

UNIVERSITÉ DE MONS  
FACULTÉ POLYTECHNIQUE  
MATHÉMATIQUE ET RECHERCHE OPÉRATIONNELLE

NONNEGATIVE MATRIX FACTORIZATION  
MODELS, OPTIMIZATION PROBLEMS, ALGORITHMS  
AND APPLICATIONS

VALENTIN LEPLAT

A thesis submitted  
in partial fulfillment of the requirements for the degree of  
Doctorat en Science de l'ingénieur et technologie

Dissertation committee:

Prof. Thierry Dutoit	Université de Mons (Chair)
Prof. Nicolas Gillis	Université de Mons (Supervisor)
Prof. Xavier Siebert	Université de Mons (Co-supervisor)
Prof. Cédric Févotte	CNRS, Université de Toulouse
Prof. Yurii Nesterov	UCLouvain
Prof. Panos Patrinos	KULeuven

November, 2020

## Abstract

The low-rank approximation of a matrix is a key problem in data analysis, and is widely used for Linear Dimensionality Reduction (LDR). LDR techniques such as principal component analysis are powerful tools for the analysis of high-dimensional data. In this thesis, we explore a popular variant of LDR, namely Nonnegative Matrix Factorization (NMF), which consists in a low-rank matrix approximation problem with nonnegativity constraints. More precisely, we seek to approximate a given nonnegative matrix  $V$  with the product of two nonnegative matrices,  $W$  and  $H$ , of smaller size. Even if, at first glance, the nonnegativity requirement seems to be restrictive in terms of practical use, it is not: NMF has many applications. Indeed, nonnegativity of the solution is required in many fields such as probability, geoscience, medical imagery, computational geometry, combinatorial optimization, analytical chemistry and machine learning.

The nonnegativity constraints allow to extract easily interpretable and meaningful information from the input data. However, they make the problem much more difficult to solve (NP-hard). The contributions of this thesis are centered on NMF over four aspects: models, optimization problems, algorithms and applications.

The two first aspects are explored by proposing models and optimization problems for NMF with volume regularization, usually referred to as minimum-volume NMF. In recent years, the minimum-volume NMF has shown to be a powerful approach to compute meaningful solutions for  $W$  and  $H$ . In this thesis, we show that our new models and optimization problems, under some mild conditions, lead to identifiability, that is, the solution of the optimization problems is unique up to ambiguities that are unavoidable and, most importantly, inconsequential for the applications at hand. Further, we propose a new class of models and optimization problems, referred to as multi-resolution NMF, to tackle a common issue for many input matrices; they are generally the result of a resolution trade-off between two adversarial dimensions. We address this issue by fusing the information coming from multiple data sets with different resolutions in order to produce a factorization with high resolutions for all the dimensions. Finally we propose a novel approach to tackle a special case of NMF referred to as exact NMF by using conic programming.

On the algorithmic aspect, we propose efficient algorithms to solve the proposed optimization problems for minimum-volume NMF. In this thesis we mainly focus on two classes of optimization problems for minimum-volume NMF: the first one integrates a Frobenius norm for the data fitting term whereas the second one integrates the family of  $\beta$ -divergences, in particular we deal with the Kullback-Leibler divergence that is notorious hard to handle. Further we introduce a general framework to derive algorithms to tackle penalized  $\beta$ -divergence NMF problems under disjoint equality constraints. Finally we propose two algorithms relying on conic programming that are able to tackle problems to compute an exact NMF.

On the application aspect, we demonstrate the efficiency of our algorithms compared to state-of-the-art algorithms on hyperspectral imaging and audio source separation problems.

---

## Declaration

I declare that this thesis is my own work

Signed \_\_\_\_\_

Valentin Leplat

Wednesday 4<sup>th</sup> November, 2020

## Acknowledgement

I have many people to thank for their patience and their support and it is important to recall that this work, as humble as it is, is mainly the result of discussions and interactions with my colleagues and advisors; "Man is a social animal".

I would like to thank my advisors, Nicolas Gillis and Xavier Siebert. I have enjoyed working with them since my master's thesis, and they constantly supported and encouraged me for my different projects. Their suggestions were always insightful and helped me a lot to understand the world of research. I thank my colleagues from MARO unit for their support for my tasks of teaching assistant, I thank in particular Daniel Tuytens for his availability and for all the support he gave when I needed it. I also thank Gwendolyn Lacroix for her work that greatly facilitated our teaching load and, of course, I thank for their trust the professors of all the courses in which I had the privilege to intervene.

I am grateful to my support committee, Cédric Févotte and Thierry Dutoit, with whom I had the chance to discuss my problems, and benefit from their experience. I thank in particular Cédric Févotte for his warm welcome during my short stay in Toulouse in June 2018 and our fruit full collaboration on the multi-resolution NMF project.

I also thank Yurii Nesterov for his kind welcome in CORE during my PhD visits in 2019, his patience and his remarkable mathematical intuitions. I had the privilege to follow his lectures in UCLouvain and his works gave me the taste to pursue my research in nonlinear programming.

I would also like to thank Jeremy E. Cohen for inspiring discussions when he was a post-doctoral researcher in Mons. Then, I would like to thank my collaborator Andersen Man Shun Ang for our fruitful collaborations and motivating discussions. I think we could spend ten years arguing and discussing and we would still have innovating ideas to start research projects.

I am also grateful to the other people who have helped me in various ways, and have interacted with this work at some point and my colleagues from COLORAMAP team for the great work atmosphere they contribute to build: Arnaud Vandaele, Punit Sharma, L. T. K. Hien, Junjun Pan (thank you for helping me for my Chinese lectures), Nicolas Nadisic,

Francois Moutier, Timothy Marrinan, Pierre De Handschutter, Christophe Kervazo, and Maryam Abdolali.

Last but not least, I thank my parents. I dedicate this thesis:

- to my life-partner Ingrid; without her love, patience and support, this thesis would have stayed in my mind.
- to Monique Jeunechamps and Pierre Lousberg who gave me the taste for applied mathematics during my first engineering degree in Liège.
- to the logos, typically known as the common sense who has taken holidays, I hope he is alright and I'm looking forward to see him.

*“Les mathématiques sont une grammaire de la nature. Ce sont les chaussures de la technique. On peut marcher sans chaussures, mais on va moins loin.”*

- Jean-Marie Souriau, Grammaire de la nature.

*"Je suis convaincu d'une chose: le talent, cela n'existe pas. Le talent, c'est avoir l'envie de faire quelque-chose. Je prétends qu'un homme qui rêve tout d'un coup; il a envie de manger un homard. Il a le talent, à ce moment-là, dans l'instant, pour manger convenablement un homard, pour le savourer convenablement. Et je crois qu'avoir envie de réaliser un rêve, c'est le talent et tout le restant c'est de la sueur, c'est de la transpiration, c'est de la discipline. Je suis sûr de cela. L'art, je ne sais pas ce que c'est, les artistes; je ne connais pas. Je crois qu'il y a des gens qui travaillent à quelque-chose, et qui travaillent avec une grande énergie finalement et l'accident de la nature, je n'y crois pas.*  
"

- Jacques Brel, Interview 1971 à Knokke.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Declaration</b>	<b>iii</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 A short message from the author . . . . .	1
1.2 Thesis structure . . . . .	2
1.3 From dimensionality reduction techniques to NMF . . . . .	4
1.4 NMF: the standard problem and models definition . . . . .	7
1.5 The breakthrough experiment for NMF . . . . .	8
1.5.1 Hyperspectral imaging . . . . .	11
1.5.2 Single-channel blind source separation . . . . .	14
1.6 Geometry of NMF . . . . .	18
1.7 Computational complexity of NMF . . . . .	24
1.8 From model uniqueness to identifiable NMF . . . . .	27
1.8.1 Identifiability for NMF . . . . .	27
1.8.2 Sufficient conditions . . . . .	29
1.8.3 Separable NMF . . . . .	37
1.8.4 Minimum-volume NMF . . . . .	37
1.9 Approximate factorization . . . . .	42
1.9.1 The metrics: the $\beta$ -divergences . . . . .	43
1.9.2 The probabilistic view of NMF . . . . .	45
1.9.3 How to choose the metric in practice ? . . . . .	47
1.9.4 Distributionally Robust and Multi-Objective Nonnegative Matrix Factorization . . . . .	48
1.10 Standard optimization schemes . . . . .	49
<b>2 Minimum-Volume Rank-Deficient Nonnegative Matrix Factorizations</b>	<b>51</b>
2.1 Introduction . . . . .	51
2.2 Minimum-volume NMF in the rank-deficient case . . . . .	53

