathbf{b}^1, \cdots, \mathbf{b}^d]$ is invertible then $\mathcal{X}^* = \mathrm{span}^+(\mathbf{B}^{-1})$*

c. *If $\mathcal{Y} \subseteq \mathcal{X}$ then $\mathcal{X}^* \subseteq \mathcal{Y}^*$.*

d. *If $\mathcal{Y}$ and $\mathcal{X}$ are closed simplicial cones and $\mathcal{Y} \subset \mathcal{X}$, then $\mathcal{X}^* \subset \mathcal{Y}^*$.*

3 Dual Space and the NMF

In this section our definition of unique NMF and some general conditions for unique NMF is given. As a starting point let us assume that both $\mathbf{W}$ and $\mathbf{H}$ have full rank i.e. $r = \operatorname{rank}(\mathbf{R})$.

Let $\mathbf{W}'$ and $\mathbf{H}'$ be any matrices that fulfil, $\mathbf{R} = \mathbf{WH} = \mathbf{W}'\mathbf{H}'$. Then $\operatorname{span}(\mathbf{W}) = \operatorname{span}(\mathbf{R}) = \operatorname{span}(\mathbf{W}')$. The column vectors of $\mathbf{W}$ and $\mathbf{W}'$ are therefore both bases for the same spaces and as a result there exists a basis shift matrix $\mathbf{Q} \in \mathbb{R}^{r\times r}$ such that $\mathbf{W}' = \mathbf{WQ}$. It follows that $\mathbf{H}' = \mathbf{Q}^{-1}\mathbf{H}$. Therefore all NMF solutions where $r = \operatorname{rank}(\mathbf{R})$ are of the form $\mathbf{R} = \mathbf{WQQ}^{-1}\mathbf{H}$. In these situations, the ambiguity of the NMF is the $\mathbf{Q}$ matrix. Note that if $r > \operatorname{rank}(\mathbf{R})$ the above arguments do not hold because $\operatorname{rank}(\mathbf{W})$ can differ from $\operatorname{rank}(\mathbf{W}')$ and thereby $\operatorname{span}(\mathbf{W}) \neq \operatorname{span}(\mathbf{W}')$.

Example 1 *The following is an example of a $\mathbb{R}_+^{4\times 4}$ matrix of rank 3, where there are two NMF solutions but no $\mathbf{Q}$ matrix to connect the solutions*

$$\begin{bmatrix} 1&1&0&0\\ 1&0&1&0\\ 0&1&0&1\\ 0&0&1&1 \end{bmatrix} = \mathbf{R} = \underbrace{\mathbf{R}}_{\mathbf{W}}\underbrace{\mathbf{I}}_{\mathbf{H}} = \underbrace{\mathbf{I}}_{\mathbf{W}'}\underbrace{\mathbf{R}}_{\mathbf{H}'}. \tag{3}$$

We mention in passing that Thomas [6] uses this matrix to illustrate a related problem.

Lemma A.5 (Minc [10, lemma 1.1]) *The inverse of a non-negative matrix is non-negative if and only if it is a scaled permutation.*

Lemma A.5 shows that all NMF solution on the form $\mathbf{WQ}$ and $\mathbf{Q}^{-1}\mathbf{H}$ where $\mathbf{Q}$ is a scaled permutation are valid, and thereby that NMF only can be unique up to a permutation and a scaling. This leads to the following definition of unique NMF in this paper.

Definition A.6 *A matrix $\mathbf{R}$ has a **unique NMF** if the ambiguity is a permutation and a scaling of the columns in $\mathbf{W}$ and rows in $\mathbf{H}$.*

The scaling and permutation ambiguity in the uniqueness definition is a well-known ambiguity that occurs in many blind source separation problems. With this definition of unique NMF, it is possible to make the following two characterizations of the unique NMF.

Theorem A.7 *If $r = \operatorname{rank}(\mathbf{R})$, an NMF is unique if and only if the positive orthant is the only r-order simplicial cone $\mathcal{Q}$ such that $\operatorname{span}^+(\mathbf{W}^T) \subseteq \mathcal{Q} \subseteq \operatorname{span}^+(\mathbf{H})^*$.*

Proof. The proof follows the analysis of the $\mathbf{Q}$ matrix above in combination with Lemma A.4.b. The theorem can also be proved by following the steps of the proof in Thomas [6]. ∎

Theorem A.8 (Donoho and Stodden [7]) *The NMF is unique if and only if there is only one r-order simplicial cone $\mathcal{Q}$ such that $\mathrm{span}^{+}(\mathbf{R}) \subseteq \mathcal{Q} \subseteq \mathcal{P}$, where $\mathcal{P}$ is the positive orthant.*

Proof. It follows directly from the definitions. ■ The first characterization is inspirited by [6] and the second characterization is implicit introduced in [7]. Note that the two characterizations of the unique NMF analyze the problem from two different viewpoints. Theorem A.7 takes a known $\mathbf{W}$ and $\mathbf{H}$ pair as the starting point and looks at the solution from the "inside", i.e. the r dimensional space of row vectors in $\mathbf{W}$ and column vectors in $\mathbf{H}$. Theorem A.8 looks at the problem from the "outside", i.e. the n dimensional column space of $\mathbf{R}$.

4 Matrix Conditions

If $\mathbf{R} = \mathbf{WH}$ is unique then both $\mathbf{W}$ and $\mathbf{H}$ have to be unique respectively, i.e. there is only one NMF of $\mathbf{W}$ and $\mathbf{H}$ namely $\mathbf{W} = \mathbf{WI}$ and $\mathbf{H} = \mathbf{IH}$. In this section, a necessary condition for $\mathbf{W}$ and $\mathbf{H}$ is given and a sufficient condition is shown.

The following definition will be shown to be a necessary condition for both the set of row vectors in $\mathbf{W}$ and column vectors in $\mathbf{H}$.

Definition A.9 *A set $\mathcal{S}$ of vectors in $\mathbb{R}^d_+$ is called **boundary close** if for all $j \neq i$ and $k > 0$ there is an element $\mathbf{s} \in \mathcal{S}$ such that*

$$\mathbf{s}_j < k\mathbf{s}_i.$$

In the case of closed sets, the boundary close condition is that $\mathbf{s}_j = 0$ and $\mathbf{s}_i \neq 0$. In this section the sets will be finite (and therefore closed), but in Section 7 the general definition above is needed.

Theorem A.10 *The set of row vectors in $\mathbf{W}$ have to be boundary close for the corresponding NMF to be unique*

Proof. If the set of row vectors in $\mathbf{W}$ are not boundary close there exist indexes $j \neq i$ and $k > 0$ such that the j'th element is always more than k times larger than the i'th element in the row vectors in $\mathbf{W}$. Let $\mathcal{Q} = \mathrm{span}^{+}(\mathbf{q}^1, \cdots, \mathbf{q}^r)$ where

$$\mathbf{q}^n = \begin{cases} \mathbf{e}^i + k\mathbf{e}^j & \text{if } n = i \\ \mathbf{e}^n & \text{otherwise} \end{cases} \tag{4}$$

and $\mathbf{e}^n$ denotes the n'th standard basis vector. This set fulfils the condition $\mathrm{span}^{+}(W^T) \subseteq \mathcal{Q} \subset \mathcal{P}$ and therefore by Theorem A.7 we conclude that the NMF cannot be unique. ■
That not only the row vectors of $\mathbf{W}$ with small elements determine the uniqueness can be seen from the following example.

Example 2 *The following is an example where $\bar{\mathbf{W}}$ is not unique but $\mathbf{W} = \left[\begin{smallmatrix} & \bar{\mathbf{W}} & \\ 3 & 1 & 1 \end{smallmatrix}\right]$ is unique.*

Let

$$\bar{\mathbf{W}} = \left[\begin{smallmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{smallmatrix}\right].$$

Here $\bar{\mathbf{W}}$ is boundary close but not unique since $\bar{\mathbf{W}} = \bar{\mathbf{W}}\mathbf{I} = \mathbf{I}\bar{\mathbf{W}}$. The uniqueness of $\mathbf{W} = \left[\begin{smallmatrix} & \bar{\mathbf{W}} & \\ 3 & 1 & 1 \end{smallmatrix}\right]$ can be verified by plotting the matrix as shown in Figure 1, and observe that the conditions of Theorem A.7 are fulfilled.

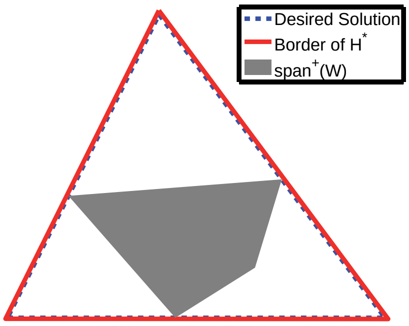


Figure 1: A three dimensional space is scaled such that the vectors are in the hyper plane: $\{\mathbf{p} | [1\,1\,1]\mathbf{p} = 1\}$. By the mapping to the hyper plane, a plane in $\mathbb{R}^3$ is mapped to a line and a simplicial cone is mapped to an area. In the figure, it can be observed that the dashed triangle (desired solution) is the only triangle (third order simplicial cone) that contains the shaded area (positive span of $\mathbf{W}$) while being within the solid border (the dual of $\mathbf{H}$). The NMF can be concluded to be unique by Theorem A.7.

In three dimensions, as in Example 2, it is easy to investigate whether a boundary close $\mathbf{W}$ is unique – if $\mathbf{W} = \mathbf{W}'\mathbf{H}'$ then $\mathbf{H}'$ can only have two types of structure: Either the trivial (desired) solution where $\mathbf{H}' = \mathbf{I}$ or a solution where only the diagonal of $\mathbf{H}'$ is zero. In higher dimensions, the number of combinations of non-trivial solutions increases and it becomes more complicated to investigate all possible non-trivial structures. For example, if $\bar{\mathbf{W}}$ is the matrix from Example 2, then the matrix

$$\widetilde{\mathbf{W}} = \left[\begin{smallmatrix} \bar{\mathbf{W}} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{W}} \end{smallmatrix}\right]$$

is boundary close and can be decomposed in several ways, e.g.

$$\widetilde{\mathbf{W}} = \left[\begin{smallmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{W}} \end{smallmatrix}\right] \left[\begin{smallmatrix} \bar{\mathbf{W}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{smallmatrix}\right] = \left[\begin{smallmatrix} \bar{\mathbf{W}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{smallmatrix}\right] \left[\begin{smallmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{W}} \end{smallmatrix}\right] = \left[\begin{smallmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{smallmatrix}\right] \left[\begin{smallmatrix} \bar{\mathbf{W}} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{W}} \end{smallmatrix}\right].$$

Instead of seeking necessary and sufficient conditions for a unique $\mathbf{W}$, a sufficient condition not much stronger than the necessary is given. In this sufficient condition we only focus on the row vectors of $\mathbf{W}$ with a zero (or very small) element.

Definition A.11 *A set of vectors $\mathcal{S}$ in $\mathbb{R}_+^d$ is called* ***strongly boundary close*** *if it is boundary close, and there exists a $z > 0$ and a numbering of the elements in the vectors such that for all $k > 0$ and $n \in \{1, \cdots, d-1\}$ there are $d-n$ vectors from $\mathcal{S}$, $\{\mathbf{s}^1, \cdots, \mathbf{s}^{d-n}\}$ that fulfil:*

1. $\mathbf{s}_n^j < k \sum_{i>n} \mathbf{s}_i^j$ *for all j and*
2. $\kappa_2\big(\big[\mathbf{b}^1, \quad \cdots \quad, \mathbf{b}^{d-n}\big]\big) \le z$*, where $\kappa_2(\cdot)$ is the "condition number" of the matrix defined as the ratio between the largest and smallest singular values [11, p. 81], $\mathbf{b}^j = \mathbf{P}_n \mathbf{s}^j$ and $\mathbf{P}_n \in \mathbb{R}^{d-n \times d}$ is a projection matrix that picks the $d-n$ last element of a vector in $\mathbb{R}^d$.*

Theorem A.12 *If $\mathrm{span}^+(\mathbf{W}^T)$ is strongly boundary close, then $\mathbf{W}$ is unique.*

The proof is quite technical and is therefore given in the Appendix. The most important to notice is that the necessary condition in Theorem A.10 and the sufficient conditions in Theorem A.12 are very similar. The first item in the strongly boundary close definition states that there has to be several vectors that has the small value. The second item ensures that the vectors with small value are linear independent in the last elements.

5 Uniqueness of $\mathbf{R}$

In this section, a condition for unique $\mathbf{V}$ is analyzed. First, Example 3 is used to investigate when a strongly boundary close $\mathbf{W}$ and $\mathbf{H}$ pair is unique. The section ends with a constraint for $\mathbf{W}$ and $\mathbf{H}$ that results in a unique NMF.

Example 3 *This is an investigation of uniqueness of $\mathbf{R}$ when $\mathbf{W}$ and $\mathbf{H}$ are given as:*

$$\mathbf{H} = \begin{bmatrix} \alpha & 1 & 1 & \alpha & 0 & 0 \\ 1 & \alpha & 0 & 0 & \alpha & 1 \\ 0 & 0 & \alpha & 1 & 1 & \alpha \end{bmatrix} \tag{5}$$

$$\mathbf{W} = \mathbf{H}^T, \tag{6}$$

where $0 < \alpha < 1$. Both $\mathbf{W}$ and $\mathbf{H}$ are strongly boundary close and the z parameter can be calculated as

$$z = \kappa_2\big(\big[\mathbf{b}^1, \quad \cdots \quad, \mathbf{b}^{d-n}\big]\big) \tag{7}$$

$$= \kappa_2\left(\begin{bmatrix} \alpha & 1 \\ 1 & \alpha \end{bmatrix}\right) = \frac{1+\alpha}{1-\alpha}. \tag{8}$$

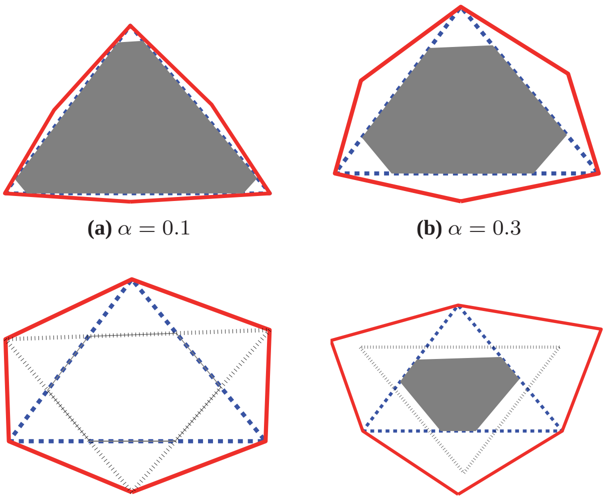
The equation above shows that small α will result in a z close to one and an α close to one results in a large z. In Figure 2, the matrix $\mathbf{R} = \mathbf{W}\mathbf{H}$ is plotted for $\alpha \in \{0.1,\ 0.3,\ 0.5,\ 0.7\}$. The dashed line is the desired solution and is repeated in all figures. It is seen that the shaded area $\mathrm{span}^+\left(\mathbf{W}^T\right)$ is decreasing when α increase, *and the solid border* $\operatorname{span}^{+}(\mathbf{H})^{*}$ *increases when* α *increases. For all* α*-values, both the shaded area and the solid border intersect with the dashed triangle. Therefore, it is not possible to get another solution by simply increase/decrease the desired solution. The figure shows that the NMF is unique for* $\alpha \in \{0.1,\ 0.3\}$ *and not unique for* $\alpha \in \{0.5,\ 0.7\}$ *where the alternative solution is shown with a dotted line. That the NMF are not unique for* $\alpha \in \{0.5,\ 0.7\}$ *can also be verified by selecting the* $\mathbf{Q}$ *to be the symmetric orthonormal matrix*

$$\mathbf{Q} = \mathbf{Q}^{T} = \mathbf{Q}^{-1} = \frac{1}{3}\begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}, \tag{9}$$

and see that both $\mathbf{WQ}$ *and* $\mathbf{Q}^{-1}\mathbf{H}$ *are non-negative. If* $\alpha = 0.3$ *then the matrix* $\mathbf{R}$ *is given by*

$$\mathbf{R} = \frac{1}{100}\begin{bmatrix} 109 & 60 & 30 & 9 & 30 & 100 \\ 60 & 109 & 100 & 30 & 9 & 30 \\ 30 & 100 & 109 & 60 & 30 & 9 \\ 9 & 30 & 60 & 109 & 100 & 30 \\ 30 & 9 & 30 & 100 & 109 & 60 \\ 100 & 30 & 9 & 30 & 60 & 109 \end{bmatrix}. \tag{10}$$

This shows that $\mathbf{R}$ *needs no zeros for the NMF to be unique.*



(a) $\alpha = 0.1$

(b) $\alpha = 0.3$

(c) $\alpha = 0.5$

(d) $\alpha = 0.7$

Figure 2: The figure shows data constructed as in Example 3 plotted in the same manner as in Figure 1 i.e. the dashed triangle is the desired solution, the solid line is the border of the dual of $\mathbf{H}$ and the shaded area is the positive span of $\mathbf{W}$. It can be seen that the NMF is unique when α equals 0.1 or 0.3 but not when α equals 0.5 or 0.7. In the cases where the NMF is not unique an alternative solution is shown with a dotted line.

In the example above, $\mathbf{W}$ equals $\mathbf{H}^T$ and thereby fulfils the same constraints. In many applications, the meaning of $\mathbf{W}$ and $\mathbf{H}$ differ, e.g. in music analysis where the column vectors of $\mathbf{W}$ are spectra of notes and $\mathbf{H}$ is a note activity matrix [12].

Next, it is investigated how to make an asymmetric uniqueness constraint.

Definition A.13 *A set of vectors in $\mathbb{R}^d$ is called **sufficiently spread** if for all j and $k > 0$ there is an element $\mathbf{s} \in \mathcal{S}$ such that*

$$\mathbf{s}_j > k \sum_{i \neq j} \mathbf{s}_i.$$

Note that in the definition for sufficiently spread the j'th element is larger than the sum in contrast to the strongly boundary close definition where the j'th element is smaller than the sum.

Lemma A.14 *The dual space to a sufficiently spread set is the positive orthant.*

Proof. A sufficiently spread set is non-negative and the positive orthant is therefore part of the dual set for any sufficiently spread set. Let $\mathbf{b}$ be a vector with a negative element in the j'th element and select

$$k = \frac{\sum_{i \neq j} |\mathbf{b}_i|}{-\mathbf{b}_j}. \tag{11}$$

Then there is an $\mathbf{s}$ in any sufficiently spread set such that $\mathbf{s}_j > k \sum_{i \neq j} \mathbf{s}_i$ and therefore

$$\begin{aligned} \mathbf{s}^T\mathbf{b} &= \mathbf{s}_j\mathbf{b}_j + \sum_{i \neq j} \mathbf{s}_i\mathbf{b}_i \leq \mathbf{s}_j\mathbf{b}_j + \Big(\sum_{i \neq j} \mathbf{s}_i\Big)\Big(\sum_{i \neq j} |\mathbf{b}_i|\Big) \\ &= -\mathbf{b}_j(-\mathbf{s}_j + k \sum_{i \neq j} \mathbf{s}_i) < 0. \end{aligned} \tag{12}$$

The $\mathbf{b}$ is therefore not in the dual to any sufficiently spread set. ■

In the case of finite sets, the sufficiently spread condition is the same as the requirement for a scaled version of all the standard basis vectors to be part of the sufficiently spread set. It is easy to verify that a sufficiently spread set also is strongly boundary close and that the z parameter is one.