2.3	Algorithm for minimum-volume NMF . . . . .	55
2.4	Numerical Experiments . . . . .	56
2.5	Faster Algorithm for minimum-volume NMF . . . . .	59
2.6	Conclusions . . . . .	63
<b>3</b>	<b>Minimum-Volume <math>\beta</math>-NMF for blind audio source separation</b>	<b>65</b>
3.1	Introduction: NMF for audio source separation . . . . .	66
3.2	Minimum-volume $\beta$ -NMF . . . . .	66
3.2.1	Geometry and min-vol $\beta$ -NMF problem . . . . .	67
3.2.2	Normalization and identifiability . . . . .	67
3.3	Algorithm for min-vol $\beta$ -NMF . . . . .	71
3.3.1	Separable auxiliary functions for $\beta$ -divergences . . . . .	72
3.3.2	A separable auxiliary function for the minimum-volume regularizer . . . . .	72
3.3.3	Auxiliary function for min-vol $\beta$ -NMF . . . . .	75
3.3.4	Algorithm for min-vol KL-NMF . . . . .	75
3.3.5	Algorithm for min-vol IS-NMF . . . . .	78
3.4	Numerical experiments . . . . .	78
3.5	Conclusion and Perspectives . . . . .	88
<b>4</b>	<b>Multi-resolution <math>\beta</math>-NMF for blind spectral unmixing</b>	<b>90</b>
4.1	Introduction . . . . .	90
4.2	Problem formulation . . . . .	92
4.3	Algorithm for MR- $\beta$ -NMF problem . . . . .	95
4.4	Numerical experiments on audio data sets . . . . .	99
4.4.1	Experimental setup and evaluation . . . . .	100
4.4.2	Results . . . . .	102
4.5	Numerical experiments on HS-MS fusion . . . . .	105
4.5.1	Test setup and criteria . . . . .	105
4.5.2	Experimental results . . . . .	112
4.6	Conclusions and outlooks . . . . .	112
<b>5</b>	<b>Multiplicative Updates for NMF with <math>\beta</math>-divergences Under Disjoint Equality Constraints</b>	<b>118</b>
5.1	Introduction . . . . .	118
5.2	General framework to design MU for $\beta$ -NMF under disjoint linear equality constraints and penalization . . . . .	120
5.2.1	Separable majorizer for the objective function . . . . .	121
5.2.2	Dealing with equality constraints via Lagrange dual variables . . . . .	123
5.2.3	MU for $\beta$ -NMF with disjoint linear equality constraints without penalization . . . . .	127
5.3	Showcase 1: Simplex-structured $\beta$ -NMF . . . . .	130

---

5.4	Showcase 2: minimum-volume KL-NMF . . . . .	133
5.4.1	Problem formulation and algorithm . . . . .	134
5.4.2	Numerical experiments . . . . .	135
5.5	Extension to quadratic disjoint constraints . . . . .	137
5.5.1	Problem formulation and algorithm . . . . .	138
5.5.2	Numerical experiments . . . . .	140
5.6	Conclusion . . . . .	142
<b>6</b>	<b>Exact NMF with conic programming</b>	<b>148</b>
6.1	Introduction and preliminaries . . . . .	148
6.1.1	Some properties of the nonnegative rank . . . . .	149
6.1.2	Conic programming . . . . .	150
6.2	Problem formulations for exact NMF . . . . .	153
6.2.1	Problem formulation via exponential cones . . . . .	153
6.2.2	Problem formulation via rotated quadratic cones . . . . .	154
6.3	Algorithm . . . . .	155
6.3.1	Sparsity Pattern Integration . . . . .	158
6.4	Numerical experiments . . . . .	161
6.4.1	Benchmark Nonnegative Matrices for Exact NMF . . . . .	161
6.4.2	The largest biclique in a bipartite graph . . . . .	164
6.5	Conclusion . . . . .	166
<b>7</b>	<b>Conclusion</b>	<b>168</b>
	<b>Bibliography</b>	<b>172</b>
	<b>Appendix</b>	<b>185</b>
1	Symbols . . . . .	185
2	Acronyms . . . . .	187
3	A brief introduction to convergence theory of popular BCD schemes . . . . .	188
4	Behaviour of the nonnegative rank under perturbations . . . . .	193
5	Convexity, concavity and complete monotonicity for a convex-concave decomposition of the discrete $\beta$ -divergence . . . . .	196



## List of Figures

1.1	The breakthrough experiment that puts NMF on stage. . . . .	10
1.2	Data cube for images. . . . .	12
1.3	Hyperspectral data cube. . . . .	12
1.4	The linear mixture model for blind spectral unmixing. . . . .	14
1.5	Process for generating an amplitude spectrogram. . . . .	16
1.6	Application of NMF to blind audio source separation. . . . .	18
1.7	Geometric illustration of Exact NMF for $K = F = 3$ and $N = 25$ . We observe that $\text{cone}(V) \subseteq \text{cone}(W) \subseteq \mathbb{R}_+^F$ . . . . .	20
1.8	Geometric illustration of Exact NMF w.r.t. the nested convex hulls on the data set from Figure 1.7. We observe that $\text{conv}(\Pi_{\Delta^F}(V)) \subseteq \text{conv}(\Pi_{\Delta^F}(W)) \subseteq \Delta^F$ . . . . .	22
1.9	An illustration of the ill posedness of NMF for $K = 3 = F$ based on the nested convex hulls interpretation. . . . .	30
1.10	Illustration of the second-order cone $\mathcal{C}$ . . . . .	34
1.11	Illustration of the SSC and Separable conditions. . . . .	35
1.12	Geometric illustration of separable NMF. . . . .	38
1.13	Geometric illustration of the standard exact NMF model (a) and exact NMF model with $H$ being column-stochastic (b), referred to as SSNMF, for $F = K = 3$ . . . . .	40
1.14	Geometric intuition of min-vol NMF. . . . .	41
1.15	Graph of the $\beta$ -divergences $d_\beta(x = 1 y)$ for $\beta = [-1, 0, 1, 2, 3]$ . . . . .	44
2.1	Function $\frac{\log(x^2+\delta)-\log(\delta)}{\log(1+\delta)-\log(\delta)}$ for different values of $\delta$ , $\ell_1$ norm ( $=  x $ ) and $\ell_0$ norm ( $= 0$ for $x = 0$ , $= 1$ otherwise). . . . .	54
2.2	Synthetic data set and recovery. (Only the first three entries of each four-dimensional vector are displayed.) . . . . .	57
2.3	Evolution of the recovery of the true $W$ depending on the noise $N = \epsilon \text{rand}(F,N)$ using Algorithm 2 ( $\tilde{\lambda} = 0.01$ , $\delta = 0.1$ , $\text{maxiter} = 100$ ). . . . .	57
2.4	Abundance maps extract by Algorithm 2 using only five bands of the San Diego airport HSI. From left to right, top to bottom: vegetation (grass and trees), three different types of roof tops, four different types of road surfaces. . . . .	58
2.5	logdet function evolution in the 1-dimensional case with $\delta = 1$ . . . . .	60
2.6	Synthetic data set and recovery for "min-vol" and "fast-min-vol" algorithms. (Only the first three entries of each four-dimensional vector are displayed.) . . . . .	61

2.7	Rates of convergence for "min-vol" and "fast-min-vol" algorithms on synthetic data set . . . . .	62
2.8	Rates of convergence for the objective function of problem (2.4) and the Frobenius norm along iterations for fast-min vol and min-vol algorithms in the rank deficient case for San Diego Dataset. . . . .	62
2.9	Rates of convergence for the objective function of problem (2.4) and the Frobenius norm along iterations for fast-min vol and min-vol algorithms in the high-dimensional case for San Diego Dataset. . . . .	63
3.1	Musical score of "Mary had a little lamb". . . . .	79
3.2	Comparative study of baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to "Mary had a little lamb" amplitude spectrogram with $K=3$ . . . . .	80
3.3	Masking coefficients obtained with baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to "Mary had a little lamb" amplitude spectrogram with $K=3$ . . . . .	81
3.4	Comparative study of baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to "Mary had a little lamb" amplitude spectrogram with $K=7$ . . . . .	83
3.5	Masking coefficients obtained with baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to "Mary had a little lamb" amplitude spectrogram with $K=7$ . . . . .	84
3.6	Factors matrices $W$ and $H$ obtained with min-vol KL-NMF LS with factorization rank $K=7$ for the third audio sample. . . . .	85
3.7	Musical score of the third audio sample. . . . .	86
3.8	Musical score of the sample "Prelude and Fugue No.1 in C major". . . . .	87
3.9	Factors matrices $W$ and $H$ obtained with min-vol KL-NMF LS with factorization rank $K=16$ on the sample "Prelude and Fugue No.1 in C major". . . . .	87
3.10	Validation of the estimate sequence obtained with min-vol KL-NMF LS with factorization rank $K=16$ on the sample "Prelude and Fugue No.1 in C major". . . . .	87
4.1	Columns of $W_{\#}$ , $W$ , $W_Y$ and $W_X$ in semi-log scale. Top, middle and bottom sub-figures show the spectral content respectively for $C_4$ , $D_4$ and $E_4$ . . . . .	103
4.2	Rows of $H$ obtained with MR-KL-NMF with $K = 5$ for $\lambda = 0.001$ and $\lambda = 1$ . . . . .	106
4.3	Columns of $W$ ( $\log_{10}$ scale) obtained with MR-KL-NMF with $K = 5$ for $\lambda = 0.001$ and $\lambda = 1$ . . . . .	107
4.4	Landsat 4 TM relative spectral responses. . . . .	108
4.5	Urban data set for HS-MS fusion problem. . . . .	109
4.6	SAM maps for the different hyperspectral images. . . . .	113