Theorem A.15 *If a pair $[\mathbf{W}^T, \ \mathbf{H}]$ is sufficiently spread and strongly boundary close, then the NMF of $\mathbf{R} = \mathbf{W}\mathbf{H}$ is unique.*

Proof. Lemma A.14 states that the dual set of a sufficiently spread set is the positive orthant,

$$\operatorname{span}^+(\mathbf{H})^* = \mathcal{P} = \operatorname{span}^+(\mathbf{I})^*. \tag{13}$$

Theorem A.12 state that $\mathbf{WI}$ is unique and by using Equation 13 and Theorem A.7 we conclude that $\mathbf{R} = \mathbf{WH}$ is unique. ■

Theorem A.15 is a stronger version of the results of Donoho and Stodden [7, Theorem 1]. The theorem in [7] also assumes that $\mathbf{H}$ is sufficiently spread but the condition for $\mathbf{W}^T$ is stronger than the strongly boundary close assumption.

6 Perturbation Analysis

In the previous sections, we have analyzed situations where there has been a unique solution. In this section, it is shown that in some situations the non-uniqueness can be seen as estimation noise on $\mathbf{W}$ and $\mathbf{H}$. The error function that describes how close an estimated $[\mathbf{W}', \mathbf{H}]'$ pair is to the true $[\mathbf{W}, \mathbf{H}]$ pair is

$$J_{(\mathbf{W},\mathbf{H})}(\mathbf{W}', \mathbf{H}') \quad = \quad \min_{\mathbf{P},\mathbf{D}} \big(\|\mathbf{W} - \mathbf{W}'(\mathbf{DP})\|_F \; + \; \big\|\mathbf{H} - (\mathbf{DP})^{-1}\mathbf{H}'\big\|_F \big), \tag{14}$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}$ is a diagonal matrix.

Theorem A.16 *Let $\mathbf{R} = \mathbf{WH}$ be a unique NMF. Given some $\epsilon > 0$ there exists a $\delta > 0$ such that any non-negative $\mathbf{V} = \mathbf{R} + \mathbf{N}$ where $\|\mathbf{N}\|_F < \delta$ fulfils*

$$J_{(\mathbf{W},\mathbf{H})}(\mathbf{W}', \mathbf{H}') < \epsilon, \tag{15}$$

where

$$[\mathbf{W}', \mathbf{H}'] = \underset{\mathbf{W}' \in \mathbb{R}_+^{n \times r},\, \mathbf{H}' \in \mathbb{R}_+^{r \times m}}{\arg\min} \|\mathbf{V} - \mathbf{W}'\mathbf{H}'\|_F \,. \tag{16}$$

The proof is given in the appendix. The theorem states that if the observation is corrupted by additive noise then it will result in noisy estimation of $\mathbf{W}$ and $\mathbf{H}$. Moreover, Theorem A.16 shows that if the noise is small then it will result in small estimation errors[1].

Example 4 *This example investigates the connection between the additive noise in $\mathbf{V}$ and the estimation error on $\mathbf{W}$ and $\mathbf{H}$. The column vectors in $\mathbf{W}$ are basis pictures of a man, a dog and the sun as shown in Figure 3 a-c. In Figure 3 d, the sum of the three basis pictures are shown. The matrix $\mathbf{H}$ is the set of all combinations of the pictures, i.e.*

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}.$$

Theorem A.15 can be used to conclude that the NMF of $\mathbf{R} = \mathbf{WH}$ is unique because both $\mathbf{W}^T$ and $\mathbf{H}$ are sufficiently spread and thereby also strongly boundary close.

[1] In this section the Frobenius norm is used in (14) and (16) to make Theorem A.16 concrete. Theorem A.16 is also valid with the same proof if any continuous metric is used instead of the Frobenius norm in those equations.

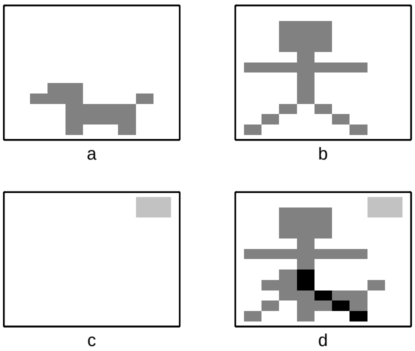


Figure 3: The three basis pictures: (a) a dog, (b) a man and (c) the sun, from Example 4 individually and summed (d).

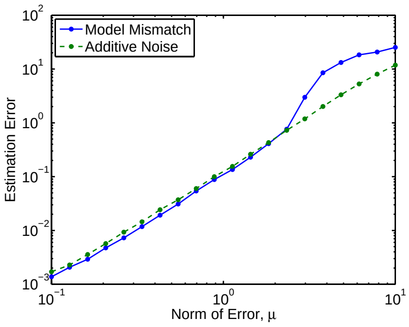


Figure 4: The graph shows the connection between the norm of the additive error $\|\mathbf{N}\|_F$ and estimation error of the underlying model $J_{(\mathbf{W},\mathbf{H})}(\mathbf{W}',\mathbf{H}')$. The two noise matrices from Example 4, $\mathbf{N}_N$ and $\mathbf{N}_M$, are plotted. In this example, the curves are aligned for small errors and for larger errors the model error $\mathbf{N}_M$ results in much larger estimation errors.

In the example, two different noise matrices, $\mathbf{N}_N$ and $\mathbf{N}_M$, are used. The $\mathbf{N}_N$ matrix model noisy observation and has elements that are random uniform IID. The $\mathbf{N}_M$ matrix contains elements that are minus one in the positions where $\mathbf{R}$ has elements that are two and zero elsewhere i.e. $\mathbf{N}_M$ is minus one in the positions where the dog and the man are overlapping. In this case the error matrix $\mathbf{N}_M$ simulates a model mismatch that occurs in the following two types of real world data. If the data set is composed of pictures, the basis pictures will be overlapping and a pixel in $\mathbf{V}$ will consist of one basis picture and not a mixture of the overlapping pictures. If the data is a set of amplitude spectra, the true model is addition of complex values and not, addition of the amplitudes.

The estimation error of the factorization $J_{(\mathbf{W},\mathbf{H})}(\mathbf{W}',\mathbf{H}')$ is plotted in Figure 4, *when the norm of the error matrix is μ i.e. $\mathbf{V} = \mathbf{WH} + \frac{\mathbf{N}}{\|\mathbf{N}\|_F}\mu$. An estimate of the $[\mathbf{W}', \mathbf{H}']$ pair, is calculated by using the iterative algorithm for Frobenius norm minimizing by Lee and Seung [3]. The algorithm is run for 500 iterations and is started from 100 different positions. The decomposition that minimize $\|\mathbf{V} - \mathbf{W}'\mathbf{H}'\|_F$ is chosen, and $J_{(\mathbf{W},\mathbf{H})}(\mathbf{W}', \mathbf{H}')$ is calculated numerically. Figure 4 shows that when the added error is small, it is possible to estimate the underlying parameters. When the norm of added noise matrix increases, the behaviour of the two noise matrices, $\mathbf{N}_N$ and $\mathbf{N}_M$, differ. For $\mathbf{N}_N$, the error of the estimate increases slowly with the norm of the added matrix while the estimation error for $\mathbf{N}_M$ increases dramatically when the norm is larger than 2.5. In the simulation we have made the following observation that can explain the difference in the performance of the two type of noise. When $\mathbf{N}_N$ is used, the basis pictures remain noisy versions of the man, the dog and the sun. When $\mathbf{N}_M$ is used and the norm is larger than 2.5, the basis pictures are the man excluding the overlap, the dog excluding the overlap and the overlap. Another way to describe the difference is that the rank of $\mathbf{N}_M$ is one and the disturbanc is in one dimension, where $\mathbf{N}_N$ is full rank and disturbanc is in many dimensions.*

Corollary A.17 *Let $\mathbf{R} = \mathbf{WH}$ be a unique NMF and $\mathbf{V} = \tilde{\mathbf{W}}\tilde{\mathbf{H}}$ where $\tilde{\mathbf{W}} = \mathbf{W} + \mathbf{N}_W$ and $\tilde{\mathbf{H}} = \mathbf{H} + \mathbf{N}_H$. Given $\mathbf{R}$ and $\epsilon > 0$ there exists a $\delta > 0$ such that if the largest absolute value of both $\mathbf{N}_W$ and $\mathbf{N}_H$ is smaller than δ then*

$$J_{(\tilde{\mathbf{W}},\tilde{\mathbf{H}})}(\mathbf{W}', \mathbf{H}') < \epsilon, \tag{17}$$

where $\mathbf{W}'$, $\mathbf{H}'$ are any NMF of $\mathbf{V}$.

Proof. This follows directly from Theorem A.16. ■ The corollary can be used in situations where there are small elements in $\mathbf{W}$ and $\mathbf{H}$ but no (or not enough) zero elements – as in the following example.

Example 5 *Let $\mathbf{R} = \mathbf{WH}$, where $\mathbf{W}$, $\mathbf{H}$ is generated as in Example 3. Let all elements in both $\mathbf{N}_W$ and $\mathbf{N}_H$ be equal to η. In Figure 5, $\mathbf{V}$ is plotted when $\alpha = 0.3$ and $\eta = \{0.01,\ 0.05,\ 0.10,\ 0.15\}$. In this example neither the shaded area nor the solid border intersects with the desired solution. Therefore, it is possible to get other solutions by simply increasing/decreasing the desired solution. For $\eta = \{0.01,\ 0.05\}$ the corners of the solutions are close to the corners of the desired solution. When $\eta = 0.1$, the corners can be placed most places on the solid border and still form a triangle that contains the shaded area. When $\eta = 0.15$ the corners can be anywhere on the solid border.*

7 Probability and Uniqueness

In this section, the row vectors of $\mathbf{W}$ and the column of $\mathbf{H}$ are seen as results of two random variables. Characteristics of the sample space (the possible outcomes) of a random variables that leads to unique NMF will be investigated.

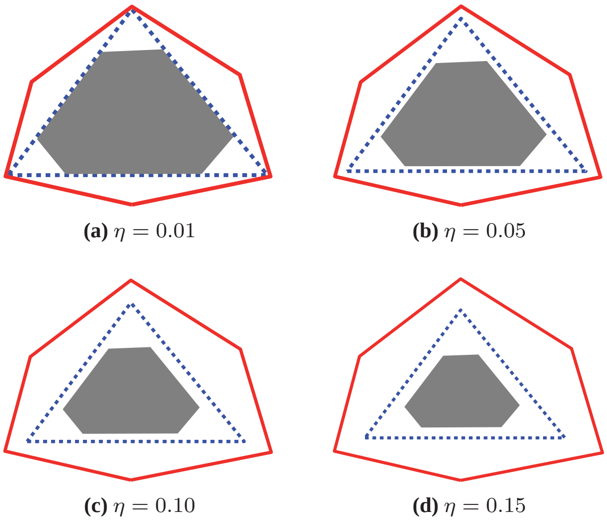


Figure 5: Data constructed as in Example 5 plotted in the same manner as in Figure 1 i.e. the dashed triangle is the desired solution, the solid line is the border of the dual of $\mathbf{H}$ and the shaded area is the positive span of $\mathbf{W}$. In all the plots, α equals 0.3 and η equals 0.01, 0.05, 0.1 and 0.15.

Theorem A.18 *Let the row vectors of $\mathbf{W}$ be generated by the random variable $\mathscr{X}_W$ and the column vectors of $\mathbf{H}$ be generated by a random variable $\mathscr{X}_H$. If the sample space of $\mathscr{X}_W$ is strongly boundary close and the sample space of $\mathscr{X}_H$ is sufficiently spread, then for all $\epsilon > 0$ and $k < 1$ there exist $N\epsilon$ and $M\epsilon$ such that*

$$p\left(\min_{\mathbf{D},\mathbf{P}}(\|\mathbf{DPQ} - \mathbf{I}\|_F) < \epsilon\right) > k, \tag{18}$$

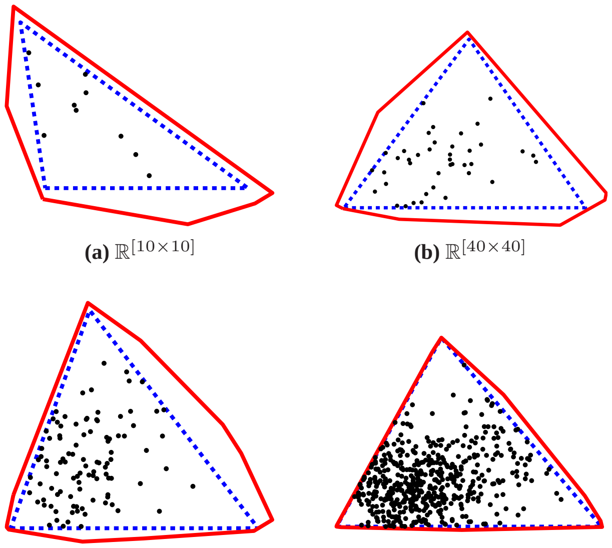
where $\mathbf{Q}$ is any matrix such that $\mathbf{WQ}$ and $\mathbf{Q}^{-1}\mathbf{H}$ are non-negative and the data size $\mathbf{R} \in \mathbb{R}_+^{n\times m}$ is such that $n > N\epsilon$ and $m > M\epsilon$.

Proof. If the data is scaled, $\mathbf{D}_1\mathbf{R}\mathbf{D}_2$, it does not change the non-uniqueness of the solutions if it is measured by the $\mathbf{Q}$ matrix. The proof is therefore done on the normalized versions of $\mathbf{W}$ and $\mathbf{H}$. Let $\mathscr{Y}_W$ and $\mathscr{Y}_H$ be the normalized version of $\mathscr{X}_W$ and $\mathscr{X}_H$. There exist a finite sets $\bar{\mathbf{W}}$ and $\bar{\mathbf{H}}$ of vectors in the closure of $\mathscr{Y}_W$ and $\mathscr{Y}_H$ that are strongly boundary close and sufficiently spread. By Theorem A.15 it is known that $\bar{\mathbf{V}} = \bar{\mathbf{W}}\bar{\mathbf{H}}$ is unique. By increasing the number vectors sampled from $\mathscr{Y}_W$ and $\mathscr{Y}_H$, for any $\epsilon' > 0$ there will be two subset of the vectors, $\mathbf{W}'$ and $\mathbf{H}'$, that with a probability larger that any $k < 1$ will fulfil

$$\epsilon' > \left\|\bar{\mathbf{W}} - \mathbf{W}'\right\|_F + \left\|\bar{\mathbf{H}} - \mathbf{H}'\right\|_F .$$

It is possible to use Corollary A.17 on this subset. The fact that limiting $\min_{\mathbf{D},\mathbf{P}}(\|\mathbf{DPQ}-\mathbf{I}\|_F)$ is equivalent to limiting (17) when the vectors are normalized finishes the proof. $\blacksquare$

Example 6 *Let all the elements in* $\mathbf{H}$ *be exponential IID and therefore generated with a sufficiently spread sample space. Additionally, let each row in* $\mathbf{W}$ *be exponential IID plus a random vector with the sample space* $\left\{\begin{pmatrix}0\\1\\1\end{pmatrix},\begin{pmatrix}1\\0\\1\end{pmatrix},\begin{pmatrix}1\\1\\0\end{pmatrix}\right\}$ *and thereby strongly boundary close. In Figure 6, the above variables are shown for the following four matrix size* $\mathbf{R}\in\{\mathbb{R}^{10\times10},\mathbb{R}^{40\times40},\mathbb{R}^{100\times100},\mathbb{R}^{500\times500}\}$.



(a) $\mathbb{R}^{[10\times10]}$

(b) $\mathbb{R}^{[40\times40]}$

(c) $\mathbb{R}^{[100\times100]}$

(d) $\mathbb{R}^{[500\times500]}$

Figure 6: The figure shows data constructed as in Example 6 plotted in the same manner as the previous figure with the exception that each row vector of $\mathbf{W}$ is plotted instead of the positive span of the vectors. The size of $\mathbf{R}$ is shown under each plot.

8 Discussion

The approach in this paper is to investigate when non-negativity leads to uniqueness in connection with NMF, $\mathbf{V}\approx\mathbf{R}=\mathbf{WH}$. Non-negativity is the only assumption for the theorems, and the theorems therefore cannot be used as argument for a NMF to be non-unique if there is additional information about $\mathbf{W}$ or $\mathbf{H}$. An example where there are stronger uniqueness results is the sparse NMF algorithm of Hoyer [13] built on the assumption that the row vectors in $\mathbf{H}$ have known ratios between the L_1 norm and the L_2 norm. Theis et al. [14] have investigated uniqueness in this situation and shown strong uniqueness results. Another example is data matrices with an added constant on each row. For this situation the Affine NMF algorithm [15] can make NMF unique even though the setup violates Theorem A.10 in this paper.

As shown in Figure 4, the type of noise influences greatly on the error curves. In applications where noise is introduced because the additive model does not hold, as for example when $\mathbf{V}$ is pictures or spectra, it is possible to influence the noise by making a non-linear function on the elements of $\mathbf{V}$. Such a non-linear function is introduced in [16] and experiments show that it improves the results. A theoretical framework to finding good non-linear functions will be interesting to investigate.

The sufficiently spread condition defined in Section 5 has an important role for unique NMF due to Lemma A.14. The sufficiently spread assumption is seen indirectly in related areas where it also leads to unique solutions, e.g. in [8] where the groundedness assumption leads to variables with a sufficiently spread sample space. If the matrix $\mathbf{H}$ is sufficiently spread then the columns in $\mathbf{W}$ will occur (almost) alone as columns in $\mathbf{V}$. Deville [17] uses the “occur alone” assumption, and thereby sufficiently spread assumption, to make blind source separation possible.

9 Conclusion

We have investigated the uniqueness of NMF from three different viewpoints,

- uniqueness in noise free situations,
- the estimation error of the underlying model when a matrix with unique NMF is added with noise and
- the random processes that lead to matrices where the underlying model can be estimated with small errors.

By doing this we have shown that it is possible to make many novel and useful characterisations that can be used as a theoretical underpinning for using the numerous NMF algorithms. There are several open issues in all the three viewpoints that, if addressed, will give a better understanding of Non-negative Matrix Factorization.

10 Appendix

Proof. [Proof of Theorem A.12] The theorem state that $\mathbf{W} = \mathbf{W}\mathbf{I}$ is a unique NMF. To proof this it is shown that the condition for Theorem A.7 is fulfilled. The positive orthant is self dual ($\mathbf{I} = \mathbf{I}^{-1}$) and thereby $\mathcal{Q} \subseteq \mathcal{P}$ where $\mathcal{Q}$ is an r order simplicial cone that contains $\operatorname{span}^+(\mathbf{W}^T)$. Let the set of row vectors in $\mathbf{W}$ be denoted $\mathcal{W}$. An r order simplicial cone, like $\mathcal{Q}$, is a closed set, and it therefore needs to contain the closure of $\mathcal{W}$ denoted by $\bar{\mathcal{W}}$. The two items in the definition A.11 of strongly boundary close can be reformulated for $\bar{\mathcal{W}}$ that contains the border:

1. $\mathbf{s}_n^j = 0$ for all j
2. the vectors $\left[\mathbf{b}^1, \quad \cdots \quad , \mathbf{b}^{d-n}\right]$ are linearly independent.