5.1	Averaged objective functions over 20 random initializations obtained for Algorithm 5 (red line with circle markers) and the GR-NMF (black dashed line) applied to the three data sets detailed in the text for 300 iteration. The comparison is performed for different values of $\beta$ , from top to bottom: $\beta = 2$ , $\beta = 3/2$ , and $\beta = 1$ . Logarithmic scale for y axis. . . . .	132
5.2	Averaged objective functions over 20 random initializations obtained for Algorithm 7 with 300 iterations (red line with circle markers), and the heuristic $\beta$ -SNMF from [119] (black dashed line). . . . .	141
5.3	Samson data set (a) and results ((b) and (c)) for the Abundance maps estimated using Algorithm 7 for the three endmembers: #1 Tree, #2 Soil and #3 Water. Two average sparsity levels considered: 0.25 (b) and 0.5 (c). . .	143
5.4	Jasper Ridge data set (a) and results ((b) and (c)) for the Abundance maps estimated using Algorithm 7 for the four endmembers: #1 Road, #2 Tree, #3 Water and #4 Soil. Two average sparsity levels are considered: 0.25 (b) and 0.5 (c). . . . .	144
5.5	Urban data set (a) and results ((b) and (c)) for the Abundance maps estimated using Algorithm 7 for the six endmembers: #1 Soil, #2 Tree, #3 Grass, #4 Roof, #5 Road/Asphalt and #6 Roof2/shadows. Two average sparsity levels are considered: 0.25 (b) and 0.5 (c). . . . .	145
5.6	Baseline abundances for the endmembers obtained for Samson data extracted from [154]: #1 Soil, #2 Tree and #3 Water. . . . .	146
5.7	Baseline abundances for the endmembers obtained for Jasper Ridge data extracted from [154]: #1 Road, #2 Soil, #3 Water and #4 Tree. . . . .	146
5.8	Baseline abundances for the endmembers obtained for Urban data extracted from [154]: #1 Asphalt, #2 Grass, #3 Tree, #4 Roof1, #5 Roof2/Shadow and #6 Soil. . . . .	146
6.1	Boundary of the exponential cone $\mathcal{K}_{exp}$ in the case $n = 3$ . . . . .	151
6.2	Boundaries of $\mathcal{Q}^3$ and $\mathcal{Q}_r^3$ , reproduced from MOSEK doc. . . . .	153
6.3	Evolution of $\frac{\ V-WH\ _F}{\ V\ _F}$ along iterations; SPI is activated in the iterations interval [400, 500]. . . . .	160
6.4	An illustration of the maximum-edge biclique problem for a graph that corresponds to the biadjacency matrix (6.20). . . . .	165
1	Illustration of the MM principle on a 1-dimensional problem. . . . .	190

## List of Tables

3.1	Differentiable convex-concave-constant decomposition of the $\beta$ -divergence under the form (3.6) [45]. . . . .	72
3.2	Multiplicative update for min-vol KL-NMF. . . . .	76
3.3	SDR, SIR and SAR metrics comparison for results obtained with baseline KL-NMF and min-vol KL-NMF LS on a synthetic mix of bass and drums . . . . .	86
3.4	Runtime performance in seconds of baseline KL-NMF, min-vol KL-NMF LS (Algorithm 3) and sparse KL-NMF [119]. The table reports the average and standard deviation over 20 random initializations for three experimental setups described in the text. . . . .	88
4.1	Multiplicative updates for MR- $\beta$ -NMF. . . . .	99
4.2	Comparison of MR- $\beta$ -NMF with baseline $\beta$ -NMF in terms of SNR on the activations and the dictionary vectors with respect to true factors on the dataset 1. The table reports the average, standard deviation and the best SNR over 100 random initializations for $W$ and $H$ . Bold numbers indicate the highest SNR. . . . .	103
4.3	Comparison of MR- $\beta$ -NMF with baseline $\beta$ -NMF in terms of SNR on the activations and the dictionary vectors with respect to true factors on the dataset 2. The table reports the average, standard deviation and the best SNR over 100 random initializations for $W$ and $H$ . Bold numbers indicate the highest SNR. . . . .	104
4.4	Comparison of MR- $\beta$ -NMF with state-of-the-arts methods for HS-MS fusion problem on dataset HYDICE Urban. The table reports the average, standard deviation for the quantitative quality assessments over 20 trials. Bold, underlined and italic to highlight the three best algorithms. . . . .	114
4.5	Comparison of MR- $\beta$ -NMF with state-of-the-arts methods for HS-MS fusion problem on dataset HYDICE Washington DC Mall. The table reports the average, standard deviation for the quantitative quality assessments over 20 trials. Bold, underlined and italic to highlight the three best algorithms. . . . .	115
4.6	Comparison of MR- $\beta$ -NMF with state-of-the-arts methods for HS-MS fusion problem on dataset AVIRIS Indian Pines. The table reports the average, standard deviation for the quantitative quality assessments over 20 trials. Bold, underlined and italic to highlight the three best algorithms. . . . .	116

5.1	Cases where (5.21) can be computed in closed form. They are indicated by the degree of the corresponding polynomial equation, otherwise the symbol $\pm$ is used. The constants $L_{kn}$ 's is the one needed in Assumption 5.2.1 for the penalization functions $\Phi_1(H)$ and $\Phi_2(W)$ ; see (5.6). . . . .	127
5.2	Runtime performance in seconds and final value of objective function $F_{\text{end}}(W, H)$ for Algorithm 5 and the GR-NMFreported for $\beta \in \{0, \frac{1}{2}, 1, \frac{3}{2}, 2\}$ . The table reports the average and standard deviation over 20 random initializations with a maximum of 300 iterations for three hyperspectral data sets. . . . .	133
5.3	Runtime performance in seconds of baseline KL-NMF, Algorithm 3 and Algorithm 6. The table reports the average and standard deviation over 20 random initializations. . . . .	137
5.4	Final values for $D_\beta$ and the penalized objective $\Psi$ from (3.1) obtained with Algorithm 3 and Algorithm 6. The table reports the average and standard deviation over 20 random initializations for three experimental setups. . . . .	137
5.5	Runtime performance in seconds and final value of objective function $\Phi_{\text{end}}(W, H)$ for Algorithm 7 and $\beta$ -SNMF. The table reports the average and standard deviation over 20 random initializations with a maximum of 300 iterations for three hyperspectral data sets. . . . .	140
5.6	Final value of objective function values $\Phi_{\text{end}}(W, H)$ for Algorithm 7 and the heuristic from [119]. The table reports the average and standard deviation over 20 random initializations for an equal computational time that corresponds to 300 iterations of Algorithm 7. . . . .	141
6.1	Important convex cones and the associated case of CP . . . . .	150
6.2	Comparison of Algorithm 8 with algorithm from [133] with "ms1" and "rbr" heuristic for 10 attempts to compute the factorizations of matrices described in the text. In bold we specify the matrices for which SCCAE-NMF is the only one to find exact NMF's. . . . .	164
6.3	Comparison of Algorithm 8 with algorithm from [59] for finding the maximum-edge biclique from a bipartite graph. The table reports the number of edges for the largest biclique extracted by the two methods from bipartite graphs defined by random binary matrices $V$ of size $N \times N$ . . . . .	166
1	domain of $\frac{\partial d_\beta(x,y)}{\partial y}$ depending on the values of $x$ and $\beta$ . . . . .	192
2	Proposed concave-convex decomposition of the discrete $\beta$ -divergence. . . . .	196



# 1 Introduction

*"You know that the beginning is the most important part of any work,..."*

- Plato, The republic, Book II.