The rest of the proof follows by induction. If $r = 2$ then $\bar{\mathcal{W}} = \mathcal{P}$ and therefore unique. Let therefore $r > 2$. There are $r-1$ linearly independent vectors in $\bar{\mathcal{W}}$ that has zero as the first element and $r-1$ of the basis vectors therefore have zero in the firs elements. In other words, is there only one basis vector with a non-zero first element. Let's us call this vector $\mathbf{b}^1$. For all $j > 1$ there is a vector in $\bar{\mathcal{W}}$ that is non-negative in the first element and zero the j'th element, so all the elements in $\mathbf{b}^1$ except the first have to be zero. The proof is completed by seeing that if the first element is removed from the vectors in $\bar{\mathcal{W}}$, it is still strongly boundary close and the problem is therefore the $r-1$ dimensional problem. ■

Proof. [Proof of Theorem A.16] Let $\mathcal{G}$ be the open set of all $\mathbf{W}'$, $\mathbf{H}'$ pairs that are close to $\mathbf{W}$ and $\mathbf{H}$

$$\mathcal{G} = \left\{[\mathbf{W}', \mathbf{H}'] \middle| J_{(\mathbf{W},\mathbf{H})}(\mathbf{W}', \mathbf{H}') < \epsilon\right\}. \tag{19}$$

Let $\bar{\mathcal{G}}$ be the set of all non-negative $\breve{\mathbf{W}}, \breve{\mathbf{H}}$ pairs that are not in $\mathcal{G}$ and where $\max(\breve{\mathbf{W}}, \breve{\mathbf{H}}) \leq \sqrt{1+\max(\mathbf{R})}$. The uniqueness of $\mathbf{R}$ ensures that

$$\left\|\mathbf{R} - \breve{\mathbf{W}}\breve{\mathbf{H}}\right\|_F > 0, \tag{20}$$

for all $[\breve{\mathbf{W}}, \breve{\mathbf{H}}] \in \bar{\mathcal{G}}$. The fact that the Frobenius norm is continuous, $\bar{\mathcal{G}}$ is a closed bounded set and the statement above is positive ensure that

$$\min_{[\breve{\mathbf{W}},\breve{\mathbf{H}}]\in\bar{\mathcal{G}}}(\left\|\mathbf{R} - \breve{\mathbf{W}}\breve{\mathbf{H}}\right\|_F) = \delta' > 0, \tag{21}$$

since a continuous function attains its limits on a closed bounded set [18, Theorem 4.28]). The $\breve{\mathbf{W}}, \breve{\mathbf{H}}$ pairs that are not in $\mathcal{G}$ and where $\max(\breve{\mathbf{W}}, \breve{\mathbf{H}}) > \sqrt{1+\max(\mathbf{R})}$ can either be transformed by a diagonal matrix into a matrix pair from $\bar{\mathcal{G}}$, $[\breve{\mathbf{W}}\mathbf{D}, \mathbf{D}^{-1}\breve{\mathbf{H}}] \in \bar{\mathcal{G}}$, having the same product ($\breve{\mathbf{W}}\breve{\mathbf{H}}$) or it can be transformed into a pair where both $\breve{\mathbf{W}}$ and $\breve{\mathbf{H}}$ have large elements, i.e.

$$\max(\breve{\mathbf{W}}\breve{\mathbf{H}}) > \sqrt{1+\max(\mathbf{R})}^2 = 1 + \max(\mathbf{R}),$$

and thereby $\left\|\mathbf{R} - \breve{\mathbf{W}}\breve{\mathbf{H}}\right\|_F > 1$.

Select δ to be be $\delta = \min(1, \delta')/2$. The error of the desired solution $\mathbf{R} = \mathbf{W}\mathbf{H}$ can be bounded by $\|\mathbf{V} - \mathbf{R}\|_F = \|\mathbf{N}\|_F < \delta$. Let $\breve{\mathbf{V}}$ be any matrix constructed by a non-negative matrix pair from $\bar{\mathcal{G}}$. Because of the way δ is selected $\left\|\breve{\mathbf{V}} - \mathbf{R}\right\|_F \geq 2\delta$. By the triangle inequality, we get

$$\begin{aligned}\left\|\breve{\mathbf{V}} - \mathbf{V}\right\|_F + \left\|\mathbf{V} - \mathbf{R}\right\|_F &\geq \left\|\breve{\mathbf{V}} - \mathbf{R}\right\|_F \\ \left\|\breve{\mathbf{V}} - \mathbf{V}\right\|_F &\geq \left\|\breve{\mathbf{V}} - \mathbf{R}\right\|_F - \left\|\mathbf{V} - \mathbf{R}\right\|_F \\ &> 2\delta - \delta = \delta > \left\|\mathbf{V} - \mathbf{R}\right\|_F .\end{aligned}$$

All solutions that are in $\bar{\mathcal{G}}$ therefore have a larger error than $\mathbf{WH}$ and will not be the minimizer of the error. ■

References

[1] H. Laurberg, "Uniqueness of non-negative matrix factorization," in *Proc. IEEE Workshop on Statistical Signal Processing*, 2007, pp. 44–48.

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

[3] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2000, pp. 556–562.

[4] M. W. Berry, M. Browne, A. N. Langville, P. V. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics and Data Analysis*, vol. 52, pp. 155–173, September 2007.

[5] A. Berman, "Problem 73-14, rank factorization of nonnegative matrices," *SIAM*, vol. 15, p. 655, 1973.

[6] L. Thomas, "Solution to problem 73-14, rank factorizations of nonnegative matrices," *SIAM*, vol. 16, pp. 393–Ű394, 1974.

[7] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*.

[8] M. Plumbley, "Conditions for nonnegative independent component analysis," *IEEE Signal Processing Lett.*, vol. 9, no. 6, pp. 177–180, 2002.

[9] R. T. Rockafellar, *Convex Analysis*, 1st ed. Princeton University Press, 1970.

[10] H. Minc, *Nonnegative Matrices*, 1st ed. John Wiley and Sons, 1988.

[11] G. H. Golub and C. F. V. Loan, *Matrix computations*, 3rd ed. Johns Hopkins University Press, 1996.

[12] P. Smaragdis and J. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2003, pp. 177–180.

[13] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, Nov 2004.

[14] F. Theis, K. Stadlthanner, and T. Tanaka, “First results on uniqueness of sparse non-negative matrix factorization,” in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2005.

[15] H. Laurberg and L. Hansen, “On affine non-negative matrix factorization,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. II, 2007, pp. 653–656.

[16] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, “Wind noise reduction using non-negative sparse coding,” in *Proc. IEEE Workshop on Machine Learning for Signal Processing*, Aug 2007, pp. 431–436.

[17] Y. Deville, “Temporal and time-frequency correlation-based blind source separation methods,” Nara, Japan, Apr. 2003, pp. 1059–1064.

[18] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Addison-Wesley, 1974.

Publication B

Non-negative Matrix Factorization with Gaussian Process Priors

Mikkel N. Schmidt and Hans Laurberg

The paper has been published in
Computational Intelligence and Neuroscience, 2008.

The layout has been revised.

Abstract

We present a general method for including prior knowledge in a non-negative matrix factorization (NMF), based on Gaussian process priors. We assume, that the non-negative factors in the NMF are linked by a strictly increasing function to an underlying Gaussian process, specified by its covariance function. This allows us to find NMF decompositions, that agree with our prior knowledge of the distribution of the factors, such as sparseness, smoothness, and symmetries. The method is demonstrated with an example from chemical shift brain imaging.

1 Introduction

Non-negative matrix factorization (NMF) [1, 2] is a recent method for factorizing a matrix as the product of two matrices, in which all elements are non-negative. NMF has found widespread application in many different areas including pattern recognition [3], clustering [4], dimensionality reduction [5], and spectral analysis [6, 7]. Many physical signals, such as pixel intensities, amplitude spectra, and occurence counts, are naturally represented by non-negative numbers. In the analysis of mixtures of such data, non-negativity of the individual components is a reasonable constraint. Recently, a very simple algorithm [8] for computing the NMF was introduced. This has initiated much research aimed at developing more robust and efficient algorithms.

Efforts have been made to enhance the quality of the NMF by adding further constraints to the decomposition, such as sparsity [9], spatial localization [10, 11], and smoothness [11, 12], or by extending the model to be convolutive [13, 14]. Many extended NMF methods are derived by adding appropriate constraints and penalty terms to a cost function. Alternatively, NMF methods can be derived in a probabilistic setting, based on the distribution of the data [6, 15–17]. These approaches have the advantage that the underlying assumptions in the model are made explicit.

In this paper we present a general method for using prior knowledge to improve the quality of the solutions in NMF. The method is derived in a probabilistic setting, and it is based on defining prior probability distributions of the factors in the NMF model in a Gaussian process framework. We assume that the non-negative factors in the NMF are linked by a strictly increasing function to an underlying Gaussian process, specified by its covariance function. By specifying the covariance of the underlying process, we can compute NMF decompositions that agree with our prior knowledge of the factors, such as sparseness, smoothness, and symmetries. We refer to the proposed method as non-negative matrix factorization with Gaussian process priors, or GPP-NMF for short.

2 NMF with Gaussian Process Priors

In the following we derive a method for including prior information in an NMF decomposition by assuming Gaussian process priors (for a general introduction to Gaussian processes, see e.g. Rasmussen and Williams [18].) In our approach, the Gaussian process priors are linked to the non-negative factors in the NMF by a suitable link function. To set up the notation, we start by deriving the standard NMF method as a maximum likelihood (ML) estimator and then move on to the maximum a posteriori (MAP) estimator. Then we discuss Gaussian process priors and introduce a change of variable that gives better optimization properties. Finally, we discuss the selection of the link function.

2.1 Maximum Likelihood NMF

The NMF problem can be stated as

$$\boldsymbol{X} = \boldsymbol{D}\boldsymbol{H} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{K\times L}$ is a data matrix that is factorized as the product of two elementwise non-negative matrices, $\boldsymbol{D} \in \mathbb{R}_+^{K\times M}$ and $\boldsymbol{H} \in \mathbb{R}_+^{M\times L}$, where $\mathbb{R}_+$ denotes the non-negative reals. The matrix $\boldsymbol{N} \in \mathbb{R}^{K\times L}$ is the residual noise.

There exists a number of different algorithms [8, 15–17, 19–21] for computing this factorization, some of which can be viewed as maximum likelihood methods under certain assumptions about the distribution of the data. For example, least squares NMF corresponds to i.i.d. Gaussian noise [17] and Kullback-Leibler NMF corresponds to a Poisson process [6].

The ML estimate of $\boldsymbol{D}$ and $\boldsymbol{H}$ is given by

$$\{\boldsymbol{D}_{\mathrm{ML}}, \boldsymbol{H}_{\mathrm{ML}}\} = \underset{\boldsymbol{D},\boldsymbol{H}\geq 0}{\arg\min}\, \mathcal{L}_{X|D,H}(\boldsymbol{D}, \boldsymbol{H}), \tag{2}$$

where $\mathcal{L}_{X|D,H}(\boldsymbol{D}, \boldsymbol{H})$ is the negative log likelihood of the factors.

Example 7 (Least squares NMF) *An example of a maximum likelihood NMF is the least squares estimate. If the noise is i.i.d. Gaussian with variance σ_N^2, the likelihood of the factors $\boldsymbol{D}$ and $\boldsymbol{H}$ can be written as*

$$p^{\mathrm{LS}}_{X|D,H}(\boldsymbol{X}|\boldsymbol{D}, \boldsymbol{H}) = \frac{1}{\left(\sqrt{2\pi}\sigma_N\right)^{KL}} \exp\left(-\frac{||\boldsymbol{X} - \boldsymbol{D}\boldsymbol{H}||_F^2}{2\sigma_N^2}\right). \tag{3}$$

The negative log likelihood, which serves as a cost function for optimization, is then

$$\mathcal{L}^{\mathrm{LS}}_{X|D,H}(\boldsymbol{D}, \boldsymbol{H}) \propto \frac{1}{2\sigma_N^2}||\boldsymbol{X} - \boldsymbol{D}\boldsymbol{H}||_F^2, \tag{4}$$

where we use the proportionality symbol to denote equality subject to an additive constant. To compute a maximum likelihood estimate of $\boldsymbol{D}$ and $\boldsymbol{H}$, the gradient of the negative log likelihood is useful

$$\nabla_H \mathcal{L}^{\mathrm{LS}}_{X|D,H}(\boldsymbol{D},\boldsymbol{H}) = \frac{1}{\sigma_N^2}\boldsymbol{D}^\top(\boldsymbol{D}\boldsymbol{H}-\boldsymbol{X}), \tag{5}$$

and the gradient with respect to $\boldsymbol{D}$, which is easy to derive, is similar because of the symmetry of the NMF problem. ▮

The ML estimate can be computed by multiplicative update rules based on the gradient [8], projected gradient descent [19], alternating least squares [20], Newton-type methods [21], or any other appropriate constrained optimization method.

2.2 Maximum a Posteriori NMF

In this paper, we propose a method to build prior knowledge into the solution of the NMF problem. We choose a prior distribution $\mathrm{p}_{D,H}(\boldsymbol{D},\boldsymbol{H})$ over the factors in the model, that captures our prior beliefs and uncertainties of the solution we seek. We then compute the maximum a posteriori (MAP) estimate of the factors. Using Bayes rule, the posterior is given by

$$\mathrm{p}_{D,H|X}(\boldsymbol{D},\boldsymbol{H}|\boldsymbol{X}) = \frac{\mathrm{p}_{X|D,H}(\boldsymbol{X}|\boldsymbol{D},\boldsymbol{H})\mathrm{p}_{D,H}(\boldsymbol{D},\boldsymbol{H})}{\mathrm{p}_X(\boldsymbol{X})}. \tag{6}$$

Since the numerator is constant, the negative log posterior is the sum of a likelihood term that penalizes model misfit, and a prior term that penalizes solutions that are unlikely under the prior

$$\mathcal{L}_{D,H|X}(\boldsymbol{D},\boldsymbol{H}) \propto \mathcal{L}_{X|D,H}(\boldsymbol{D},\boldsymbol{H}) + \mathcal{L}_{D,H}(\boldsymbol{D},\boldsymbol{H}). \tag{7}$$

The MAP estimate of $\boldsymbol{D}$ and $\boldsymbol{H}$ is

$$\{\boldsymbol{D}_{\mathrm{MAP}},\boldsymbol{H}_{\mathrm{MAP}}\} = \operatorname*{arg\,min}_{\boldsymbol{D},\boldsymbol{H}\geq 0} \mathcal{L}_{D,H|X}(\boldsymbol{D},\boldsymbol{H}), \tag{8}$$

and it can again be computed using any appropriate optimization algorithm.

Example 8 (Non-negative sparse coding) *An example of a MAP NMF is non-negative sparse coding (NNSC) [9, 22], where the prior on $\boldsymbol{H}$ is i.i.d. exponential, and the prior on $\boldsymbol{D}$ is flat with each column constrained to have unit norm*

$$p^{\mathrm{NNSC}}_{D,H}(\boldsymbol{D},\boldsymbol{H}) = \prod_{i,j}\lambda\exp\left(-\lambda\boldsymbol{H}_{i,j}\right), \quad ||\boldsymbol{D}_k|| = 1\ \forall k, \tag{9}$$

where $||\boldsymbol{D}_k||$ is the Euclidean norm of the k'th column of $\boldsymbol{D}$. This corresponds to the following penalty term in the cost function

$$\mathcal{L}_{D,H}^{\text{NNSC}}(\boldsymbol{D},\boldsymbol{H}) \propto \lambda \sum_{i,j} \boldsymbol{H}_{i,j}. \tag{10}$$

The gradient of the negative log prior with respect to $\boldsymbol{H}$ is then

$$\nabla_H \mathcal{L}_{D,H}^{\text{NNSC}} = \lambda, \tag{11}$$

and the gradient with respect to $\boldsymbol{D}$ is zero, with the further normalization constraint given in Equation (9). ▮

2.3 Gaussian Process Priors

In the following, we derive the MAP estimate under the assumption that the non-negative matrices $\boldsymbol{D}$ and $\boldsymbol{H}$ are independently determined by a Gaussian process [18] connected by a link function. The Gaussian process framework provides a principled and practical approach to the specification of the prior probability distribution of the factors in the NMF model. The prior is specified in terms of two functions: i) a covariance function that describes corellations in the factors and ii) a link function, that transforms the Gaussian process prior into a desired distribution over the non-negative reals.

We assume that $\boldsymbol{D}$ and $\boldsymbol{H}$ are independent, so that we may write

$$\mathcal{L}_{D,H}(\boldsymbol{D},\boldsymbol{H}) = \mathcal{L}_D(\boldsymbol{D}) + \mathcal{L}_H(\boldsymbol{H}). \tag{12}$$

In the following, we consider only the prior for $\boldsymbol{H}$, since the treatment of $\boldsymbol{D}$ is equivalent due to the symmetry of the NMF problem. We assume that there is an underlying variable vector, $\boldsymbol{h} \in \mathbb{R}^{LM}$, which is zero mean multivariate Gaussian with covariance matrix $\boldsymbol{\Sigma}_h$

$$\mathrm{p}_h(\boldsymbol{h}) = \left(2\pi|\boldsymbol{\Sigma}_h|^2\right)^{-\frac{1}{2}NL} \exp\left(-\frac{1}{2}\boldsymbol{h}^\top \boldsymbol{\Sigma}_h^{-1}\boldsymbol{h}\right), \tag{13}$$

and linked to $\boldsymbol{H}$ via a link function, f_h: $\mathbb{R}_+ \to \mathbb{R}$

$$\boldsymbol{h} = f_h\big(\operatorname{vec}(\boldsymbol{H})\big), \tag{14}$$

which operates element-wise on its input. The $\operatorname{vec}(\cdot)$ function in the expression stacks its matrix operand column by column. The link function should be strictly increasing, which ensures that the inverse exists. Later, we will further assume that the derivatives of f_h and f_h^{-1} exist. Under these assumptions, the prior over $\boldsymbol{H}$ is given by (using the change of variables theorem)

$$\begin{aligned}
\mathrm{p}_H(\boldsymbol{H}) &= \mathrm{p}_h\Big(f_h\big(\operatorname{vec}(\boldsymbol{H})\big)\Big) \; \Big|\mathcal{J}\Big(f_h\big(\operatorname{vec}(\boldsymbol{H})\big)\Big)\Big| && (15)\\
&\propto \exp\left(-\frac{1}{2}f_h\big(\operatorname{vec}(\boldsymbol{H})\big)^\top \boldsymbol{\Sigma}_h^{-1} f_h\big(\operatorname{vec}(\boldsymbol{H})\big)\right) \prod_i \big|f_h'\big(\operatorname{vec}(\boldsymbol{H})\big)\big|_i, && (16)
\end{aligned}$$

where $\mathcal{J}(\cdot)$ denotes the Jacobian determinant and f'_h is the derivative of the link function. The negative log prior is then

$$\mathcal{L}_H(\boldsymbol{H}) \propto \frac{1}{2} f_h\big(\mathrm{vec}\,(\boldsymbol{H})\big)^\top \boldsymbol{\Sigma}_h^{-1} f_h\big(\mathrm{vec}\,(\boldsymbol{H})\big) - \sum_i \log\big|f'_h\big(\mathrm{vec}\,(\boldsymbol{H})\big)\big|_i. \tag{17}$$

This expression can be combined with an appropriate likelihood function, such as the least squares likelihood in Equation (4), and be optimized to yield the MAP solution; however, in our experiments, we found that a more simple and robust algorithm can be obtained by making a change of variable as explained next.

2.4 Change of Optimization Variable

Instead of optimizing over the non-negative factors $\boldsymbol{D}$ and $\boldsymbol{H}$, we introduce the variables $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$, which are related to $\boldsymbol{D}$ and $\boldsymbol{H}$ by

$$\boldsymbol{D} = g_d(\boldsymbol{\delta}) = \mathrm{vec}^{-1}\left(f_d^{-1}(\boldsymbol{C}_d^\top \boldsymbol{\delta})\right), \quad \boldsymbol{H} = g_h(\boldsymbol{\eta}) = \mathrm{vec}^{-1}\left(f_h^{-1}(\boldsymbol{C}_h^\top \boldsymbol{\eta})\right), \tag{18}$$

where the $\mathrm{vec}^{-1}(\cdot)$ function maps its vector input into a matrix of appropriate size. The matrices $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$ are the matrix square roots (Cholesky decompositions) of the covariance matrices $\boldsymbol{\Sigma}_d$ and $\boldsymbol{\Sigma}_h$, such that $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are standard i.i.d. Gaussian.

This change of variable serves two purposes. First of all, we found that optimizing over the transformed variables was faster, more robust, and less prone to getting stuck in local minima. Second, the transformed variables are not constrained to be non-negative, which allows us to use existing unconstrained optimization methods to compute their MAP estimate.

The prior distribution of the transformed variable $\boldsymbol{\eta}$ is

$$\mathrm{p}_\eta(\boldsymbol{\eta}) = \mathrm{p}_H\big(g_h(\boldsymbol{\eta})\big)\,\big|\mathcal{J}\big(g_h(\boldsymbol{\eta})\big)\big| = \frac{1}{(2\pi)^{\frac{LM}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{\eta}^\top\boldsymbol{\eta}\right), \tag{19}$$

and the negative log prior is

$$\mathcal{L}_\eta(\boldsymbol{\eta}) \propto \frac{1}{2}\boldsymbol{\eta}^\top\boldsymbol{\eta}. \tag{20}$$

To compute the MAP estimate of the transformed variables, we must combine this expression for the prior (and a similar expression for the prior of $\boldsymbol{\delta}$) with a likelihood function, in which the same change of variable is made

$$\mathcal{L}_{\delta,\eta|X}(\boldsymbol{\delta},\boldsymbol{\eta}) = \mathcal{L}_{X|D,H}\big(g_d(\boldsymbol{\delta}), g_h(\boldsymbol{\eta})\big) + \frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\eta}^\top\boldsymbol{\eta}. \tag{21}$$

Then the MAP solution can be found by optimizing over $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$

$$\{\boldsymbol{\delta}_{\mathrm{MAP}}, \boldsymbol{\eta}_{\mathrm{MAP}}\} = \arg\min_{\boldsymbol{\delta},\boldsymbol{\eta}} \mathcal{L}_{\delta,\eta|X}(\boldsymbol{\delta},\boldsymbol{\eta}), \tag{22}$$

and, finally, estimates of $\boldsymbol{D}$ and $\boldsymbol{H}$ can be computed using Equation (18).

Example 9 ***Least squares non-negative matrix factorization with Gaussian process priors (GPP-NMF)***
If we use the least squares likelihood in Equation (4), the posterior distribution in Equation (21) is given by

$$\mathcal{L}^{\textit{LS-GPP}}_{\delta,\eta|X}(\boldsymbol{\delta},\boldsymbol{\eta}) = \frac{1}{2}\left(\sigma_N^{-2}||\boldsymbol{X} - g_d(\boldsymbol{\delta})g_h(\boldsymbol{\eta})||_F^2 + \boldsymbol{\delta}^\top\boldsymbol{\delta} + \boldsymbol{\eta}^\top\boldsymbol{\eta}\right) \tag{23}$$

The MAP estimate of $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ is found by minimizing this expression, for which the derivative is useful

$$\begin{aligned}&\nabla\eta\mathcal{L}^{\textit{LS-GPP}}_{\delta,\eta|X}(\boldsymbol{\delta},\boldsymbol{\eta}) = \\ &\quad\sigma_N^{-2}\Big(\operatorname{vec}\left(g_d(\boldsymbol{\delta})^\top(g_d(\boldsymbol{\delta})g_h(\boldsymbol{\eta}) - \boldsymbol{X})\right) \odot (f_h^{-1})'(\boldsymbol{C}_h^\top\boldsymbol{\eta})\Big)^\top\boldsymbol{C}_h + \boldsymbol{\eta},\end{aligned} \tag{24}$$

where $\odot$ denotes the Hadamard (element-wise) product. The derivative with respect to $\boldsymbol{\delta}$ is similar due to the symmetry of the NMF problem. ▮

2.5 Link Function

Any strictly increasing link function that maps the non-negative reals to the real line can be used in the proposed framework; however, in order to have a better probabilistic interpretation of the prior distribution, we propose a simple principle for choosing the link function. We choose the link function such that f_h^{-1} maps the marginal distribution of the elements of the underlying Gaussian process vector $\boldsymbol{h}$ into a specifically chosen marginal distribution of the elements of $\boldsymbol{H}$. Such an inverse function can be found as $f_h^{-1}(\boldsymbol{h}_i) = \mathrm{P}_H^{-1}\big(\mathrm{P}_h(\boldsymbol{h}_i)\big)$ where $\mathrm{P}(\cdot)$ denotes the marginal cumulative distribution functions (cdf).

Since the marginals of a Gaussian process are Gaussian, $\mathrm{P}_h(\boldsymbol{h}_i)$ is the Gaussian cdf, and, using Equation (13), the inverse link function is given by

$$f_h^{-1}(\boldsymbol{h}_i) = \mathrm{P}_H^{-1}\left(\frac{1}{2} + \frac{1}{2}\Phi\left(\frac{\boldsymbol{h}_i}{\sqrt{2}\sigma_i}\right)\right) \tag{25}$$

where σ_i^2 is the i'th diagonal element of $\boldsymbol{\Sigma}_h$ and $\Phi(\cdot)$ is the error function.

Example 10 (Exponential-to-Gaussian link function) *If we choose to have exponential marginals in $\boldsymbol{H}$, as in NNSC described in Example 8, we select P_H as the exponential cdf. The inverse link function is then*

$$f_h^{-1}(\boldsymbol{h}_i) = -\frac{1}{\lambda}\log\left(\frac{1}{2} - \frac{1}{2}\Phi\left(\frac{\boldsymbol{h}_i}{\sqrt{2}\sigma_i}\right)\right), \tag{26}$$

where λ is an inverse scale parameter. The derivative of the inverse link function, which is needed for the parameter estimation, is given by

$$(f_h^{-1})'(\boldsymbol{h}_i) = \frac{1}{\sqrt{2\pi}\sigma_i\lambda} \exp\left(\lambda f_h^{-1}(\boldsymbol{h}_i) - \frac{\boldsymbol{h}_i^2}{2\sigma_i^2}\right). \tag{27}$$

▮

Example 11 (Rectified-Gaussian-to-Gaussian link function) *Another interesting non-negative distribution is the rectified Gaussian given by*

$$p(x) = \begin{cases} \frac{2}{\sqrt{2\pi}s} \exp\left(-\frac{x^2}{2s^2}\right) & , \quad x \geq 0 \\ 0 & , \quad x < 0 \end{cases} \tag{28}$$

Using this pdf in Equation (25), the inverse link function is

$$f_h^{-1}(\boldsymbol{h}_i) = \sqrt{2}s\Phi^{-1}\left(\frac{1}{2} + \frac{1}{2}\Phi\left(\frac{\boldsymbol{h}_i}{\sqrt{2}\sigma_i}\right)\right), \tag{29}$$

and the derivative of the inverse link function is

$$(f_h^{-1})'(\boldsymbol{h}_i) = \frac{s}{2\sigma_i} \exp\left(\frac{f_h^{-1}(\boldsymbol{h}_i)^2}{2s^2} - \frac{\boldsymbol{h}_i^2}{2\sigma_i^2}\right). \tag{30}$$

▮

2.6 Summary of the GPP-NMF Method

The GPP-NMF method can be summarized in the following steps.

1. Choose a suitable negative log likelihood function $\mathcal{L}_{X|D,H}(\boldsymbol{D}, \boldsymbol{H})$ based on knowledge of the distribution of the data or the residual.

2. For each of the non-negative factors $\boldsymbol{D}$ and $\boldsymbol{H}$, choose suitable link and covariance functions according to your prior beliefs. If necessary, draw samples from the prior distribution to examine its properties.

3. Compute the MAP estimate of $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ by Equation (22) using any suitable unconstrained optimization algorithm.

4. Compute $\boldsymbol{D}$ and $\boldsymbol{H}$ using Equation (18).

Our Matlab implemention of the GPP-NMF method is available online [23].

3 Experimental Results

We will demonstrate the proposed method on two examples, first a toy example, and second an example taken from the chemical shift brain imaging literature.

In our experiments we use the least squares GPP-NMF described in Example 9 and the link functions described in Example 10–11. The specific optimization method used to compute the GPP-NMF MAP estimate is not the topic of this paper, and any unconstrained optimization algorithm could in principle be used. In our experiments we used a simple gradient descent with line search to perform a total of 1000 alternating updates of $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$, after which the algorithm had converged. For details of the implementation, see the accompanying Matlab code [23].

3.1 Toy Example

We generated a 100×200 data matrix, $\boldsymbol{Y}$, by taking a random sample from the GPP-NMF model with two factors. We chose the generating covariance function for both $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ as a Gaussian radial basis function (RBF),

$$\phi(i,j) = \exp\left(-\frac{(i-j)^2}{\beta^2}\right), \tag{31}$$

where i and j are two sample indices, and the length scale parameter, which determines the smoothness of the factors, was $\beta^2 = 100$. We set the covariance between the two factors to zero, such that the factors were uncorrelated. For the matrix $\boldsymbol{D}$ we used the rectified-Gaussian-to-Gaussian link function with $s = 1$, and for $\boldsymbol{H}$ we used the exponential-to-Gaussian link function with $\lambda = 1$. Finally, we added independent Gaussian noise with variance $\sigma_N^2 = 25$, which resulted in a signal-to-noise ratio of approximately -7 dB. The generated data matrix is shown in Figure 1.

We then decomposed the generated data matrix using four different methods:

1. **LS-NMF:** Standard least squares NMF [8]. This algorithm does not allow negative data points, so these were set to zero in the experiment.

2. **CNMF:** Constrained NMF [6, 7], which is a least squares NMF algorithm that allows negative observations.

3. **GPP-NMF: Correct prior:** The proposed method with correct link-functions, covariance matrix, and parameter values.

4. **GPP-NMF: Incorrect prior:** To illustrate the sensitivity of the method to prior assumptions, we evaluated the proposed method with incorrect prior assumptions: We switched the link functions, such that for $\boldsymbol{D}$ we used the exponential-to-Gaussian, and for $\boldsymbol{H}$ we used the rectified-Gaussian-to-Gaussian. We used an RBF covariance function with $\beta^2 = 10$ and $\beta^2 = 1000$ for $\boldsymbol{D}$ and $\boldsymbol{H}$ respectively.

The results of the decompositions are shown as reconstructed data matrices in Figure 1. All four methods find solutions that visually appear to fit the underlying data. Both LS-NMF and CNMF find non-smooth solutions, whereas the two GPP-NMF results are smooth in accordance with the priors. In the GPP-NMF with incorrect prior, the dark areas (high pixel intensities) appear too wide in the first axis direction and too narrow in the section axis direction, due to the incorrect setting of the covariance function. The GPP-NMF with correct prior is visually almost equal to the true underlying data.

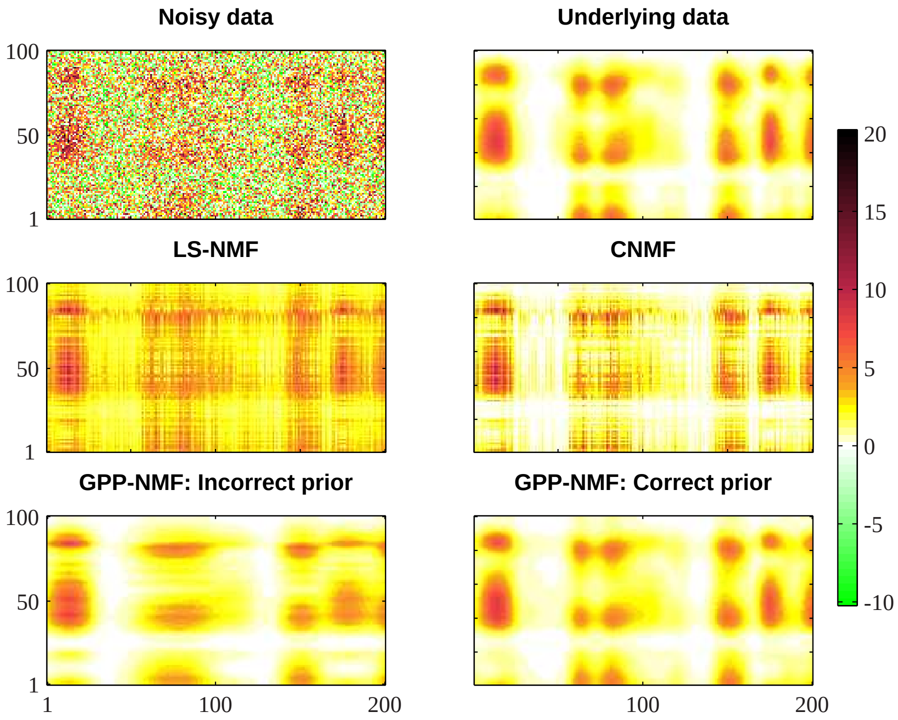


Figure 1: Toy example data matrix (upper left), underlying noise-free non-negative data (upper right), and estimates using the four methods described in the text. The data has a fairly large amount of noise and the underlying non-negative factors are smooth in both directions. The LS-NMF and CNMF decomposition are non-smooth, since these methods do not model of correlations in the factors. The GPP-NMF, which uses a smooth prior, finds a smooth solution. When using the correct prior, the soulution is very close to the true underlying data.

Plots of the estimated factors are show in Figure 2. The factors estimated by the LS-NMF and the CNMF methods appear noisy and are non-smooth, whereas the factors estimated by the GPP-NMF are smooth. The factors estimated by the LS-NMF method have a positive bias, because of the truncation of negative data. The GPP-NMF with incorrect prior has too many local extrema in the $\boldsymbol{D}$ factor and too few in the $\boldsymbol{H}$ factor due to the incorrect covariance functions. There are only minor difference between the result of the GPP-NMF with the correct prior and the underlying factors.

Measures of root mean squared error (RMSE) of the four decompositions are given in Figure 3. All four methods fit the noisy data almost equally well. (Note that, due to the additive noise with variance 25, a perfect fit to the underlying factors would result in a RMSE of 5 with respect to the noisy data.) The LS-NMF fits the data worst due to the truncation of negative data points, and the CNMF fits the data best, due to overfitting. With respect to the noise free data and the underlying factors, the RMSE is worst for the LS-NMF and best for the GPP-NMF with correct prior. The GPP-NMF with incorrect prior is better than both LS-NMF and CNMF in this case. This shows, that in this situation it better to use a prior which is not perfectly correct, compared to using no prior as in the LS-NMF and CNMF methods, (which corresponds to a flat prior over the non-negative reals and no correlations.)

3.2 Chemical Shift Brain Imaging Example

Next, we demonstrate the GPP-NMF method on ^{1}H decoupled ^{31}P chemical shift imaging data of the human brain. We use the data set from Ochs et al. [24], which has also been analyzed by Sajda et al. [6, 7]. The data set, which is shown in Figure 4, consists of 512 spectra measured on an $8 \times 8 \times 8$ grid in the brain.

Ochs et al. [24] use PCA to determine, that the data set is adequately described by two sources (which correspond to brain and muscle tissue.) They propose a bilinear Bayesian approach, in which they use a smooth prior over the constituent spectra, and force to zero the amplitude of the spectral shape corresponding to muscle tissue at 12 positions deep inside the head. Their approach produces physically plausible results, but it is computationally very expensive and takes several hours to compute.

Sajda et al. [6, 7] propose an NMF approach that is reported also to produce physically plausible results. Their method is several orders of magnitude faster, taking less than a second to compute. The disadvantage of the method of Sajda et al. compared to the Bayesian approach of Ochs et al. is, that it provides no mechanism for using prior knowledge to improve the solution.

The GPP-NMF approach we propose in this paper bridges the gap between the two previous approaches, in the sense that it is a relatively fast NMF approach, in which priors over the factors can be specified. These priors are specified by the choice of the link and covariance functions. We used prior predictive sampling to find reasonable settings of the the function parameters: We drew random samples from the prior distribution and examined properties of the factors and reconstructed data. We then manually adjusted the parameters of the prior to match our prior beliefs. An example of a random draw from the prior distribution is shown in Figure 5, with the parameters set as described below.

We assumed that the factors are uncorrelated, so the covariance between factors is

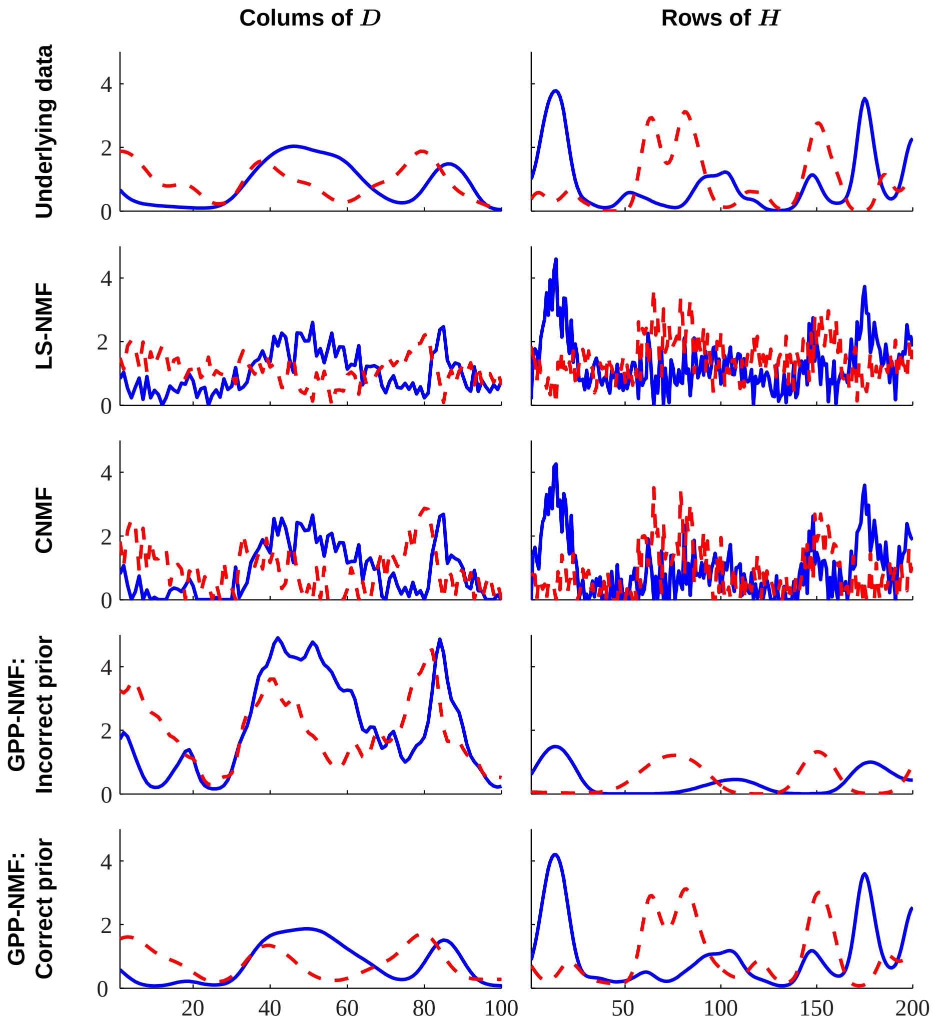


Figure 2: Underlying non-negative factors in the toy example: Columns of $\boldsymbol{D}$ (left) and rows of $\boldsymbol{H}$ (right). The factors found by the LS-NMF and the CNMF algorithm are noisy, whereas the factors found by the GPP-NMF method are smooth. When using the correct prior, the factors found are very similar to the true factors.