## 1.1 A short message from the author

This thesis summarizes my research conducted from January 2018 until September 2020 which has been dedicated to the development of models and algorithms for a widely used and popular Linear Dimensionality Reduction (LDR) called Nonnegative Matrix Factorization (NMF). NMF belongs to the field of data sciences and has shown a great deal of interest since many years as, whether we want it or not, we are surrounded by data. The data continuously grows year after year and processing such amount of data becomes a hard challenge. In the context of this thesis, we are interested by identifying the underlying structure of a data set and extracting meaningful information. Indeed, data does not imply information. Data can be raw, not structured, incomplete and not exploitable by Human beings or automatic systems. Information, on the opposite, should be coherent, meaningful and helpful for good diagnostics. This thesis has been written in an inflection period for recent Human history: the rate of data exchanges is huge and the way to extract information from data and the way to present it is a source of power and significant influence. We have seen the impact of repeated and unchecked information in this COVID period on our fellow citizens, we have seen inefficient <sup>1</sup> liberticide political decisions based on unreliable predictive models (from Neil Ferguson and his team <sup>2</sup>) and it is clear that we need now to understand the consequence of the data and demand sound public debates when it comes to interpret such data and take political decisions. I close this interlude by writing down a couple of advices I usually give to students in Statistics which seem, at first glance, trivial but easily put on the side when emotions enter the game: always ask the relative and the absolute values for the statistics, do not rely on first-order moment only, demand the standard deviation at least, consider with the highest degree of vigilance any mono-variate studies, and machine learning is not the synonym of "truth". Behind every model, there are parameters to choose, potentially with ideological biases or business interests. The beauty of mathematics belongs into the rationale and the sound logic, what the ancient Greeks called the logos. The logos, initially and formally discussed by Plato and Aristotle, is a transparent, universal, time and space invariant method that ensures

---

<sup>1</sup><https://bit.ly/2Tzt2J7>

<sup>2</sup><https://bit.ly/34A6f62>

a fair and correct thinking, here lies the key to exactitude (truth is too pretentious), here lies the key to "Le Vrai, le Bien, le Beau". Now, you know what is the core-stone of my training and my motivation, it is time to dive in the technical aspects.

## 1.2 Thesis structure

This thesis is the concatenation of the research outputs of the authors and its coauthors that consist in one journal paper [90], two conference papers [93, 89], three journal preprints [62, 91, 92] under review and one working paper on the content of Chapter 6. Most of the chapters of this thesis appear in these papers. The present thesis is organized as follows:

- **Chapter 1 Introduction:** Some theoretical background needed throughout this thesis is presented. This Chapter serves as the technical backbone for Chapters 2 to 6. In particular:
  - Section 1.3 introduces NMF as a special case of LDR techniques.
  - Section 1.4 formally defines the notion of NMF models and the standard optimization problems associated to these models that require to be solved to compute the NMF of an input data matrix.
  - Sections 1.5 presents the breakthrough experiment that put NMF on stage in 2000. It shows in particular that NMF is able to extract easily representable and meaningful results from the data. Finally sections 1.5.1 and 1.5.2 present the two main applications that we consider in this thesis to test our models and algorithms.
  - Section 1.6 presents two geometrical interpretations of NMF. These geometrical interpretation will be useful to understand:
    - \* one of the main issue about NMF models, that is, the nonuniqueness. We discuss in Section 1.8 the recent results and models from the literature useful to address this issue. The material of Section 1.8 is the backbone for the theoretical contributions presented in Chapter 3.
    - \* to understand the main motivation for the NMF problems introduced in Chapters 2 and 3 that rely on a "minimum-volume" regularization.
- **Chapter 2:** Nonnegative matrix factorization (NMF) with volume regularization has been shown to be a powerful approach to identify the latent factors that generated the data for many applications such as hyperspectral unmixing, document classification, etc. In this Chapter, we show that minimum-volume NMF can also be used when the basis matrix is rank deficient, which is a reasonable scenario for some real-world NMF problems (e.g., for unmixing multispectral images). We propose an efficient algorithm to tackle the optimization problems for minimum-volume NMF and we



show its efficiency for NMF applications for which the basis matrix is rank-deficient. The material of this Chapter appears in [89].

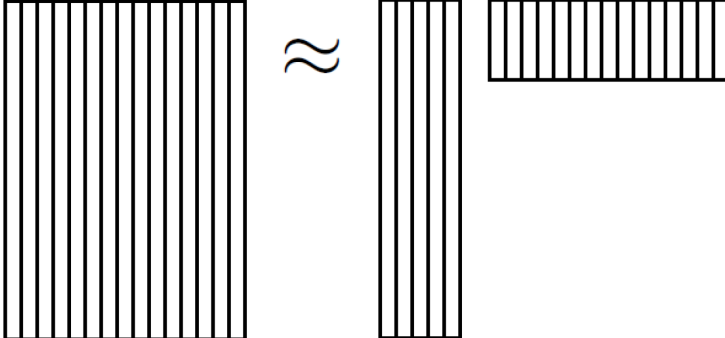
- **Chapter 3:** This Chapter presents a new class of model and associated optimization problems, dubbed as min-vol  $\beta$ -NMF, that integrate a  $\beta$ -divergence and a minimum-volume regularization. We show that, in the noiseless case and under mild conditions, this model and the associated optimization problems provably identify the latent factors. We introduce an efficient algorithm to tackle these optimization problems. We showcase our algorithm to tackle the blind audio source separation problem. The material of this Chapter appears in [93, 90].
- **Chapter 4:** We first give a brief introduction to the Blind spectral unmixing problem. This problem is typically tackled with NMF-based methods by factorizing a data matrix that is the result of a resolution trade-off between two adversarial dimensions. In this Chapter, we propose a new NMF-based framework, dubbed multi-resolution  $\beta$ -NMF (MR- $\beta$ -NMF), to address this issue by fusing the information coming from multiple data with different resolutions in order to produce a factorization with high resolutions for all the dimensions. The material of this Chapter appears in [91].
- **Chapter 5:** This Chapter introduces a general framework to design multiplicative updates (MU) for NMF problems based on  $\beta$ -divergences with disjoint equality constraints, and with penalty terms in the objective function. Our MU satisfy the set of constraints after each update of the variables during the optimization process, while guaranteeing that the objective function decreases monotonically. The material of this Chapter appears in [92].
- **Chapter 6:** We discuss in this Chapter how to use conic programming to compute an exact NMF for an input matrix. We introduce a novel framework that includes two approaches for computing an exact NMF. Each of the proposed approaches relies on the construction and the resolution of a specific optimization problem. For each optimization problem we introduce a particular change of variables that enables the use of two special cases of conic constraints, that are the exponential and second-order conic constraints. Then we propose a general algorithm that is able to tackle both problems in a unified manner and that solves a sequence of conic problems. We finally show that our algorithm is able to compute exact NMF for several classes of nonnegative matrices, we also show that our framework is flexible and can be used to tackle other problems such the maximum-edge biclique problem.
- **Chapter 7 Conclusion:** We conclude the thesis by summarizing the contributions of the Chapters 2 to 6, we attempt to put them in perspective, recall the open problems and give some directions for further research.

### 1.3 From dimensionality reduction techniques to NMF

The extraction of the underlying structure within a data set is a problem of major importance in data science. One of the first techniques is the LDR that consists in the transformation of a set of data points that belong into a high-dimensional space into a low-dimensional linear subspace so that the low-dimensional representation reveals the meaningful properties of the original data (the underlying structure). Each data point is represented as a linear combination of a small number of basis elements. Then, from a set of data points, we want to extract a basis and find the coordinates of the data points in this basis. Mathematically, given a set of  $N$  vectors  $v_n \in \mathbb{R}^F$  ( $1 \leq n \leq N$ ), LDR searches for  $K$  basis vectors  $w_k \in \mathbb{R}^F$  ( $1 \leq k \leq K$ ) such that each data point is well estimated by a linear combination of the basis vectors, that is,

$$v_n \approx \sum_{k=1}^K w_k h_{kn} \quad \text{for all } n,$$

where the scalar  $h_{kn}$  are the components of each data point expressed in the basis  $\langle w_1, \dots, w_k, \dots, w_K \rangle$ . By concatenating these approximations for all  $n$  under a matrix form, we can write:

$$V = [v_1 v_2 \dots v_N] \approx [w_1 w_2 \dots w_K] [h_1 h_2 \dots h_N]$$


Therefore one can clearly see that LDR is equivalent to Low-Rank Matrix Approximation (LRMA) of matrix  $V \in \mathbb{R}^{F \times N}$  by matrices  $W \in \mathbb{R}^{F \times K}$  and  $H \in \mathbb{R}^{K \times N}$  such that:

- each column of  $W$ , denoted  $w_k$  ( $1 \leq k \leq K$ ), is a basis vector,
- each column of  $H$ , denoted  $h_n$  ( $1 \leq n \leq N$ ), gives the coordinates of data point  $v_n$  in the basis  $W$ ,
- in the case  $W$  is full-column rank, that is,  $\text{rank}(W) = K$ ,  $K$  corresponds to the dimension of the vector subspace spanned by the columns of  $W$ ,

In the applications discussed in this thesis, the number of basis vectors is significantly smaller than the space dimension  $F$  and the number of data points  $N$ , that is,  $K \ll \min(F, N)$ . Another remark of major importance is that LRMA is a model, it means that we consider that our input data matrix is well approximated by the product of  $W$  and  $H$ ,

in other words we consider this linear approximation meaningful and representative of the reality. Let us put the emphasis that we live with models and all the models are arguably wrong but retains a part of exactitude.

The goal is then to compute the "best" matrices  $W$  and  $H$  for the approximation of the input data matrix  $V$ . To achieve this goal, we need to define an error measure that characterizes the level of accuracy of the approximation of  $V$  by  $WH$  w.r.t. a particular metric. The error measure, denoted  $D(V|WH)$ , concerns each entry  $V_{fn}$  of  $V$  approximated by  $[WH]_{fn}$ . We define a "local" scalar error, denoted  $d(V_{fn}|[WH]_{fn})$ , that is a function of the  $(f, n)$ -th entries of  $V$  and  $WH$  and in some specific cases, it is a function of the so-called residual matrix  $V - WH$ . Typically, we do not give a priori greater importance to the approximation's accuracy of an entry rather than another. Therefore, the global error measure usually boils down to the sum over indices  $f, n$  of all the local errors, that is,

$$D(V|WH) = \sum_f^F \sum_n^N d(V_{fn}|[WH]_{fn}).$$

By choosing a specific expression for the local scalar error  $d(V_{fn}|[WH]_{fn})$ , we choose a metric. Typically, this error measure  $D(V|WH)$  corresponds to an entrywise norm of the residual matrix  $V - WH$ . Let us cite the most popular one; the squared Euclidean distance  $d(V_{fn}|[WH]_{fn}) = ([V - WH]_{fn})^2$ . In this case, the entrywise norm of the residual matrix  $V - WH$  corresponds to the well-known squared Frobenius norm and LRMA is equivalent to principal component analysis (PCA) [76], which can be solved using the singular-value-decomposition (SVD)[67] of  $V$ , keeping the first  $K$  singular values and setting

$$\begin{aligned} W &= U\Sigma(:, 1 : K)^{1/2}, \\ H &= \Sigma(:, 1 : K)^{1/2}V^T, \end{aligned}$$

for instance. The metrics that we consider in this thesis are extensively discussed in Section 1.9.1.

LRMA models have gained more and more interest in the two last decades as data analysis and information extraction are at the center of the attention nowadays. Even if LRMA models are apparently simple, they are powerful tools since many high-dimensional data sets are well approximated by low-rank matrices [132]. Many variants of the LRMA models have been used recently. They mainly differ in two ways (i) the metric used, (ii) the different constraints imposed on  $W$  and  $H$ . For (i), the main reason for choosing a specific metric is linked to the noise statistic that we assume on the data. For example, minimizing the Frobenius norm implicitly assumes independent and identically distributed (i.i.d.) Gaussian noise on each entry of  $V$ , see Section 1.9.2 for more details. From the acquisition to the saving in memory of the data, noise can randomly occur at each stage of the acquisition process. It is important in practice to assume some reasonable probability density function for these random variables, the reason is mostly twofold; build a representative model of the reality and facilitate the computation of meaningful solutions. For (ii),

the constraints depend on the application. For some applications, we want that each data point is approximated at most by  $r$  basis vectors such as  $r < K$ , therefore each column of  $H$  should have at least  $K - r$  zeros. This LRMA variant is referred to as sparse dictionary learning. This model yields a more compact and easily interpretable decomposition.

Among all variants of LRMA, we are interested in NMF. NMF requires the factors  $W$  and  $H$  to be component-wise nonnegative. These constraints are denoted  $W \geq 0$  and  $H \geq 0$ . As mentioned above, NMF is not the only LDR or LRMA technique, we mentioned the principal component analysis (PCA), we can also cite independent component analysis, sparse PCA, low-rank matrix completion, to cite a few. Then the question arises: why focusing on NMF only? NMF is a rich and variate topic, it is at the intersection between various major disciplines: continuous optimization (convex and non-convex), linear algebra, signal processing, machine learning and data mining. Also, the nonnegative constraints allow us to interpret the factors  $W$  and  $H$  meaningfully, for example when they correspond to nonnegative physical quantities. Even if, at first glance, the nonnegativity requirement seems to be restrictive in terms of practical use, it is not: NMF has many applications. This is either due to the fact that, for many applications, the input data is physically nonnegative, or the mathematical modeling of the problem requires nonnegativity. As popular applications for NMF, we can cite the identification of topics in a set of documents, the identification of materials and their localization in hyperspectral images, the audio spectral unmixing, the detection of communities in large networks, the analysis of medical images, see [56, 61] and the references therein for others examples. For applications mentioned above, the input data is nonnegative. Further, one can show the connection between NMF and topics in mathematics and computer science such as the minimum biclique cover of a bipartite graph [46] and the nested polytopes problem [61] for which the modeling requires nonnegativity.

Now that the context is settled, let us give a formal definition of the standard NMF problem.

## 1.4 NMF: the standard problem and models definition

The standard NMF problem can be formulated as follows:

### Problem 1.4.1: Nonnegative Matrix Factorization

Given a nonnegative matrix  $V \in \mathbb{R}_+^{F \times N}$ , a factorization rank  $K$  and a metric  $d(x|y)$  between two scalars  $x$  and  $y$ , NMF aims to compute two nonnegative matrices  $W \in \mathbb{R}_+^{F \times K}$  and  $H \in \mathbb{R}_+^{K \times N}$  such that the error measure  $D(V|WH) = \sum_f^F \sum_n^N d(V_{fn}|[WH]_{fn})$  is minimized. NMF requires to solve:

$$\min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} D(V|WH) = \sum_{fn} d(V_{fn}|[WH]_{fn})$$

subject to  $H \geq 0, W \geq 0,$

where  $d(x|y)$  defined for all  $x, y \geq 0$  is such that:

- $d(x|y)$  is continuous over  $x$  and  $y$ ,
- $d(x|y) \geq 0$  for all  $x, y \geq 0$ ,
- $d(x|y) = 0$  if and only if  $x = y$ .

The choice of  $d(x|y)$  is crucial as it leads to different properties such as the differentiability and  $L$ -smoothness (the gradient is Lipschitz-continuous) of the error measure  $D(V|WH)$  on its active domain so that different optimization schemes are needed to tackle Problem 1.4.1; see Sections 1.9.1 and 1.10 for more details.

Let us now introduce the two linear models associated to problem 1.4.1: (i) the exact NMF model and (ii) the approximate NMF model. For (i), we are looking for nonnegative matrices  $W$  and  $H$  such that  $V = WH$ , major aspects of the exact NMF models are presented in Sections 1.6 to 1.8. For (ii), the exactness is not required and we are searching for an approximate decomposition, that is,  $V \approx WH$ . The reason is the presence of noise, and the linear model being in most cases only an approximate model. To sum up, we made in this thesis the distinction between the NMF model and the optimization problem that is associated to the model and that is solved to compute  $(W, H)$ . The two standard NMF models are:

$$V = WH \text{ such that } W \in \mathbb{R}_+^{F \times K}, H \in \mathbb{R}_+^{K \times N} \text{ with } K \ll \min(F, N), \quad (1.1)$$

$$V \approx WH \text{ such that } W \in \mathbb{R}_+^{F \times K}, H \in \mathbb{R}_+^{K \times N} \text{ with } K \ll \min(F, N). \quad (1.2)$$

The standard problem 1.4.1 is solved in order to compute the solutions  $(W, H)$  for the NMF model at hand.

The exact NMF model (1.1) is useful to compute the nonnegative rank of  $V$ , denoted  $\text{rank}_+(V)$ , which is defined as the smallest integer  $K$  such that an exact NMF of  $V$  exists,

mathematically, we write:

$$\text{rank}_+(V) = \min \left\{ K \in \mathbb{N} \mid V = \sum_{k=1}^K A_k, A_k \geq 0, \text{rank}_+(A_k) = 1 \text{ for all } k \right\}.$$

The computation of the nonnegative rank is NP-hard [135] and determining the value for  $K$  such that  $V$  has an exact nonnegative factorization is a research topic on its own, referred to as the Model Order Selection (MOS) problem. For the recent progress on computing the value of the nonnegative rank in NMF, see [56, 37] and the references therein. The MOS will come back further for the minimum-volume  $\beta$ -NMF problem we present in Chapter 3 that is able to perform automatic model order selection in the case we overestimate  $K$ .