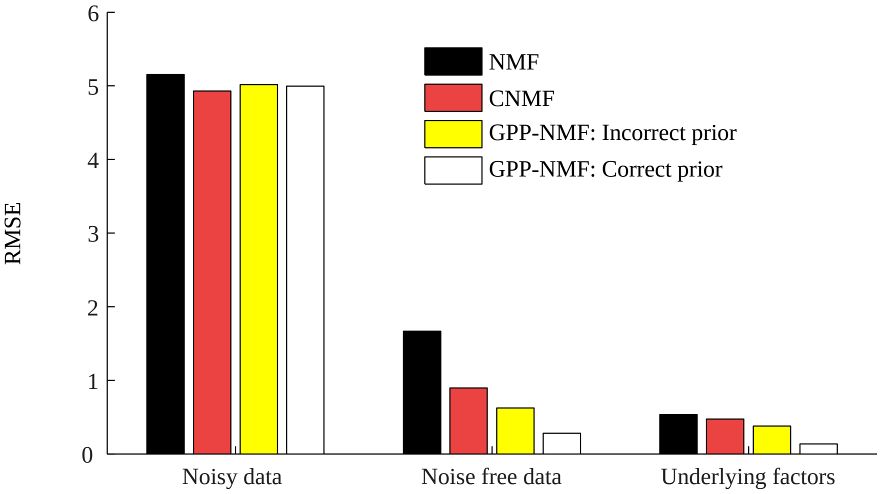


Figure 3: Toy example: Root mean squared error (RMSE) with respect to the noisy data, the underlying noise free data, and the true underlying non-negative factors. The CNMF solution fits the noisy data slightly better, but the GPP-NMF solution fits the underlying data much better.

zero. We used a Gaussian RBF covariance function for the constituent spectra, with a length scale of $\beta = 0.3$ parts per million (ppm), and we used the exponential-to-Gaussian link function with $\lambda_d = 1$. This gave a prior for the spectra that is sparse with narrow smooth peaks. For the amplitude at the 512 voxels in the head, we used a Gaussian RBF covariance function on the 3-D voxel indices, with length scale $\beta = 2$. Furthermore, we centered the left-to-right coordinate axis in the middle of the brain, and computed the RBF kernel on the absolute value of this coordinate, so that a left-to-right symmetry was introduced in the prior distribution. Again, we used the exponential-to-Gaussian link function, and we chose $\lambda_h = 2 \cdot 10^{-4}$ to match the overall magnitude of the data. This gave a prior for the amplitude distribution that is sparse, smooth, and symmetric. The noise variance was set to $\sigma_N^2 = 10^8$ which corresponds to the noise level in the data set.

We then decomposed the data set using the proposed GPP-NMF algorithm and, for comparison, reproduced the results of Sajda et al. [7] using their CNMF method. The results of the experiments are shown in Figure 4. An example of a random draw from the prior distribution is shown in Figure 5. The results of the CNMF is shown in Figure 6, and the results of the GPP-NMF is shown in Figure 7. The figures show the constituent spectra and the fifth axial slice of the spatial distribution of the spectra. The 8×8 spatial distributions are smoothed in the illustration, similar to the way the results are visualized in the literature.

The results show that both methods give physically plausible results. The main dif-

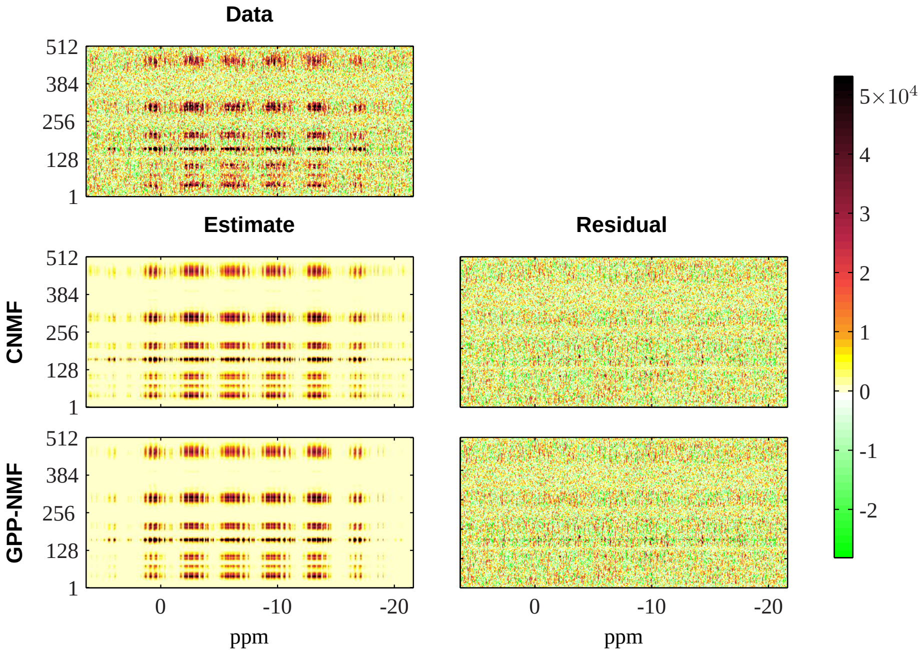


Figure 4: Brain imaging data matrix (top) along with the estimated decomposition and residual for the CNMF (middle) and GPP-NMF (bottom) method. In this view the results of the two decompositions are very similar, the data appears to be modeled equally well and the residuals are similar in magnitude.

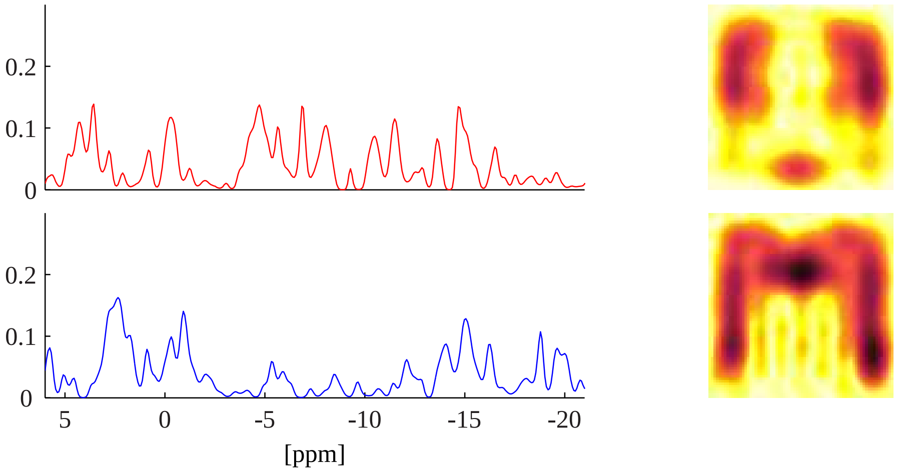


Figure 5: Brain imaging data: Random draw from the prior distribution with the parameters set as described in the text. The prior distribution of the constituent spectra (left) is exponential and smooth and the spatial distribution (right) in the brain is exponential, smooth, and has a left-to-right symmetry.

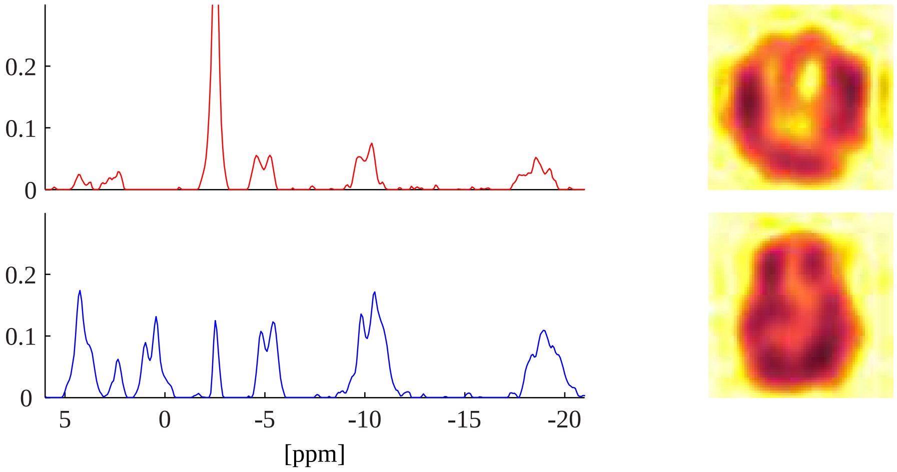


Figure 6: CNMF decomposition result. The recovered spectra are physically plausible, and the spatial distribution in the brain for the muscle (top) and brain (bottom) tissue is somewhat separated. Muscle tissue is primarily found near the edge of the skull, whereas brain tissue is primarliy found at the inside of the head.

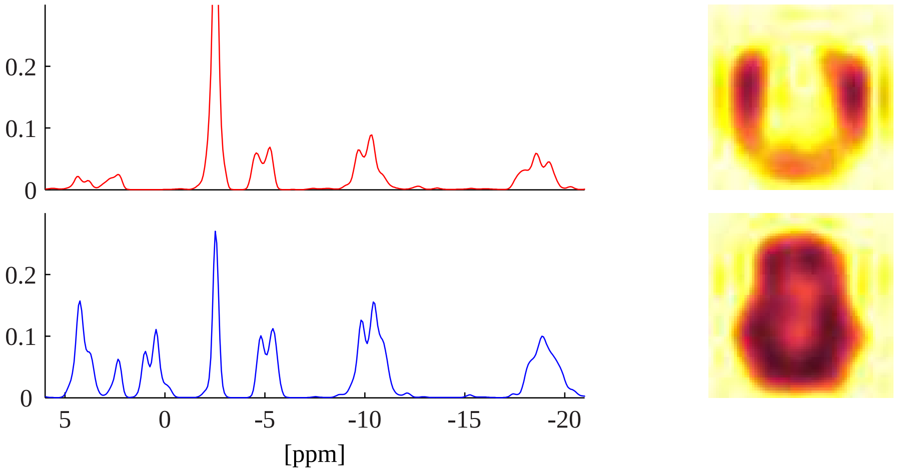


Figure 7: GPP-NMF decomposition result. The recovered spectra are very similar to the spectra found by the CNMF method, but they are slightly more smooth. The spatial distribution in the brain is highly separated between brain and muscle tissue, and it is more symmetric than the CNMF solution.

ference is that the spatial distribution of the spectra corresponding to muscle and brain tissue is much more separated in the GPP-NMF result, which is due to the exponential, smooth, and symmetric prior distribution. By including prior information, we obtain a solution, where the factor corresponding to muscle tissure is clearly located on the edge of the skull.

4 Conclusions

We have introduced a general method for exploiting prior knowledge in non-negative matrix factorization, based on Gaussian process priors, linked to the non-negative factors by a link function. The method can be combined with any existing NMF cost function that has a probabilistic interpretation, and any existing unconstrained optimization algorithm can be used to compute the maximum a posteriori estimate.

Experiments on toy data show, that with a suitable selection of the prior distribution of the non-negative factors, the GPP-NMF method gives much better results in terms of estimating the true underlying factors, both when compared to traditional NMF and CNMF.

Experiments on chemical shift brain imaging data show that the GPP-NMF method can be successfully used to include prior knowledge of the spectral and spatial distribution, resulting in better spatial separation between spectra corresponding to muscle and brain tissue.

5 Acknowledgments

We would like to thank Paul Sajda and Truman Brown for making the brain imaging data available to us. This research was supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–02–0092 and partly also by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), contract no. 506778.

References

[1] P. Paatero and U. Tapper, “Positive matrix factorization: A nonnegative factor model with optimal utilization of error-estimates of data values,” *Environmetrics*, vol. 5, pp. 111–126, 1994.

[2] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[3] L. Weixiang, Z. Nanning, and Y. Qubo, “Nonnegative matrix factorization and its applications in pattern recognition,” *Chinese Science Bulletin*, vol. 51, no. 1, pp. 7–18, Jan 2006.

[4] C. Ding, X. He, and H. D. Simon, “On the equivalence of nonnegative matrix factorization and spectral clustering,” in *Data Mining, Proceedings of SIAM International Conference on*, 2005, pp. 606–610.

[5] S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kita, “Dimensionality reduction using non-negative matrix factorization for information retrieval,” in *Systems, Man, and Cybernetics, IEEE International Conference on*, vol. 2, 2001, pp. 960–965.

[6] P. Sajda, S. Du, and L. Parra, “Recovery of constituent spectra using non-negative matrix factorization.” in *Wavelets: Applications in Signal and Image Processing, Proceedings of SPIE*, vol. 5207, 2003, pp. 321–331.

[7] 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,” *Medical Imaging, IEEE Transactions on*, vol. 23, no. 12, pp. 1453–1465, Dec 2004.

[8] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Neural Information Processing Systems, Advances in (NIPS)*, 2000, pp. 556–562.

[9] P. Hoyer, “Non-negative sparse coding,” in *Neural Networks for Signal Processing, IEEE Workshop on*, 2002, pp. 557–565.

[10] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. Cheng, “Learning spatially localized, parts-based representation,” in *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, vol. 1, Dec 2001, pp. 207–212.

[11] Z. Chen and A. Cichocki, “Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints,” Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep., 2005.

[12] T. Virtanen, “Sound source separation using sparse coding with temporal continuity objective,” in *International Computer Music Conference, ICMC*, 2003.

[13] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sourses from monophonic inputs,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science (LNCS), vol. 3195. Springer, Sep 2004, pp. 494–499.

[14] M. N. Schmidt and M. Mørup, “Nonnegative matrix factor 2-d deconvolution for blind single channel source separation,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science (LNCS). Springer, Apr 2006, vol. 3889, pp. 700–707.

[15] O. Winther and K. B. Petersen, “Bayesian independent component analysis: Variational methods and non-negative decompositions,” *Digital Signal Processing*, vol. 17, no. 5, 2007.

[16] T. Hofmann, “Probabilistic latent semantic indexing,” in *Research and Development in Information Retrieval, Proceedings of the International SIGIR Conference on*, 1999.

[17] A. Cichocki, R. Zdunek, and S. ichi Amari, “Csiszar’s divergences for non-negative matrix factorization: Family of new algorithms,” in *Lecture Notes in Computer Science (LNCS)*, vol. 3889. Springer, 2006, pp. 32–39.

[18] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[19] C.-J. Lin, “Projected gradient methods for non-negative matrix factorization,” *Neural Computation*, vol. 19, pp. 2756–2779, 2007.

[20] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, “Algorithms and applications for approximate nonnegative matrix factorization,” *Computational Statistics and Data Analysis*, 2006.

[21] D. Kim, S. Sra, and I. S. Dhillon, “Fast newton-type methods for the least squares nonnegative matrix approximation problem,” in *Data Mining, Proceedings of SIAM Conference on*, 2007.

[22] J. Eggert and E. Körner, “Sparse coding and NMF,” in *Neural Networks, IEEE International Conference on*, vol. 4, 2004, pp. 2529–2533.

[23] M. N. Schmidt. (2008) Non-negative matrix factorization with gaussian process priors. [Online]. Available: http://www.mikkelschmidt.dk/cin2008

[24] M. F. Ochs, R. S. Stoyanova, F. Arias-Mendoza, and T. R. Brown, “A new method for spectral decomposition using a bilinear bayesian approach,” *Journal of Magnetic Resonance*, vol. 137, pp. 161–176, 1999.

Publication C

Structured Non-negative Matrix Factorization with Sparsity Patterns

Hans Laurberg, Mikkel N. Schmidt, Mads Græsbøll Christensen, and Søren Holdt Jensen

Submitted to Proc. *Asilomar*, 2008.

Abstract

In this paper, we propose a novel algorithm for monaural blind source separation based on non-negative matrix factorization (NMF). A shortcoming of most source separation methods is the need for training data for each individual source. The algorithm proposed in this paper is able separate sources even when there is no training data for the individual sources. The algorithm makes use of models trained on mixed signals and uses training data where more than one source is active at the time. This makes the algorithm applicable to situations where recordings of the individual sources are unavailable. The key idea is to construct a structure matrix that indicates where each source is active, and we prove that this structure matrix, combined with a uniqueness assumption, is sufficient to ensure that results are equivalent to training on isolated sources. Our theoretical findings is backed up by simulations on music data that show that the proposed algorithm trained on mixed recordings performs as well as existing NMF source separation methods trained on solo recordings.

1 Introduction

Separation of a single source in a monaural recording, such as a single instrument in polyphonic music or the cocktail party problem [1] is a difficult task. An unsupervised approach is to decompose the signal into basic "atoms", and then group these to form auditory objects—se e.g. [2–6]. Another unsupervised approach is to form flexible source models, and fit these to the mixture—se e.g. [7–9]. A supervised approach is to learn source models from isolated recordings of each source, and use these to separate the mixture subsequently. These source models can be based on, e.g., neural networks [10, 11], factorial hidden Markov models [12, 13], vector quantization [14, 15], independent component analysis [16, 17], or non-negative matrix factorization [1, 18].

When a reasonable amount of training data with isolated sources is available, supervised, model based methods generally yield very good results; however, there are many applications where suitable training data cannot be obtained—for example in instrument separation where many instruments and singers never occur alone. Thus, to use model based methods to separate sources, it is desirable to learn source models directly from the available mixture.

In this paper, we propose a method for learning models of individual sources directly from mixture, in a single-channel source separation framework [18] based on non-negative matrix factorization (NMF). We show that, under certain conditions, training on mixtures works equally well as training on isolated sources. There has been proposed algorithms to learn source models directly from mixtures, by locating areas in the training data, where only one source is active [19]. Our approach does not require this; however, we do require areas, in which each source is *inactive*. The proposed algorithm is successfully tested on music data.

The paper is structured as follows. In Section 2, we introduce NMF and discuss its computation. Next, in Section 3, we describe a general framework for single-channel source separation based on NMF. Our proposed method for learning source models directly from mixed recordings is described in Section 4 and experimentally evaluated on music recordings in Section 5. Finally, we conclude with our conclusions in Section 6 and a detailed description of the simulations in Appendix 1.

2 Non-negative Matrix Factorization

Non-negative matrix factorization[1] (NMF) is the approximate factorization of a non-negative matrix, $\mathbf{V} \in \mathbb{R}_+^{n\times m}$, into the product of two non-negative matrices, $\mathbf{W} \in \mathbb{R}_+^{n\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times m}$

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

In [20] a simple iterative NMF algorithm has been proposed, that minimizes

$$E(\mathbf{W},\mathbf{H}) = \|\mathbf{V} - \mathbf{WH}\|_F^2 \,, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Further, they have proven [21], that each iteration reduces the objective function. In addition to the Frobenius norm, numerous NMF cost functions have been suggested [22, 23], and many different algorithms for computing the NMF have been proposed—for an overview, see [24]. Much effort has been put into finding solutions that are sparse, starting with the sparse NMF method proposed by [25][2]. Many papers from different areas report that sparse NMF algorithms outperform traditional NMF algorithms, which indicates that data in those papers are sparse—se e.g. [1, 25–28]. We believe that there are two reasons for the success of sparse NMF. Firstly, the NMF research has started in areas where it is known that there are understandable underlying data (which often means sparse underlying data). Secondly, if the underlying data is not sparse at all (no elements are close to zero) the NMF is not unique [29].

In this paper, we will use the sparse NMF formulation of [27] that is based on the following cost function

$$C(\mathbf{W},\mathbf{H}) = \frac{1}{2}\left\|V - \overline{\mathbf{W}}\mathbf{H}\right\|_F^2 + \lambda \sum_{i,j} \mathbf{H}_{i,j} \tag{3}$$

$$\overline{\mathbf{W}}_n = \frac{\mathbf{W}_n}{\|\mathbf{W}_n\|} \quad, n \in \{1, \ldots, r\}, \tag{4}$$

where $\mathbf{W}_n$ is the n'th column vector in $\mathbf{W}$, and the parameter λ controls the trade-off between sparsity of $\mathbf{H}$ and approximation error, $E(\mathbf{W},\mathbf{H})$.

[1] In some literature NMF is also called non-negative matrix approximation and positive matrix factorization.

[2] In the work of Hoyer, the method is called non-negative sparse coding.