In Section 1.4, we make the hypothesis that the factorization rank is given. In practice when we deal with real-life data, we do not know the "correct" value for this parameter. In other words, we do not know the embedding dimension of  $V$  in the column space of  $W$ . In the literature, it is stated that that this dimension is closely related to the nonnegative rank of  $V$  but from our point of view, this link is weak and circumstantial as:

- the nonnegative rank is a particular case of factorization rank, namely the smallest one such that an exact NMF exists.
- The nonnegative rank makes sense as soon as an exact NMF model is valid, that is, in the noiseless case.
- The nonnegative rank is a mathematical concept on its own as it is not only related to exact NMF, it has meaning for other topics such as the nested polytope problem and the decomposition of a bivariate probability matrix into a convex combination of independent bivariate probability matrices [31].
- The factorization rank for an approximate NMF model has meaning only in the paradigm of the application at hand. For instance, it could be referred to as the number of materials present within a scenery, the number of audio sources in an audio signal, the number of communities in a social network, etc.

Therefore, for this thesis and when it comes to deal with approximate NMF models,  $K$  is an input parameter that we choose based on prior information (from the literature for instance) and depending on the application. In Section 6, we propose new optimization problems associated to the exact NMF model but we consider that the nonnegative rank is known.

The following section presents the breakthrough experiment that puts NMF on stage as the most popular applications for NMF.

## 1.5 The breakthrough experiment for NMF

In 2000, Lee and Seung [87] wrote the seminal paper that really puts NMF on stage by showing its remarkable ability to automatically extract sparse and easily interpretable

factors. Lee and Seung showcased its ability with a breakthrough experiment in which NMF is applied for an input matrix  $V$  whose columns are vectorized images of human face, see Figure 1.1. Interestingly, the columns of matrix  $W$  obtained for NMF models correspond to vectorized images of constitutive parts of a human face, such as the eyes, nose and the ears, whereas PCA learns holistic representations. Lee and Seung showed an important feature of NMF: the nonnegativity constraints typically induce sparse factors, i.e., factors with relatively many zero entries. Formally, the reason for such behavior is that stationary points  $(W, H)$  of optimization Problem 1.4.1 associated to NMF models will typically belong to the boundary of the feasible domain (the nonnegative orthant) hence will contain zero components. This can be easily explained with the first-order optimality conditions; let us consider the following simplified optimization problem with nonnegativity constraints:

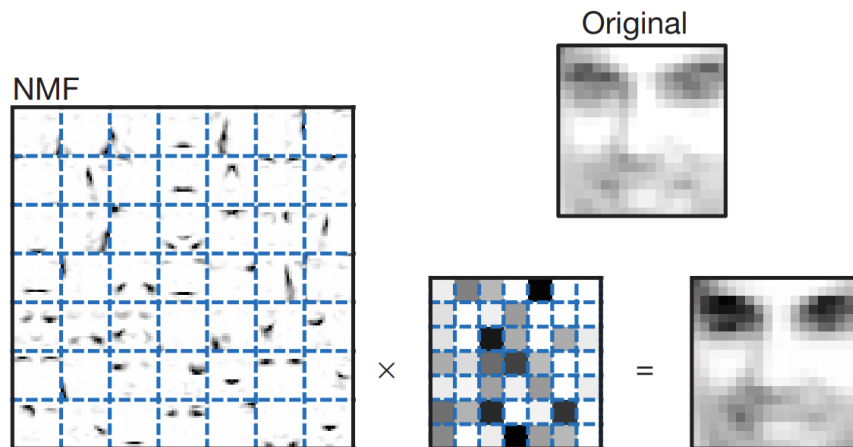
$$\min_{x \in \mathbb{R}_+^n} f(x), \quad (1.3)$$

By using the Karush–Kuhn–Tucker necessary condition, the set of stationary points for such a problem is  $\mathcal{D} = \{x \in \mathbb{R}^n | x \geq 0, \nabla f(x) \geq 0, x_i [\nabla f(x)]_i = 0 \text{ for } 1 \leq i \leq n\}$ . Hence some components of the solution can be expected to be equal to zero. Sparsity of the factors has many benefits as, in addition to reducing memory requirements to store the factors, it improves their interpretability. On the opposite, PCA do not naturally generate sparse factors.

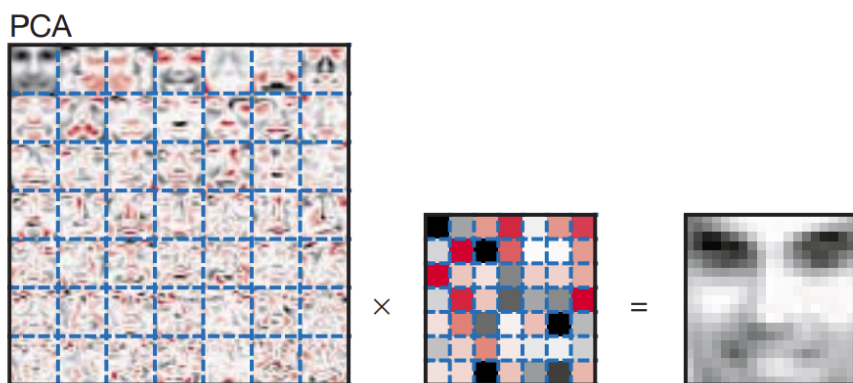
We can trace the first results showing the ability of NMF to extract interpretable information back in the 1990s in analytical chemistry for which researchers factorized spectral samples of chemical compounds and showed that the columns of matrix  $W$  correspond to the spectra of constituent elements of the chemical samples [112]. NMF also showed impressive interpretability in machine learning applications such the text mining for which NMF is able to identify the main topics contained in a set of documents (web site contents, journal papers, etc). The interpretability of NMF is closely related to its model uniqueness, or its ability to *identify* the ground-truth factors, denoted  $(W^\#, H^\#)$ , that generated the data  $V$  through the exact NMF model  $V = W^\# H^\#$  with  $W^\#, H^\# \geq 0$ . The connection between interpretability and identifiability is intuitively pleasing [50]. Indeed, in the case the data  $V$  really follows the generative models  $V = WH$ , then it is essential to find the ground-truth factors as they explain the data. The identifiability for NMF is discussed later in Section 1.8 as we first need to introduce many concepts related to the geometry of NMF (Section 1.6).

NMF is used in many others applications that we briefly list here-under:

- Community detection; based on a social graph, NMF is able to identify groups of people who have similar activities [100].
- Gene expression analysis; [24] the authors show that NMF is an efficient method for identification of distinct molecular patterns and provides a powerful method for class



(a) Results with NMF



(b) Results with PCA

**Fig. 1.1.** The breakthrough experiment that puts NMF on stage. Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas principal components analysis (PCA) learn holistic representations. Figure reproduced from [87].



discovery. They demonstrate in particular the ability of NMF to recover meaningful biological information from cancer-related microarray data.

- Identification of hidden Markov models [82].
- Prediction of epileptic seizures using electroencephalographic (EEG) signals [127].
- Recommender systems popularized by the "Netflix prize competition"; the goal is to predict the preferences of users for some items based on the preferences or taste information from many users. In the case of the Netflix problem, it boils down to predict how much someone is going to like a movie based on her/his movie ratings and the ratings of others.

This list is far from being exhaustive and we refer our readers to reference [56] for more examples for which NMF has proved to be a powerful tool to understand the underlying structure of data sets. In this thesis, we focus on two major applications for NMF:

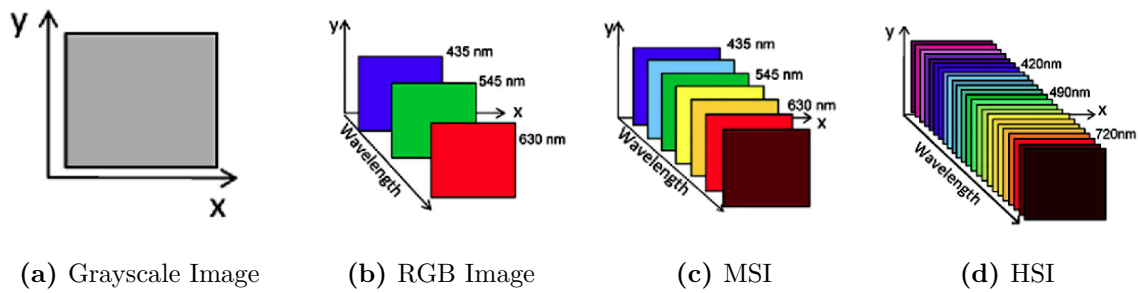
- hyperspectral unmixing (Section 1.5.1),
- audio single-channel blind source separation (Section 1.5.2).