Algorithm 1 NMF source separation

```
1: For each source, $n$, compute NMF of isolated training data,
       $\mathbf{V}'_n \approx \mathbf{W}'_n \mathbf{H}'_n.$
   Store $\mathbf{W}'_n$ and discard $\mathbf{H}'_n$.
2: Compute $\mathbf{H}_1, \cdots, \mathbf{H}_N$
       $\mathbf{V} \approx \sum_{n=1}^{N} \mathbf{V}_n = \mathbf{W}'\mathbf{H} = [\, \mathbf{W}'_1 \cdots \mathbf{W}'_N \,] \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_N \end{bmatrix}$
3: Return $\mathbf{V}_n = \mathbf{W}'_n \mathbf{H}_n$ as an estimate of the $n$'th source.
```

3 Source separation using NMF

A supervised approach [18] to source separation is described in Algorithm 1. In the first step of the algorithm, training data, consisting of isolated recordings of each source, are used to build a model of each source. Step 1 in the algorithm has only to be calculated once for each source, and the computational complexity of this step is thus not crucial. For the cost function in Equation 3, Step 2 in the algorithm can be computed efficiently using quadratic programming. To ensure that the $\mathbf{W}'$ in Algorithm 1 Step 1 can be used for separation, it is desirable that the estimated $\mathbf{W}'$ is unique up to a permutation and a scaling—for further analysis of uniqueness of NMF see [29]. In [29] a NMF is called unique if all factorisations are on the form

$$\mathbf{V} = \underbrace{\mathbf{W}'}_{=\mathbf{W}\mathbf{D}^{-1}\mathbf{P}^{-1}} \underbrace{\mathbf{H}'}_{=\mathbf{P}\mathbf{D}\mathbf{H}}, \qquad (5)$$

where $\mathbf{P}$ and $\mathbf{D}$ is a permutation and a scaling, respectively. So using this terminology, Algorithm 1 will produce reproducible results if all $\mathbf{V}'_n$ are unique.

4 Learning source models from mixed sources

To explain the new algorithm, we start by reformulating the first step in Algorithm 1. If all training data are gathered in one matrix, say $\mathbf{V}' = [\, \mathbf{V}'_1 \cdots \mathbf{V}'_N \,]$, Step 1 can be computed for all instruments by solving

$$\mathbf{V}' \approx \mathbf{W}'\mathbf{H}' = [\, \mathbf{W}'_1 \cdots \mathbf{W}'_N \,] \begin{bmatrix} \mathbf{H}'_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{H}'_N \end{bmatrix}. \qquad (6)$$

Implementing step one of Algorithm 1 in this manner is computationally inefficient, but it makes it clear, that prior knowledge of zeros in $\mathbf{H}'$ makes it possible to find $\mathbf{W}'_n$ for

Algorithm 2 Structured NMF source separation

```
Gather all training data in a data matrix
    $$\mathbf{V}_{train} = [\mathbf{V}_1 \cdots \mathbf{V}_M].$$
    Let $\mathbf{H}'$ be a structured matrix, and solve
    $$\mathbf{V}_{train} \approx [\mathbf{W}'_1 \cdots, \mathbf{W}'_N]\mathbf{H}',$$
    maintaining the structure in $\mathbf{H}$. Store $[\mathbf{W}'_1 \cdots \mathbf{W}'_N]$ and discard $\mathbf{H}'$.
Compute $\mathbf{H}_1, \cdots, \mathbf{H}_N$
    $$\mathbf{V} \approx \sum_{n=1}^{N} \mathbf{V}_n = \mathbf{W}'\mathbf{H} = \left[\, \mathbf{W}'_1 \cdots \mathbf{W}'_N \,\right] \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_N \end{bmatrix}$$
Return $\mathbf{V}_n = \mathbf{W}'_n\mathbf{H}_n$ as an estimate of the $n$'th source.
```

each source. In the following, we call a matrix $\mathbf{H}'$ with zeros in patterns a *structured* $\mathbf{H}'$ matrix, and we refer to NMF, with *structured* $\mathbf{H}$, as *structured NMF*. The following theorem shows, that most matrices $\mathbf{H}$ with structure can be used to indentify the model for each source.

Theorem C.1 *Let*

$$\mathbf{V} = \left[\, \mathbf{V}_1 \cdots \mathbf{V}_N \,\right] = \mathbf{W}\mathbf{H} = \left[\, \mathbf{W}_1 \cdots \mathbf{W}_N \,\right] \begin{bmatrix} \mathbf{H}_1^1 & & \mathbf{H}_N^1 \\ & \ddots & \\ \mathbf{H}_1^N & & \mathbf{H}_N^N \end{bmatrix}$$

be a unique NMF, where $\mathbf{H}_n^n = \mathbf{0}$ for all n, and let

$$\mathbf{V} = \hat{\mathbf{W}}\hat{\mathbf{H}} = \left[\, \hat{\mathbf{W}}_1 \cdots \hat{\mathbf{W}}_N \,\right] \begin{bmatrix} \hat{\mathbf{H}}_1^1 & & \hat{\mathbf{H}}_N^1 \\ & \ddots & \\ \hat{\mathbf{H}}_1^N & & \hat{\mathbf{H}}_N^N \end{bmatrix}$$

be any NMF of $\mathbf{V}$, where $\hat{\mathbf{H}}_n^n = \mathbf{0}$ for all n. If there are no $n \neq m$ such that $\mathbf{H}_n^m$ has a row of zeros then

(a) $\mathbf{W}_n\mathbf{H}_m^n = \hat{\mathbf{W}}_n\hat{\mathbf{H}}_m^n$*, for all n and m.*

(b) For all n, there is a permutation matrix, $\mathbf{P}_n$, and a diagonal scaling matrix, $\mathbf{D}_n$, such that $\hat{\mathbf{W}}_n = \mathbf{W}_n\mathbf{P}_n\mathbf{D}_n$.

Proof. The NMF of $\mathbf{WH}$ is unique, and therefore $\hat{\mathbf{W}} = \mathbf{WD}^{-1}\mathbf{P}^{-1}$ and $\hat{\mathbf{H}} = \mathbf{PDH}$. The proof is concluded by realizing that the permutation $\mathbf{P}$ must be block diagonal,

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{P}_N \end{bmatrix}, \tag{7}$$

in order for $\hat{\mathbf{H}}_n^n = \mathbf{0}$ for all n and therefore

$$\mathbf{PD} = \begin{bmatrix} \mathbf{P}_1\mathbf{D}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{P}_N\mathbf{D}_N \end{bmatrix}. \tag{8}$$

■

In terms of modelling sources, the theorem states, that if one wants to estimate N source models, and has N training recordings, each with one source missing, then there is a unique solution, if all source components are active in all training files where it is not missing. Theorem C.1 leads naturally to Algorithm 2. The training data used in step 1 of Algorithm 2 does not need to be isolated recordings of each and, and Theorem C.1 shows, that if the assumptions are fulfilled, the result is the same as for Algorithm 1. Note that step 2 and 3 of Algorithm 2 is the same as in Algorithm 1.

5 Results

We have constructed three tests, in which we compare Algorithm 1 and 2. Throughout the test, Algorithm 1 always has solo recordings for the training where as Algorithm 2 always use mixture recordings. The first is a simulation shows that both algorithms can separate three and four artificial sources. The second test is an example of instrument tone separation in a single channel recording of musical notes and the third test is an example of instrument separation in a single channel recording of mixed polyphonic music.

In the first test, artificial sources are separated. The data, V', is a square matrix, and each source has $\{2, 4, 6\}$ components. In Figure 1, the estimation error is shown for Algorithm 1 (trained on individual training data for each source) and Algorithm 2 (trained on mixed training data). For a detailed description of the experiment, see Appendix 1. From the simulation it can be seen, that when the amount of data is sufficient, the two algorithms perform equally well. In the second test, Algorithm 2 is used on amplitude spectra of three instruments form the Iowa Music Database [30]. Each $\mathbf{V}_n'$ consist of two instruments that both plays one note. In this test the averaged cosine of the angles between the basis vectors calculated using Algorithm 2 and the basis vectors calculated using solo recordings above 0.99, which in practise means that they are equal. Figure 2 shows an example of $\mathbf{V}_{train} = [\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3]$ and there corresponding three basis spectres are shown in figure 3. It can be seen that the basis vectors are estimated almost

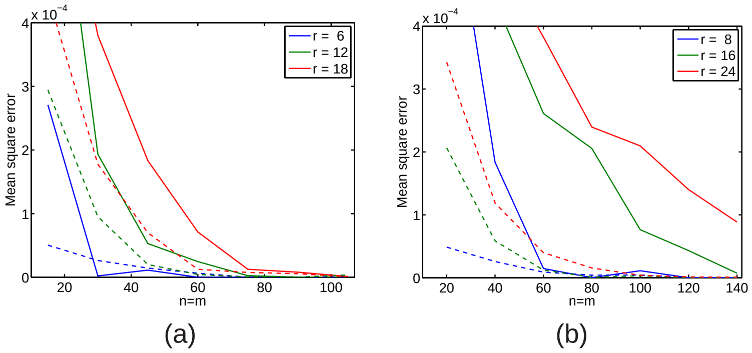


Figure 1: The mean error of separation of (a) three sources and (b) four sources, using Algorithm 1 (dashed lines) and Algorithm 2 (solid lines). The simulation is computed with different model orders r and size of training data $m = n$.

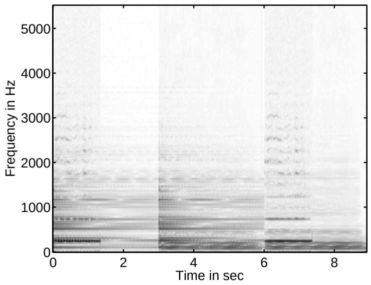


Figure 2: The amplitude spectrogram $\mathbf{V}$ of a bas, a flute and a piano that plays together two and two.

correct even though the spectres are heavily overlapping. It can also be seen that the small errors occur at a frequency in a basis vector when the there are a lot of energy in both the other basis vectors at that frequency. A reason for this is that the NMF problem might not be unique and the non-uniqueness is that it is possible to raise the energy of one basis vector by decreasing the other basis vectors when the tones starts and stops at the same time.

In the third test, Algorithm 2 is used on amplitude spectra of midi music. The instrument models were trained on three 10-second training files, each with two instruments playing. These models were used to separate the three instruments from a 10-second evaluation file, as shown in Figure 4. In this test, the mixing of the instruments is performed in the time domain, which makes the amplitude spectra non-additive, due to phase differences, when there is overlap between the spectra. In this simple experiment, it is possible to separate the three instruments with minor artefacts. In the estimated piano the artefacts do not sound like an instrument but in the estimated drum signal, it is

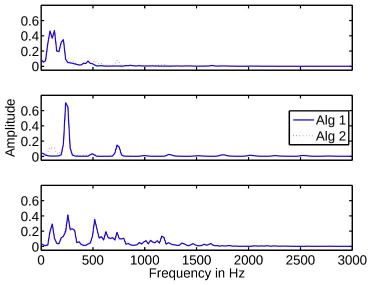


Figure 3: The estimated basis spectras of the bas (top), flute (middle) and piano (bottom) from Figur 2.

possible to hear the bas in the background and in the estimated bas signal there is the piano in the background. It is possible to download the sound files from our website (http://kom.aau.dk/ hla/structuredNMF).

6 Conclusion

An algorithm for source separation based on training source models on mixed audio recordings was presented. In contrast to existing algorithms, the proposed algorithm uses training data where more than one source is active, which makes the algorithm applicable to situations, where individual recordings of sources are unavailable.

The proposed algorithm is based on the non-negative matrix factorization (NMF), and can be used with most NMF algorithms. The novel idea in this paper is to construct a structure matrix, that indicates where each source is active, and the proof that this structure matrix, together with a uniqueness assumption, is enough to ensure, that results are equivalent to training on isolated sources. The theoretical results are backed up by simulations that show that the proposed algorithm performs as well as existing NMF source separation methods, when sufficient training data is available.

1 Simulation details

In the first test $\mathbf{W}$, $\mathbf{H}_{train}$ and $\mathbf{H}_{test}$ are generated as uniform IID values raised to the power of 8. All NMF calculation in this simulation use the sparse NMF algorithm [27] with $\lambda = 0.001$, 200 iterations and 20 different starting points. The error plotted in Figure 1 is a Monte Carlo simulation of the mean square error of between the test sources and the estimated test sources. There are used 20 Monte Carlo runs in the simulation. In order to make the plot more dense, the error is divided by the number of basis vectors r to compensate for different amplitudes of the matrices.

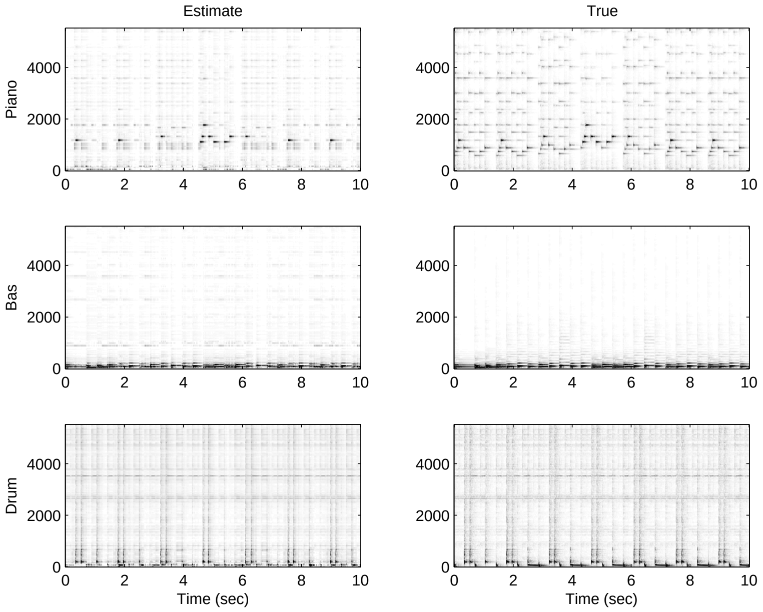


Figure 4: The figure shows the spectrograms of the separation of a MIDI music piece with a piano (top), a bas (middle) and a drum (bottom). The left column shows the estimate and the right column shows the spectrogram of each instruments.

In the second test, notes with the length of one to three seconds were used and the data was downsampled to 11.025 kHz. In the third test was the sampling frequency of the sound files is 44.1 kHz. The algorithm setup for both music tests is the sparse NMF with $\lambda = 0.1$, 500 iterations, one starting point and the amplitude spectrogram of the music is calculated using a (46.4 ms) Hanning window and 50% window overlap. To estimate the instrument time signal the phase of the mixed spectrogram is used directly.

References

[1] Mikkel N. Schmidt and Rasmus K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2006.

[2] Tuomas Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, Mar 2007.

[3] Francis R. Bach and Michael I. Jordan, “Blind one-microphone speech separation: A spectral learning approach,” in *Neural Information Processing Systems, Advances in (NIPS)*, 2005, pp. 65–72.

[4] Guoning Hu and DeLiang Wang, “On amplitude modulation for monaural speech segregation,” in *Neural Networks (IJCNN), Proceeding of the International Joint Conference on*, 2002, pp. 69–74.

[5] Tero Tolonen, “Methods for separation of harmonic sound sources using sinusoidal modeling,” in *Audio Engineering Society (AES) Convention*, May 1999.

[6] T. Virtanen and A. Klapuri, “Separation of harmonic sound sources using sinusoidal modeling,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, 2000, vol. 2, p. 765.

[7] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné, and Shigeki Sagayama, “Single channel speech and background segregation through harmonic-temporal clustering,” in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on (WASPAA)*, Oct 2007, pp. 279–282.

[8] Ron J. Weiss and Daniel P. W. Ellis, “Monaural speech separation using source-adapted models,” in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on (WASPAA)*, 2007, pp. 114–117.

[9] Mikkel N. Schmidt and Morten Mørup, “Nonnegative matrix factor 2-d deconvolution for blind single channel source separation,” in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, vol. 3889 of *Lecture Notes in Computer Science (LNCS)*, pp. 700–707. Springer, Apr 2006.

[10] J. Klingseisen and M. Plumbley, “Experiments on musical instrument separation using multiple cause models,” in *In Cambridge Music Processing Colloquium, Cambridge, England, 30 Sept*, 1999.

[11] J. Klingseisen and M. Plumbley, “Towards musical instrument separation using multiple-cause neural networks,” in *Proceedings of the International Workshop on Independent Component Analysis And Blind Signal Separation*, June 2000.

[12] Sam T. Roweis, “One microphone source separation,” in *Neural Information Processing Systems, Advances in (NIPS)*, 2000, pp. 793–799.

[13] Trausti Kristjansson, John Hershey, Peder Olsen, Steven Rennie, and Ramesh Gopinath, “Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system,” in *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2006, pp. 97–100.

[14] Sam T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Speech Communication and Technology, European Conference on (EUROSPEECH)*, 2003, pp. 1009–12.

[15] Daniel P. W. Ellis and Ron J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, May 2006, pp. 957–960.

[16] Gil-Jin Jang and Te-Won Lee, "A probabilistic approach to single channel source separation," in *Neural Information Processing Systems, Advances in (NIPS)*, 2003.

[17] Gil-Jin Jang and Te-Won Lee, "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, Dec 2003.

[18] Beiming Wang and Mark D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *ICA Research Network International Workshop, Proceedings of*, Sep 2006, pp. 17–20.

[19] Minje Kim and Seungjin Choi, "On spectral basis selection for single channel polyphonic music separation," in *Artificial Neural Networks, International Conference on (ICANN)*. Sep 2005, vol. 3697 of *Lecture Notes in Computer Science (LNCS)*, pp. 157–162, Springer.

[20] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[21] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2000, pp. 556–562.

[22] Raul Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2007.

[23] Andrzej Cichocki, Rafal Zdunek, and Shun ichi Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Lecture Notes in Computer Science (LNCS)*. 2006, vol. 3889, pp. 32–39, Springer.

[24] Michael W. Berry, Murray Browne, Amy N. Langville, Paul V. Pauca, and Robert J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics and Data Analysis*, vol. 52, pp. 155–173, September 2007.

[25] P. O. Hoyer, “Non-negative sparse coding,” in *IEEE Proc. Neural Networks for Signal Processing*, 2002, pp. 557–565.

[26] Bruno A. Olshausen and David J. Field, “Sparse coding of sensory inputs,” *Current Opinion in Neurobiology*, vol. 14, no. 4, pp. 481–487, Aug 2004.

[27] J. Eggert and E. Körner, “Sparse coding and NMF,” in *Proc. Int. Joint Conf. on Neural Networks*, 2004, pp. 2529–2533.

[28] Arshia Cont, “Realtime multiple pitch observation using sparse non-negative constraints,” in *Music Information Retrieval, International Conference on (ISMIR)*, October 2006.

[29] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, “Theorems on positive data: On the uniqueness of nmf,” *Computational Intelligence and Neuroscience*, 2008.

[30] “The university of iowa musical instrument samples database,” [Online]. Available: http://theremin.music.uiowa.edu.

Publication D

On Affine Non-negative Matrix Factorization

Hans Laurberg and Lars Kai Hansen

The paper has been published in
Proceedings of ICASSP, 2007.

The layout has been revised.

Abstract

We generalize the non-negative matrix factorization (NMF) generative model to incorporate an explicit offset. Multiplicative estimation algorithms are provided for the resulting sparse affine NMF model. We show that the affine model has improved uniqueness properties and leads to more accurate identification of mixing and sources.

Keywords: Non-negative matrix factorization, NMF, BSS, Sparse NMF

1 Introduction

Non-negative matrix factorization (NMF) has become a popular tool for data analysis. An often stated reason for NMF is that it leads to 'parts based' representations, hence, facilitates data analytic interpretation. However, uniqueness is important for the parts based representations to be meaningful. The NMF generative model is based on linear mixing of positive sources by positive coefficients. The positive sources may have offsets which can lead to non-uniqueness, we therefore here propose a model based on *affine mixing*, i.e., mixing with an offset. The NMF learning algorithm is straightforwardly generalized to handle the augmented model. We show that the affine model indeed has improved uniqueness properties and thus leads to more accurate identification of mixing and sources.

NMF algorithms are used to factorize a nonnegative matrix $V \in \mathbb{R}^{N\times M}$ in two nonnegative matrices $W \in \mathbb{R}^{N\times D}$ and $H \in \mathbb{R}^{D\times M}$

$$V \approx R = WH; \;\; V_{i,j} \approx R_{i,j} = \sum_{d=1}^{D} W_{i,d} H_{d,j} \tag{1}$$

Following the seminal papers by Lee and Seung [1, 2], a least squares or a Kullback-Leibler inspired cost are used. Our observations in this paper can be applied to both. For simplicity we will concentrate on the Euclidian cost in the following,

$$E(W,H) = \|V - WH\|_F^2 \,, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm. Lee and Seung [2] have shown that the following update rule will decrease $E(W,H)$

$$H \leftarrow H \otimes \frac{W^T V}{W^T R} \tag{3}$$

$$W \leftarrow W \otimes \frac{V H^T}{R H^T} \;\;, \tag{4}$$

where $\otimes$ and $\frac{(\cdot)}{(\cdot)}$ are element wise multiplication and division. This update rule is used as a reference and is shown in panel (B) of figures 1, 3, 4, 5 and 6.