### 1.5.1 Hyperspectral imaging

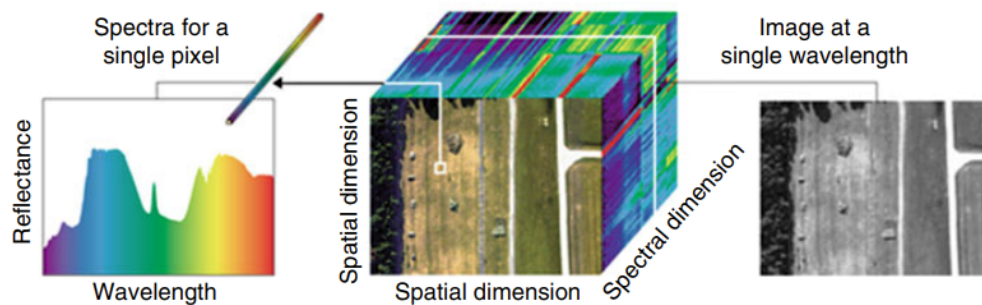
A grayscale image is an image in which the value of each pixel is a single sample representing only an amount of light; that is, it carries only intensity information. The intensity of a pixel is expressed within a given range between a minimum and a maximum value. Each pixel can be represented as a 1-dimensional vector. An RGB image is such that each of its pixel is coded over three components, the three components give the intensity of reflected light for the wavelengths corresponding to the visible red, green and blue.

An RGB image is based on additive color model in which red, green, and blue light are added together in various ways to reproduce a broad array of colors in such a way it can be perceived by an human eye. A hyperspectral image (HSI) is an image that contains information over a wide spectrum of light instead of just assigning primary colors (red, green, blue) to each pixel. The light striking each pixel is broken down into many different spectral bands; each pixel has typically between 100 and 200 components, corresponding to the reflectance (fraction of light reflected by that pixel) at many different wavelengths. In general, the spectral range of airborne hyperspectral sensors is 380–12.700 nm and for satellite sensors is 400–1.400 nm. The AVIRIS airborne hyperspectral imaging sensor, for instance, obtains spectral data over 224 continuous channels, each with a bandwidth of 10 nm over a spectral range from 400 to 2.500 nm.

The number of wavelengths measured depend on the cameras sensor used and are usually chosen depending on the application considered. The advantage of hyperspectral images is that they provide more information on what is imaged, some of it blind to the human



**Fig. 1.2.** Cube data for different classes of images: a grayscale image (Left), a RGB image (Left-Middle), a mutlispectral image (Right-Middle) and a hyperspectral image (Right). Axis  $x$  and  $y$  designate the two spatial coordinates.



**Fig. 1.3.** Hyperspectral data cube containing all spatial and spectral data for each pixel (figure extracted from [47])

eye as many wavelengths belong to the invisible light spectrum. This additional information allows one to identify and characterize the materials present in a scenery. Prior to hyperspectral images, we also have multi-spectral images which have fewer bands as compared to hyperspectral images; each pixel has typically between 4 and 10 components. In summary, an image acquired from any sensor will be in the format of a data cube; a grayscale image would then correspond to a slice of a cube. For comprehension purposes, Figure 1.2 presents side by side the cube data format for a grayscale image, an RGB image, a seven band MSI and an HSI. Figure 1.3 shows a Hyperspectral data cube that contains both spatial and spectral information from materials within a given scenery. Each pixel across a sequence of continuous, narrow spectral bands contains both spatial and spectral properties. Pixels are sampled across many narrowband images by a scanning system at a particular spatial location, resulting in a “hyperspectral data cube”.

### Blind Hyperspectral Unmixing with NMF

The algorithms and the image processing methodologies associated with HSI are a product of military research, and were primarily used to identify targets and other objects against background clutter. Now it has many civil applications, and has particularly been useful in

satellite technology, we can cite: agriculture, mineralogy, astronomy, chemical imaging. It is also a efficient tool for the assessment of tissue conditions at diagnosis and during surgery [120]. The main aim is to extract physical information from raw data collected across the spectrum, which can be easily converted to describe inherent properties of surface targets. In this thesis, we would like to extract the constitutive material present in the image, called endmembers (e.g., grass, trees, road surfaces, roof tops) and determine the abundances of the endmembers in each pixel, that is, identify which pixels contain which materials and in which quantity. To achieve this goal with NMF-based models, we need (i) to transform the hyperspectral cube data of a given scene into an input nonnegative matrix  $V$ , and (ii) assume a model for the mixing process. For (i), given a HSI with  $F$  wavelengths and a  $p \times q$  spatial resolution, let us pose  $N = p \times q$  and construct the matrix  $V \in \mathbb{R}_+^{F \times N}$  such that  $V(k, n)$  corresponds to the reflectance of  $n$ th pixel at the  $f$ th wavelength. Each column of  $V$  corresponds to the spectral signature of one particular pixel while each row corresponds to a vectorized image at a given wavelength. In practice, let us remark that when we do not have at hand a dictionary of spectral signatures for the endmembers, this problem is referred to as the Blind Hyperspectral Unmixing (BHU).

For (ii), this follows the fact that the resolution of most hyperspectral images is low, and hence most pixels potentially contain several materials. Here comes the necessity to consider a model for the mixing of the recorded reflectance. The simplest and most popular model is the linear mixing model which assumes that the spectral signature of a pixel is a linear combination of the spectral signatures of its constitutive endmembers. For such a model, the weights are given by the abundances. For instance, if a pixel contains 60 % of grass and 40% of dirt, then its spectral signature is equal to 0.6 times the spectral signature of the grass plus 0.4 times the spectral signature of the dirt.

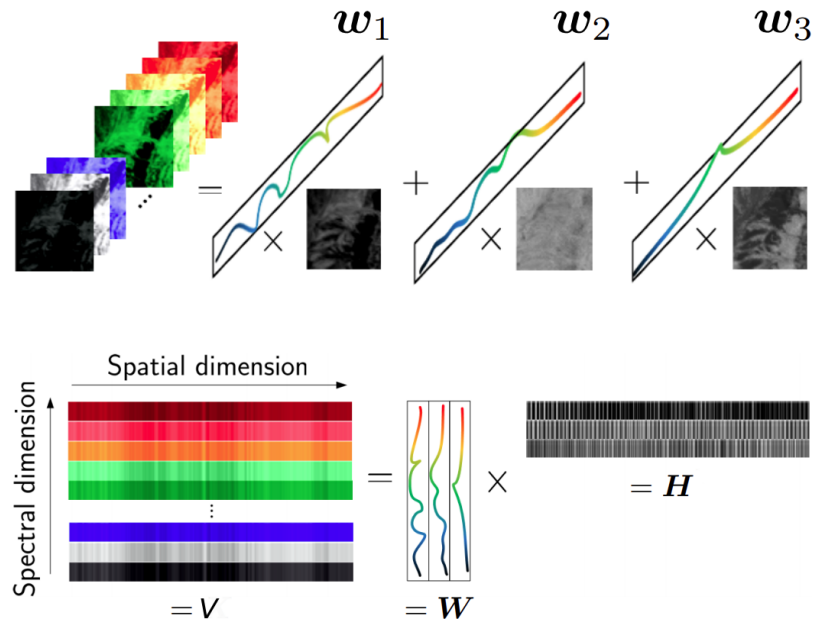
In practice, the equality is not required as there are many sources of imperfections such as the noise, the reflective surfaces arbitrary complex that induce numerous and different modes of reflection in a small area, the mirage effects and the atmospheric absorption. This model is only approximate. Finally if we assume that the combination for the linear combination are nonnegative and that the image contains a limited number  $K$  of endmembers, then the linear mixture model is mathematically equivalent to approximate NMF models, indeed, for all  $n$ :

$$V(:, n) \approx \sum_{k=1}^K W(:, k) h_{kn} = WH(:, n),$$

where  $h_{kn} \geq 0$  is the abundance of the  $k$ th material in the  $n$ th pixel. Figure 1.4 illustrates the linear mixing model for a hyperspectral data cube and the corresponding NMF model after rearranging the pixels in the case there are no imperfections.

Important remarks from [55] are reported here-under:

- using a standard algorithm such as the ones detailed in Section 1.10 will in general not lead to the sought decomposition. The reason is related to non-uniqueness for the solutions of NMF as explained in Section 1.8.1.