2 Sparse NMF

Hoyer [3] introduced sparse NMF and Eggert [4] proposed the following cost function where only the normalized version of W has impact on the cost

$$E(W,H) = \frac{1}{2}\left\|V - \overline{W}H\right\|_F^2 + \lambda \mathbf{1}^T H \mathbf{1} \tag{5}$$

$$\overline{W}_n = \frac{W_n}{\|W_n\|} \quad , n \in \{1,\ldots,N\} \tag{6}$$

where W_n is the n'th column vectorin W and $\mathbf{1}$ is a column vector where all elements are one. The length of $\mathbf{1}$ can be deduced by the context. The scalar λ is a positive parameter that controls the tradeoff between sparseness of H and approximation of V by the product of W, H. Eggert [4] argues for using the following multiplicative update

$$H \leftarrow H \otimes \frac{\overline{W}^T V}{\overline{W}^T R + \lambda} \tag{7}$$

$$W_n \leftarrow \overline{W}_n \otimes \frac{\sum_{m=1}^{M} H_{m,n}(V_n + \overline{W}_n (R_m)^T \overline{W}_n)}{\sum_{m=1}^{M} H_{m,n}(R_n + \overline{W}_n (V_m)^T \overline{W}_n)} \tag{8}$$

These update rules are used in panel (C) of figures 1, 3, 4, 5 and 6.

The normalization of W and the sparse nature of H critically constrains the solution and can improve uniqueness and lead to more accurate estimates. However, the constraints may not be consistent with the form of the mixing process and the statistics of the source signals H. In particular offsets in one or more rows of V will counteract the sparse model. If the generative model incorporates additive noise it is not clear that simple subtraction of the minimal value of each row in V will lead to a correct recovery of the generating W, H. If the noise is, e.g., Gaussian, V can be negative in the native representation, hence, one cannot estimate the 'true' offset.

2.1 Affine Sparse NMF

The above sparse NMF methods do not handle offsets, however, it is incorporated as follows with $W_0 \in \mathbb{R}^{N\times 1}$

$$V \approx R = WH + W_0 \mathbf{1}^T. \tag{9}$$

Using this augmented signal model the sparse cost function in Equation 5 becomes

$$E(W,H,W_0) = \frac{1}{2}\left\|V - \overline{W}H - W_0\mathbf{1}^T\right\|_F^2 + \lambda \mathbf{1}^T H \mathbf{1} \tag{10}$$

Following Eggert [4] the update rule for W and H remains as given in Equation 7 and 8 using the new definition of R and the update rule for W_0 (that in not normalized) is

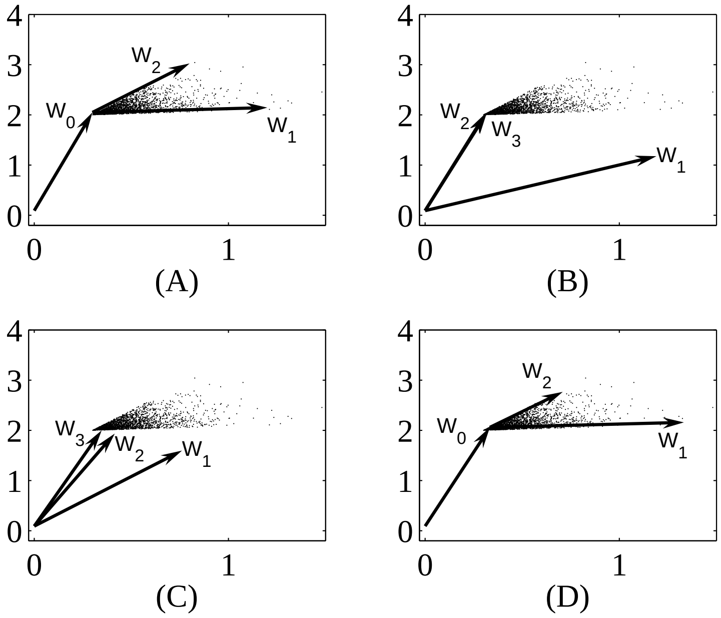


Figure 1: Simulated data where $V \in \mathbb{R}^{2\times 2000}$ is generated according to Equation 9. Each column of V is plotted as a dot. In (A) the generating W and W_0 are shown. In (B) and (C) the standard NMF and sparse NMF each find three vectors that can describe the data. Both algorithms find one vector that is a linear combination of the true W_0 and W_1 and finds two vectors that are very close to the true W_0. In (D) the 'Affine sparse NMF' method correctly estimates the structure of the W matrix.

the standard NMF update rule in Equation 4

$$W_0 \leftarrow W_0 \otimes \frac{\mathbf{1}^T V}{\mathbf{1}^T R} \tag{11}$$

The affine sparse NMF results are shown in panel (D) of figures 1, 3, 4, 5 and 6.

3 Results

How does the augmented sparse affine NMF model data? To answer this question we first visualize synthetic data as generated by the proposed model, and we show existing methods fail to reconstruct the correct parameters of the generative model. We then go on to show that two commonly used data sets have the characteristics of the proposed model and that the proposed algorithm performs better than the existing algorithms on the data. In order to get a 'fair' comparison the standard NMF and sparse NMF

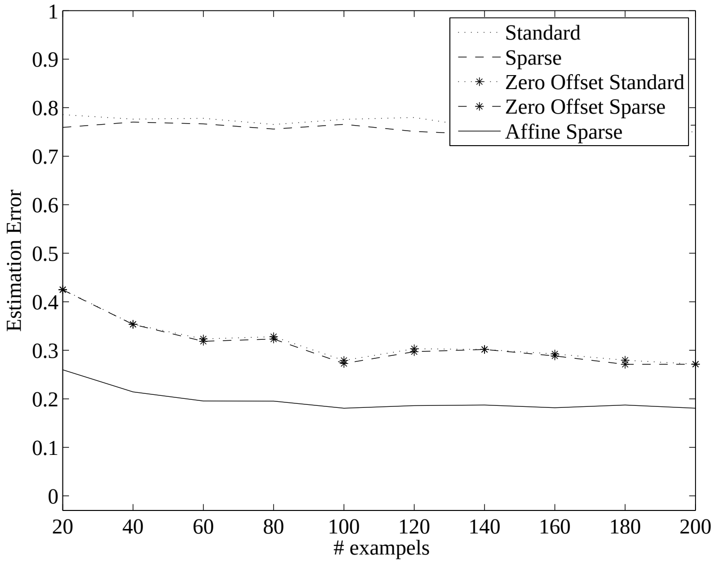


Figure 2: The variation of the relative least squares error of the NMF reconstruction of W. The error is plotted as a function of the amount of data (M). The simulated data was generated using $D = 10$ components and an off set. The 'zero offset' methods are based on the simple heuristic that data is first preprocessed to have minimum value zero in each row.

both have one column more than the sparse affine NMF method. This ensures that the maximum rank of R is the same for all methods.

Simulated Data. In Figure 1 there are $M = 2000$ elements in V. The data is generated as in Equation 9. The elements of R are exponentially distributed. The true W vectors and the column vectors of V are shown in Figure 1 panel (A). Figure 1 (B–D) shows the three different algorithms estimate of W. The standard NMF (B) finds W such that the data is in the positive span of W. The W estimated by the sparse NMF algorithm (C) also spans data but the column vectors of W point more directly towards data. Although these methods estimated W can reproduce V, they do not find the correct structure (W). The proposed method (D) finds a W that is close to the true W.

A quantitative evaluation of the different algorithms' estimate is presented in Figure 2. Data is generated as in Equation 9 where the elements of W and W_0 are uniform i.i.d. The elements of H are first generated as exponential i.i.d. samples and then each column is normalized to unit sum. In this way the elements in H describe how much each column vector of W contribute towards V. In all simulations $N = 100, D = 10$. We have run the simulation with different amounts of data examples (column in V) M. In the evaluation V is analysed as 11(=D +1) outer product $\sum_{d=0}^{D} V^{(d)} = V$, where

$V_{i,j}^{(d)} = W_{i,d} H_{d,j}$. The error in the figure is the relative least squares error of the $V^{(d)}$ estimate for each data set size

$$\frac{\sum_{d=0}^{D} \left\| V^{(d)} - R^{(d)} \right\|_F^2}{\sum_{d=0}^{D} \left\| V^{(d)} \right\|_F^2} \tag{12}$$

For completeness we have included in the performance evaluation a modification of the standard method in which data is first subtracted with constant offsets to achieve zero minimum value in each of the N variables of V. The simulation shows that the standard NMF and the sparse NMFs do not find the true W and H. The constant offset subtraction improves the performans but is outperformed by the sparse affine NMF succeeds. Notis that the two latter methods is favoured by knowing that $(H^T)_0 = \mathbf{1}$.

The Swimmer Database. The “Swimmer Database” was introduced by Donoho and Stodden [5] to discuss the uniqueness issues we have adressed in this presentation. The point was that even if NMF can represent V it may not necessarily find the right W. The database consist of 256 (32×32 pixel) black-and-white pictures of a ‘stick-man’ with 4 limbs that can be in one of 4 positions. All pictures have a ‘torso’ that represent an offset as discussed in this paper. The pictures in the dataset can be constructed by 17 ($= 4 \times 4 + 1$) non-overlapping basis pictures. In Figure 3 (A) examples from the database are shown. The algorithms described in section 2 are tested on the data set and a subset of the 17 basis pictures are shown in Figure 3(B–D). Only the proposed method is able to find the 17 non-overlapping basis pictures, the standard NMF and Sparse NMF all let the torso be a part of all basis pictures. The Swimmer simulation is further analyzed in Figure 4. The 1024 ($= 32 \times 32$) dimensional column vectors in V and W are mapped onto a two dimensional subspace to show that the structure of the swimmer database is in fact equivalent to that of Figure 1. In the plot it is seen that only the affine sparse NMF finds the true basis vectors.

Business Card Data Set. Our final example is based on a set of business card images of faculty of Aalborg University’s Department of Electronic Systems. The photographer has manually centered and scaled the pictures. The pictures are scaled to 30×40 pixel and the color map is chosen such that white is zero and black is maximum. An ‘AAU watermark’ logo has been added to all pictures in the database. A subset of the pictures are shown in Figure 5(A) and a subset of the 25 basis pictures estimated by the three algorithms is shown in Figure 5(B–D). In this simulation the sparse affine NMF algorithm estimates more sparse basis pictures and most basis pictures describe one physical object only.

A two dimensional subspace (axes formed by a picture with ‘hair’ and an picture with the AAU-logo) of the images in Figure 5 are shown in Figure 6. As above we find that none of the standard NMF’s nor sparse NMF basis vectors describe the AAU logo without also capturing ‘hair’. The basis pictures for the proposed method however are found close to the axes meaning that they either capture hair or the AAU’ logo.

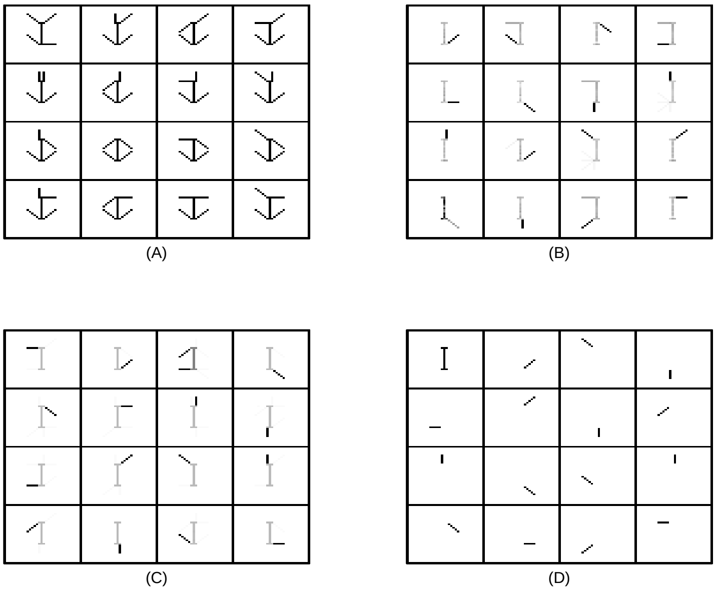


Figure 3: Subset of A: The Swimmer database B: Basis pictures using standard NMF. C: Basis pictures using sparse NMF. D: Basis pictures using sparse affine NMF.

4 Discussion and conclusion

Non-negative matrix factorization is widely applied because of the ability to create ‘parts based’ representations, hence, facilitating model interpretation. However, uniqueness is important for the parts based representations to be meaningful. Lack of uniqueness can happen in several ways, e.g., due to an offset vector W_0 as discussed here. Another mechanism resulting in lack of uniqueness is if the support of the process creating a row of H does not include $H = 0$, i.e., if there is an offset in the row variable of H. The H_0 offset can be seen as a W_0 offset with the constraint that W_0 is in the positive span of the column vectors in W

$$R = W(H + H_0\mathbf{1}^T) = WH + W_0\mathbf{1}^T, \;\; W_0 = WH_0 \qquad (13)$$

Hence, the H offset issue is a special case of the model we have discussed here: If the resulting W_0 is in the positive span of the columns of W, they can be interpreted as H offsets.

In this work we have defined the augmented non-negative linear mixing model - the sparse affine NMF. We have presented three case stories in which the new sparse affine NMF algorithm outperforms the standard algorithms and a naive solution in estimation of the underlying structure of the data.

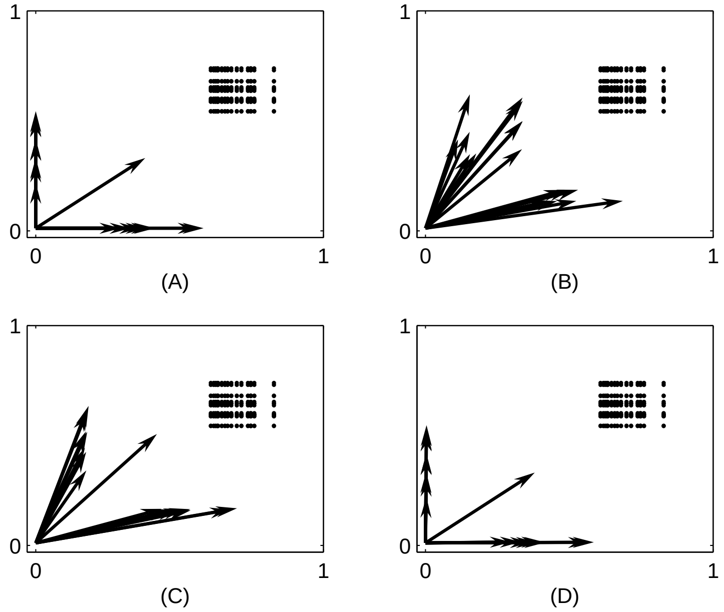


Figure 4: A two dimensional subspace of the column vectors in V (dots) and W (vectors) are shown for the Swimmer database. The 'x-axis' is a picture which is zero in the upper part and uniform random values in the lower part. The 'y-axis' is constructed the same way but with the zeros in the lower part.

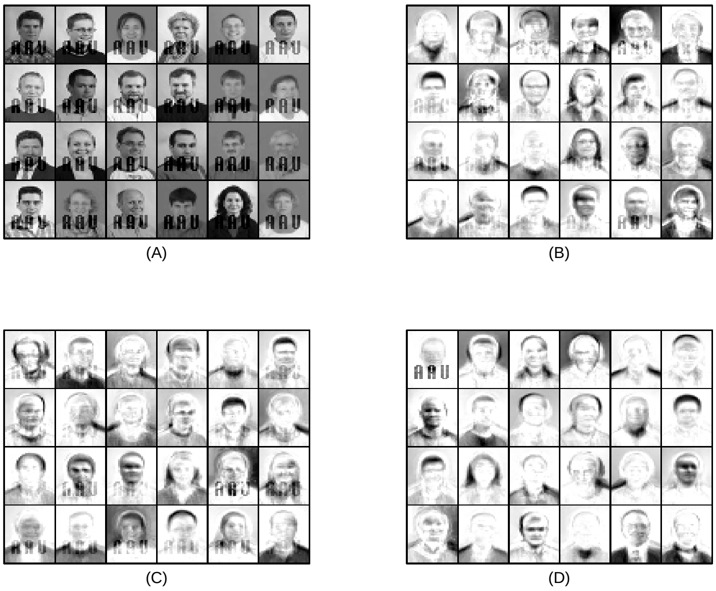


Figure 5: (A): Subset of the Picture database with 197 pictures (B – D): A subset of the basis pictures using standard NMF, sparse NMF and sparse affine NMF. The standard NMF makes very noisy basis pictures. The sparse NMF produce basis pictures where the 'AAU watermark' is visible in around 50% of the pictures, and in addition a lot of the pictures do not represent a single part of the picture. The sparse affine NMF has only one picture with the watermark (W_0) and most pictures represent only one part of the picture.

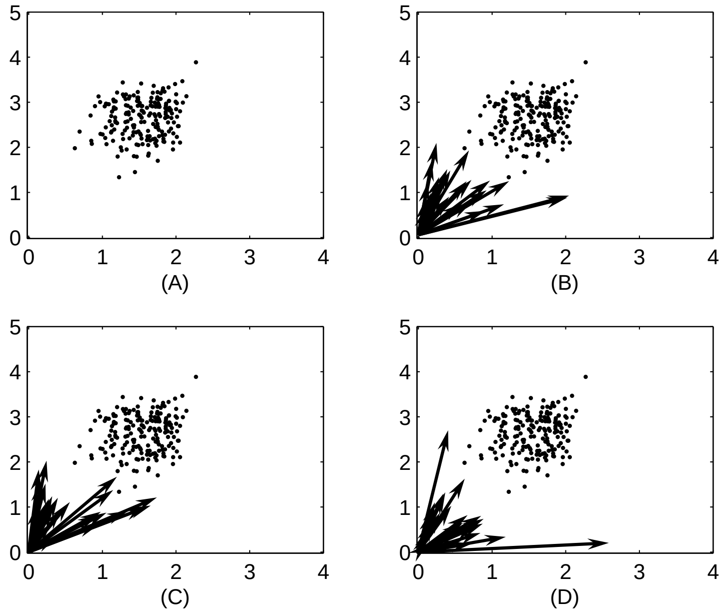


Figure 6: The business card images plotted in two dimensions to show that data and solutions have pattern like the ones in Figure 1. The x–axis is the an image of the AAU logo, and the y–axis is an image vector capturing the 'hair' region.

References

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization.” *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[2] D. D. Lee and S. H. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2000, pp. 556–562.

[3] P. O. Hoyer, “Non-negative sparse coding,” *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pp. 557–565, 2002.

[4] J. Eggert and E. Körner, “Sparse coding and NMF,” in *2004 IEEE International Joint Conference on Neural Networks*, 2004, pp. 2529–2533.

[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 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

Publication E

Uniqueness of Non-negative Matrix Factorization

Hans Laurberg

The paper has been published in
Proceedings of SSP, 2007.

Abstract

In this paper, two new properties of stochastic vectors are introduced and a strong uniqueness theorem on non-negative matrix factorizations (NMF) is introduced. It is described how the theorem can be applied to two of the common application areas of NMF, namely music analysis and probabilistic latent semantic analysis. Additionally, the theorem can be used for selecting the model order and the sparsity parameter in sparse NMFs.

Keywords: Non-negative matrix factorization (NMF), sparse NMF, non-negativity, model selection.

1 Introduction

Large quantities of positive data occur in different research areas such as music analysis, text analysis, image analysis and probability theory. Before deductive science is applied to large quantities of data it is often appropriate to reduce data by preprocessing, e.g., by rank reduction or by feature extraction. Principal component analysis is an example of such preprocessing. When the original data is non-negative it is often desirable that this property is preserved in the preprocessing. For example, elements in a power spectrogram, probabilities and pixel intensities should still be non-negative after the processing to be meaningful. This has lead to the construction of algorithms for rank reduction of matrices and feature extraction that makes a non-negative output. Many of the algorithms are on the form of non-negative matrix factorization (NMF) proposed by Lee and Seung [1, 2]. The NMF algorithms factorize a non-negative matrix $V \in \mathbb{R}^{n\times m}$ into two non-negative matrices $W \in \mathbb{R}^{n\times r}$ and $H \in \mathbb{R}^{r\times m}$

$$V \approx R = WH; \;\; V_{i,j} \approx R_{i,j} = \sum_{d=1}^{r} W_{i,d} H_{d,j}, \tag{1}$$

where $W_{i,j}$ denotes the i'th element in the j'th column. There are no closed-form solutions to the problem of finding W and H given a V, but Lee and Seung [1, 2] proposed two computationally efficient algorithms for minimizing the difference between V and WH for two different error functions. Later, numerous other algorithms have been proposed (see Berry et al. [3]).

An interesting question is whether there exists only one NMF of a matrix. The importance of this question depends on the particular application of the NMF. There can be two different starting points when using a mathematical model like the NMF – either one can believe that the model describes nature and that the variables have a physical meaning, or one can believe that the model can capture the part of interest, or its behavior, even though there is not a one to one mapping between the parameters, model and the physical system. When using NMF, one can wonder whether $V =$

$WH + G$, where G is a noise source, or whether data is constructed by another model. Or, in other words, does a ground truth W and H exist. These questions are important in evaluating whether or not it is a problem that there are other NMF solutions to the same data, i.e.,

$$V \approx R = WH = W'H'. \tag{2}$$

If NMF is used even though the data is not assumed to be generated by equation 1, it may not be a problem that there are several other solutions. On the other hand, if one assumes that there exist a ground truth, it may be a problem if that model is not detectable, i.e., it is not possible to find W and H from data.

In this paper, we analyze under which circumstances there only exists exactly one NMF of a matrix. In doing this, it is assumed that there exists a true W and H and the conditions on W and H that results in a unique NMF are explored. Here, the elements of W and H are viewed as stochastic variables and it is shown that the factorization is unique under mild conditions. The theorems in this paper deals with the situation where V is constructed as $V = WH$, i.e., the noise free situation. To the best of our knowledge, the only papers that deals with the issue of uniqueness of NMF are the paper by Donoho and Stodden [4] and the paper by Theis et al. [5]. Donoho and Stoddens paper deals with two different situations; one where $W = V$ or $H = V$ and another where $W_{n,d} \neq 0 \Rightarrow \ W_{n,d'} = 0$ for all $d' \neq d$. The paper by Theis et al. deals with the specific situation where the vectors $H_d = [H_{d,1}, \cdots, H_{d,m}]^T$ has known ratios between the L_1 norm and the L_2 norm, $\|H_d\|_1 / \|H_d\|_2$. This paper takes another approach by assuming that W and H are generated from a process and identifies the statistical properties of the process that makes the NMF unique. The work reported here is inspired by Plumbley [6] showing that it is possible to make unique blind source separation (BSS) if the source are non-negative, independent and have zero as the largest lower limit. The theorems in this paper are derived by combining the ideas of Plumbley with the ideas of Donoho and Stodden.

The remaining part of this paper is organized as follows. In Section 2, some basic definitions are introduced and a fundamental property for the detection of negative matrix elements is presented. Then, the main results on the uniqueness of NMFs are presented an analyzed in Section 3. In Section 4, we give some examples of the application of the theorems to real data analysis, where after we conclude on our work in Section 5.

2 Sufficiently Spread Data

Prior knowledge of non-negativity of a data set can be used to reduce a rotation ambiguity to a permutation ambiguity under some circumstances. The only rotation where all matrix elements are non-negative are a permutation, so if it is possible to detect negative values in a matrix, a rotation ambiguity can be restricted to a permutation ambiguity. Plumbley [6] shows that if the stochastic variables in a vector s are independent, the probability for $x = As$ having a negative element will be grater than zero if s is grounded. In other words, if the rotation matrix A is not a permutation, then there is a positive probability for having a negative element in x.

Definition E.1 *A stochastic variable s_j is called grounded if $p(s_j < k) > 0$ for all $k > 0$.*

Definition E.2 *A stochastic vector $s \in \mathbb{R}^n$ is called sufficiently spread if all the elements are non-negative and for all $\epsilon > 0$ and $j \in \{1, \cdots, n\}$ then $p(-\epsilon s_j + \sum_{i \neq j} s_i < 0) > 0$.*

In the BSS problem considered in [6], the assumption of independence of the stochastic variables was necessary for other parts of the algorithm. The following theorem shows that a stochastic vector being sufficiently spread is a necessary and sufficient condition for the detection of negative elements in a matrix.

Theorem E.3 *Let $s \in \mathbb{R}^n$ be a stochastic vector, $u \in \mathbb{R}^n$ and U be any matrix with n rows. Then the following are equivalent:*

1. *$p(\min(Us) < 0) > 0$ if and only if there is at least one negative element of U.*
2. *$p(u^T s < 0) > 0$ if and only if there is at least one negattive element of u.*
3. *s is sufficiently spread.*

Proof. The proof follows the same steps as the proof of Theorem 1 in [6]. ■ In words, a vector is sufficiently spread if it is possible that any element can be much larger than all the other elements in the vector. Below are some examples where s is sufficiently spread (A – D) and some examples where s is not sufficiently spread (E – H).

Ex. A. Let s_1 to s_{n-1} be exponential independent and identically distributed (IID) and let $s_n = 1$.

Ex. B. Let s_1 and s_2 be uniformly IID on the interval $(0, 1)$ and $s_3 = 1 - s_2$.

Ex. C. Let $t \in \mathbb{R}^n$ be sufficiently spread and $s = \frac{t}{\|t\|}$.

Ex. D. Let $t \in \mathbb{R}^n$ be sufficiently spread, P a permutation matrix, D a diagonal matrix with positive elements and $s = PDt$.

Ex. E. Let s_1 to s_{n-1} be uniformly IID on the interval $(0, 1)$ and let $s_n = 1$.

Ex. F. Let $t \in \mathbb{R}^n$ be uniformly IID on the interval $(1, 2)$, $n > 2$, $p \in \{1, \cdots, n\}$ be uniformly distributed, $s_p = t_p - 1$ and $s_q = t_q$ for all $q \neq p$.

Ex. G. Let $t \in \mathbb{R}^n$ be sufficiently spread, $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ and $s = At$.

Ex. H. Let $t \in \mathbb{R}^n$ be sufficiently spread, $A \in \mathbb{R}^{n \times n}$ be a matrix with more than one non-zero element in one row and $s = At$.

The vector in example A are sufficiently spread since the exponential distribution has no upper limit in contrast to the uniform distribution that has an upper limit and are used in example E. Example B show that the elements in s can be dependent and sufficiently spread at the same time, and example G and H shows dependent not sufficiently spread vectors. Examples C and D show that rescaling of a stochastic vector has no influence on sufficient spreadness. Examples F illustrate that a stochastic vector can not be sufficiently spread if only one element at a time can be close to zero.

3 Uniqueness and NMF

In this section, NMF is analyzed. We assume that both W and H have full rank, e.g., $r = \operatorname{rank}(V)$. Let W' and H' be any matrices of the same size as W and H respectively that fulfil $V = WH = W'H'$. Then

$$\operatorname{span}(W) = \operatorname{span}(V) = \operatorname{span}(W'), \tag{3}$$

where $\operatorname{span}()$ is the space spanned by the column vectors. The column vectors of W and W' are therefore both bases for the same space and there exists a matrix Q such that $W' = WQ$. It is straightforward to show that $H' = Q^{-1}H$. All NMF solutions where $r = \operatorname{rank}(V)$ are therefore on the form $V = WQQ^{-1}H$ and the ambiguity of the NMF can therefore be described by the Q matrix. We note that if $r > \operatorname{rank}(V)$, the above argument does not hold since $\operatorname{span}(W)$ may not equal $\operatorname{span}(W')$ (see Thomas [7]). It is possible for an NMF to be unique when $r > \operatorname{rank}(V)$, but we are here only concerned with the situation where $r = \operatorname{rank}(V)$.

Lemma E.4 *If a matrix and its inverse are both non-negative, then the matrix is a scaled permutation.*

Corollary E.5 *An NMF solution W and H can only be estimated uniquely up to a permutation and a scaling.*

Corollary E.5 leads to the following definition of unique NMF.

Definition E.6 *A NMF is called unique if the ambiguity is a permutation and a scaling of the columns in W and rows in H.*

In the following each row of W and each column of H will be seen as stochastic vectors. It will be assumed that it is possible to increase the size of V and thereby increase the number of rows in W and columns in H. If V is a spectrogram, this is the same as using more data (increasing H) and using a higher spectral resolution (increasing W). Using the previous theorems, it can be shown that if the row vectors in W and column vectors in H are sufficiently spread and statistically independent (or dependent in a nondeterministic way), the NMF factorization is unique. Theorem E.3 ensures that Q must have non-negative elements for $W' = WQ$ to be non-negative, and that Q^{-1} must have non-negative elements for $H' = Q^{-1}H$ to be non-negative. Additionally, Lemma E.4 states that if the elements of Q and Q^{-1} are non-negative, then Q is a scaled permutation matrix and therefore the factorization is unique.

In some practical situations, it is not reasonable to assume that both the row vectors in W and column vectors in H are sufficiently spread. A looser condition is therefore introduced next:

Definition E.7 *A positive stochastic vector $s \in \mathbb{R}^n$ is called boundary close if*

$$p\Big(\frac{s_j}{\|s\|} < k\Big) > 0$$

for all $j \in \{1, \cdots, n\}$ and $k > 0$.

It can be verified that the sufficiently spread property implies the boundary close property by choosing $\epsilon < k$, whereby we get

$$\underbrace{\frac{s_j}{\|s\|} \le \frac{s_j}{s_i} \le \frac{\sum_{j\neq i} s_j}{s_i} < \epsilon < k}_{\Downarrow} \tag{4}$$

$$p\Big(\frac{s_j}{\|s\|} < k\Big) > p\Big(\frac{\sum_{j\neq i} s_j}{s_i} < \epsilon\Big) \tag{5}$$

$$= p\Big(-s_i\epsilon + \sum_{j\neq i} s_j < 0\Big) > 0. \tag{6}$$

If the elements in a vector are IID, the boundary close property is equivalent to the sufficiently spread property. When the elements are not IID, the two definitions differ. If a vector is boundary close, all elements can be very small. If a vector is sufficiently spread, all elements except one can be very small. On the previous side s is boundary close in example A – D because it is sufficiently spread. Example E is not boundary close of the same reasons that it is not sufficiently spread. Example F is boundary close. Example G is also boundary close, but if $s_1 = 0$ then $s_2 = s_3$. Later, in Theorem E.8, this will be called "boundary close in a deterministic way". In example H there is not enough information to conclude if s is boundary close or not. A realization of a vector that is sufficiently spread is depicted in Figure 1 to the left and a boundary close vector is shown to the right.

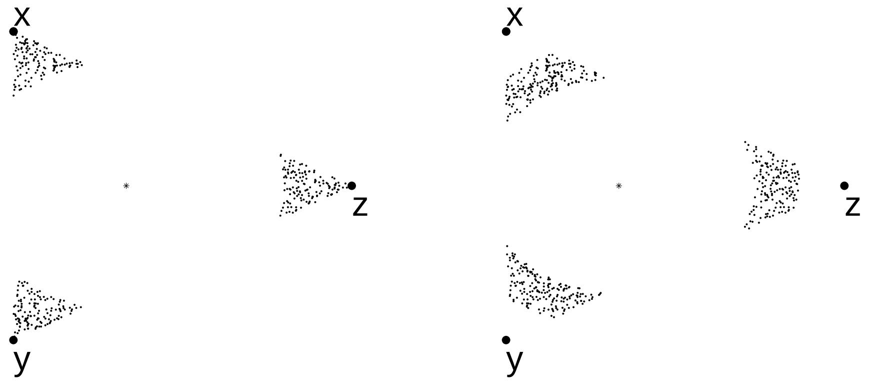


Figure 1: A three dimensional space is scaled such that the vectors are in the hyper plane: $\{p : [1\,1\,1]p = 1\}$. The big dots in the corners are the standard basis vectors for the three dimensional space. To the left, where a sufficiently spread vector is shown, it can be seen that the data fills out the area around the corners. To the right, where a boundary close vector is shown, it can be seen that the data is close to the lines that connects the corners.

Theorem E.8 *If the row vectors in W are strongly boundary close*[1]*, the column vectors in H are sufficiently spread and W and H are independent. Then, given $p < 1$ and $k > 0$, there exist m and n such that any NMF-solution W' and H' will fulfill:*

$$p\Big(\min_{D,P} \|W'DP - W\|_F + \big\|(DP)^{-1}H' - H\big\|_F\,) < k\Big) > p \qquad (7)$$

where P is a permutation matrix and D is a diagonal matrix.

Theorem E.8 states that the NMF is unique if the row vectors in W are boundary close and the column vectors in H are sufficiently spread.

Recently, it has been argued that some real world data has an inherent offset that leads to non-uniqueness for traditional NMF algorithms [8]. However, in [8], this offset was taken into account and an algorithm that can find the constructing NMF even if the data contains an offset was proposed. If the parts of the model that does not fulfill Theorem E.8 has a known structure, it is still possible to find the true W and H.

It is worth noting that the theorem specify the uniqueness of NMF from a solution W', H' and not directly from V. This might seam as a limitation for the theorem, but it can also be used directly to suggest new algorithms as explained in the following Section.

[1] Strongly boundary close is the same as being boundary close in an non-deterministic way. The extra condition is somewhat technical and is not restrictive in practice, and the exact definition is therefore left out.

4 Some Applications

We now proceed to describe how the information of a unique solution can be used to estimate parameters such as the number of constructing vectors r or a sparsity parameter. When an NMF solution results in a W that has boundary close row vectors and in an H having sufficiently spread column vectors, it can be determined that the solution is unique. Since sufficiently spread is the hardest condition, it will typically be the one that is not fulfilled. Hojer [9] introduced the sparse NMF where the update rule of NMF is changed in order to obtain an NMF that has a sparse H, meaning that H has few non-zero coefficients. From Theorem E.8 it follows that a minimum number of zeros are needed in both W and H for the factorization to be unique. Especially H need a lot of zeros in order to be sufficient spread. Algorithms that find NMF with sparse H are therefore expected to have a higher probability for returning unique factorizations. Many sparse NMF algorithms have a sparsity parameter that determines the trade-off between sparsity of H and the approximation error. One example of this is the sparse NMF algorithm of Eggert and Körner [10], where λ determines the tradeoff in the error function defined as

$$E(W,H) = \frac{1}{2}\left\|V - \overline{W}H\right\|_F^2 + \lambda \mathbf{1}^T H \mathbf{1} \tag{8}$$

$$\overline{W}_j = \frac{W_j}{\|W_j\|} \quad , j \in \{1, \ldots, n\}, \tag{9}$$

where W_j is the jth column vector of W and $\mathbf{1}$ is a column vector where all elements are one. Based on Theorem E.8, a natural criterion for selection of the sparsity parameter can be made; select λ as the smallest value that makes the solution unique. In a similar manner, a natural selection of a model order, r, is the ones that makes the NMF unique. In a third application, Theorem E.8 can be used for NMF algorithm comparison, i.e., it can be used as a basis for selecting the algorithm that makes unique solution. Next, two examples are given showing how Theorem E.8 can be used to argue that a data set has a unique NMF. Smaragdis and Brown [11] use NMF for polyphonic music transcription where V is the amplitude spectrogram. In this setup it is fair to assume that there exist a W where the column vectors are the amplitude spectrum of an instrument that play one note and a corresponding H that describe the identity of the notes. It can be seen from their experiments that W is boundary close. In music, it happens that there is a single note that is playing and in that case H is sufficiently spread. If an NMF is found for large set recordings of polyphonic music and the correct order is selected, the solution is therefore expected to be unique. Another example is probabilistic latent semantic analysis (PLSA) [12], which is a text analyzing method that gather information of several documents in a "bag of words" matrix V, where $V_{i,j}$ describes how many times the i'th keyword occurs in the j'th document. In PLSA, V is decomposed into two positive matrixes using an NMF like algorithm. Each column in W can be seen as a topic and in all topics there will be keywords that does not occur and W is therefore boundary close. If there are documents that only cover one topic, H will be sufficiently spread. The NMF on a "bag of words" matrix is therefore expected to be unique.

5 Conclusion

Uniqueness of non-negative matrix factorizations have been analyzed by combining results of Donoho and Stodden [4] and Plumbley [6]. A new condition for a stochastic vector called sufficiently spread has been introduced and it has been shown that this condition is a sufficient and necessary for the dection of a negative value in a matrix. A weaker condition, called boundary close, has also been introduced, and a theorem that states that the factorization is unique if the row vectors in W is boundary close and the column vectors in H are sufficiently spread has been described. The theorem is valid when there is no prior knowledge of W and H. In situations where there is prior knowledge, an NMF can be unique even when the conditions are not fulfilled. An analysis shows that NMF on a bag of word matrix and on music amplitude spectrograms are expected to be unique. The theorem can be used as an explanation why the sparse NMF methods tend to result in unique factorizations. The theorems make it possible to evaluate the uniqueness of a factorization and can thereby be used for choosing sparsity parameter, model order and NMF algorithm.

Acknowledgment

Most of this work was performed while the author was a visiting PhD student at the Intelligent Signal Processing Group of IMM, Technical University of Denmark. I would like to thank Søren Holdt Jensen and Mads Græsbøll Christensen for many helpful comments on the manuscript.

References

[1] D.D. Lee and S.H. Seung, “Learning the parts of objects by non-negative matrix factorization.,” *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[2] D.D. Lee and S.H. Sebastian, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2000, pp. 556–562.

[3] M. W. Berry et al., “Algorithms and applications for approximate nonnegative matrix factorization,” *Computational Statistics and Data Analysis*, 2006, in press, doi: 10.1016/j.csda.2006.11.006.

[4] D. Donoho and V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts?,” in *Advances in Neural Information Processing*

Systems 16, S. Thrun, L. Saul, and B. Schölkopf, Eds. MIT Press, Cambridge, MA, 2004.

[5] F.J. Theis, K. Stadlthanner, and T. Tanaka, “First results on uniqueness of sparse non-negative matrix factorization,” in *European Sig. Proc. Conf. (EUSIPCO)*, 2005.

[6] M. Plumbley, “Conditions for nonnegative independent component analysis,” *IEEE Signal Processing Lett.*, vol. 9, no. 6, pp. 177–180, 2002.

[7] L.B. Thomas, “Solution to problem 73-14, rank factorizations of nonnegative matrices by A. Berman and R. J. Plemmons,” *SIAM Review*, vol. 16, pp. 393–Ű394, 1974.

[8] H. Laurberg and L.K. Hansen, “On affine non-negative matrix factorization,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2007, vol. II, pp. 653–656.

[9] P. O. Hoyer, “Non-negative sparse coding,” in *IEEE Proc. Neural Networks for Signal Processing*, 2002, pp. 557–565.

[10] J. Eggert and E. Körner, “Sparse coding and NMF,” in *Proc. Int. Joint Conf. on Neural Networks*, 2004, pp. 2529–2533.

[11] P. Smaragdis and J.C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2003, pp. 177–180.

[12] T. Hofmann, “Probabilistic latent semantic analysis,” in *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, San Francisco, CA, 1999, pp. 289–29.