**Fig. 1.4.** (Top) The linear mixture model for blind spectral unmixing with the slices of the hyperspectral data cube on left hand side; on the right hand side we have the vectors  $w_k$  which contains the spectral signature of a pure material, or endmember and the abundance maps of the endmembers, which are re-arranged columns of  $H$ . (Bottom) The corresponding NMF model after re-arranging the pixels (figure extracted from [50])

- In practice, meaningful solutions can be obtained by adding constraints to NMF models such a sum-to-one constraint for the columns of the abundance matrix or sparsity constraints. We refer the reader to references [30] for the discussion on various constraints useful for blind hyperspectral applications based on NMF.

### 1.5.2 Single-channel blind source separation

Blind audio source separation concerns the techniques used to extract unknown signals called sources from a mixed audio signal  $x$ . In this thesis, we assume that the audio signal is recorded with a single microphone. Considering a mixed signal composed of various audio sources, the blind audio source separation consists in isolating and extracting each of the sources on the basis of the single recording. Usually, the only known information is the number of estimated sources present in the mixed signal. The blind source separation problem is said to be underdetermined as there are fewer sensors (only one in our case) than sources. It then appears necessary to find additional information to make the problem well posed. The most common technique used for this kind of problem is to get some form of redundancy in the mixed signal in order to make it overdetermined. This is typically done by computing the spectrogram which represents the signal in the time and frequency domains simultaneously (splitting the signals into overlapping time frames). The computation of spectrograms can be summarized as follows: short time segments are

extracted from the signal and multiplied element wise by a window function or “smoothing” window of size  $F$ . Successive windows overlap by a fraction of their length, which is usually taken as 50%. On each of these segments, a discrete Fourier transform is computed and stacked column-by-column in a matrix  $X$ . Thus, from a one-dimensional signal  $x \in \mathbb{R}^T$ , we obtain a complex matrix  $X \in \mathbb{C}^{F \times N}$  called spectrogram where  $F \times N \simeq 2T$  (due to the 50% overlap between windows). Note that the length of the window determines the shape of the spectrogram. These preliminary operations correspond to computing the short time Fourier transform (STFT), which is given by the following formula: for  $1 \leq f \leq F$  and  $1 \leq n \leq N$ ,  $X_{f,n} = \sum_{j=0}^{F-1} w_j x_{nL+j} e^{-i \frac{2\pi f j}{F}}$ , where  $w \in \mathbb{R}^F$  is the smoothing window of size  $F$ ,  $L$  is a shift parameter (also called hop size), and  $H = F - L$  is the overlap parameter. The number of rows corresponds to the frequency resolution. Letting  $f_s$  be the sampling rate of the audio signal, consecutive rows correspond to frequency bands that are  $f_s/F$  Hz apart. The STFT process is pictured on the left side of Figure 1.5.

The time-frequency representation of a signal highlights two of its fundamental properties: sparsity and redundancy. Sparsity comes from the fact that most real signals are not active at all frequencies at all time points. Redundancy comes from the fact that frequency patterns of the sources repeat over time. Mathematically, this means that the spectrogram is a low-rank matrix. This has been validated on many experiments and makes sense physically. Indeed, let us take the example of the spectrogram of an instrument which is a finite sum of the spectrograms of the each note. Usually the number of notes is small w.r.t. the dimensions of the spectrograms.

These two fundamental properties led sound source separation techniques to integrate algorithms such as nonnegative matrix factorization (NMF). Such techniques retrieve sensible solutions even for single-channel signals.

### Mixing assumptions

Given  $K$  source signals  $s^{(k)} \in \mathbb{R}^T$  for  $1 \leq k \leq K$ , we assume the acquisition process is well modelled by a linear instantaneous mixing model:

$$x(t) = \sum_{k=1}^K s^{(k)}(t) \quad \text{with } t = 0, \dots, T-1. \quad (1.4)$$

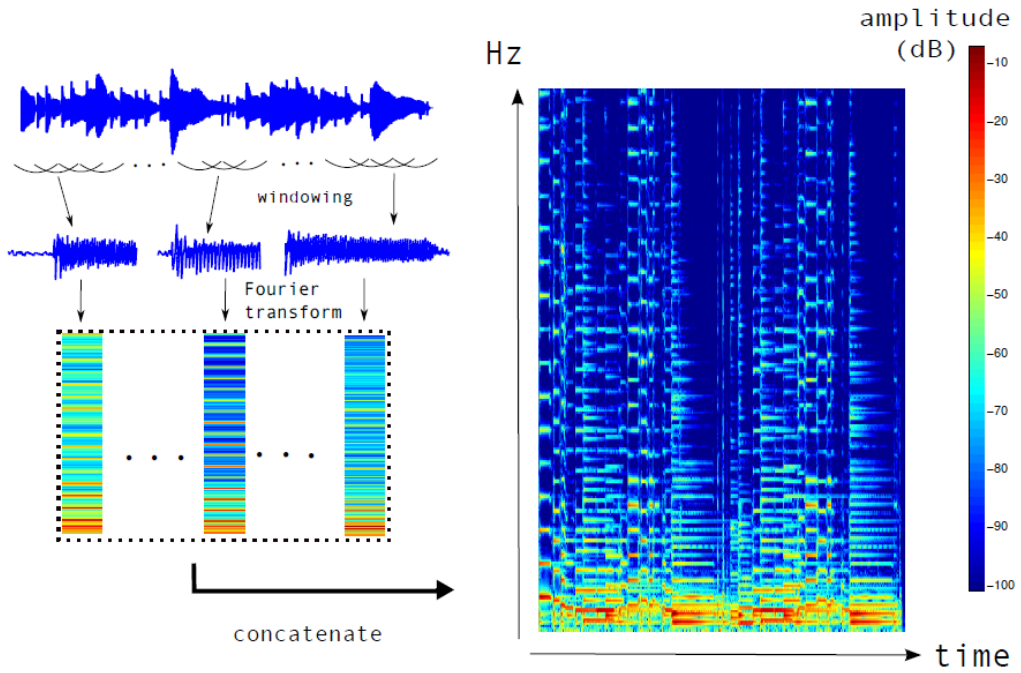
Therefore, for each time index  $t$ , the mixed signal  $x(t)$  from a single microphone is the sum of the  $K$  source signals. It is standard to assume that microphones are linear as long as the recorded signals are not too loud. If signals are too loud, they are usually clipped. The mixing process is modelled as instantaneous as opposed to convolutive used to take into account sound effects such as reverberation. The source separation problem consist in finding source estimates  $\hat{s}^{(k)}$  of  $s^{(k)}$  sources for all  $k \in \{1, \dots, K\}$ . Let us denote  $S$  the linear STFT operator, and let  $S^\dagger$  be its conjugate transpose. We have  $S^\dagger S = FI$ , where  $I$  is the identity matrix of appropriate dimension. For the remainder of this thesis,  $S^\dagger$  stands for the inverse short time Fourier transform. Note that the term inverse is not meant in a mathematical sense. Indeed the STFT is not a surjective transformation from

$\mathbb{R}^T$  to  $\mathbb{C}^{F \times N}$ . In other words, each spectrogram or each matrix with complex entries is not necessarily the STFT of a real signal; see [88] and [99] for more details. By applying the STFT operator  $S$  to (1.4), we obtain the mixing model in the time-frequency domain :

$$X = S(x(t)) = S\left(\sum_{k=1}^K s^{(k)}(t)\right) = \sum_{k=1}^K S^{(k)},$$

where  $S^{(k)}$  is the STFT of the source  $k$ , that is, the spectrogram of source  $k$ .

To identify the sources, we use in this thesis the amplitude spectrogram  $V = |X| \in \mathbb{R}_+^{F \times N}$  or the power spectrogram respectively defined as  $V_{fn} = |X_{fn}|$  and  $V_{fn} = |X_{fn}|^{(2)}$  for all  $f, n$ . We assume that  $V = \sum_{k=1}^K |S^{(k)}|$ , which means that there is no sound cancellation between the sources, which is usually the case in most signals. Figure 1.5 summarizes the process to compute the amplitude spectrogram of an input signal:



**Fig. 1.5.** Process for generating an amplitude spectrogram based on an input audio signal.

We can observe two fundamental properties of a standard amplitude spectrogram for using NMF: sparsity and low-rank structure. Figure extracted from [88].

Finally, we assume that the source spectrograms  $|S^{(k)}|$  are well approximated by non-negative rank-one matrices. Note that a source can be made of several rank-one factors in which case a post-processing step will have to recombine them a posteriori (e.g., looking at the correspondence in the activation of the sources over time). Note also that we focus in this thesis on the NMF stage of the source separation which factorizes  $V$  into the source spectrograms. For the phases reconstruction, which is a highly non-trivial problem, we consider a naive reconstruction procedure consisting in keeping the same phase as the input mixture for each source [88].

### NMF for audio source separation

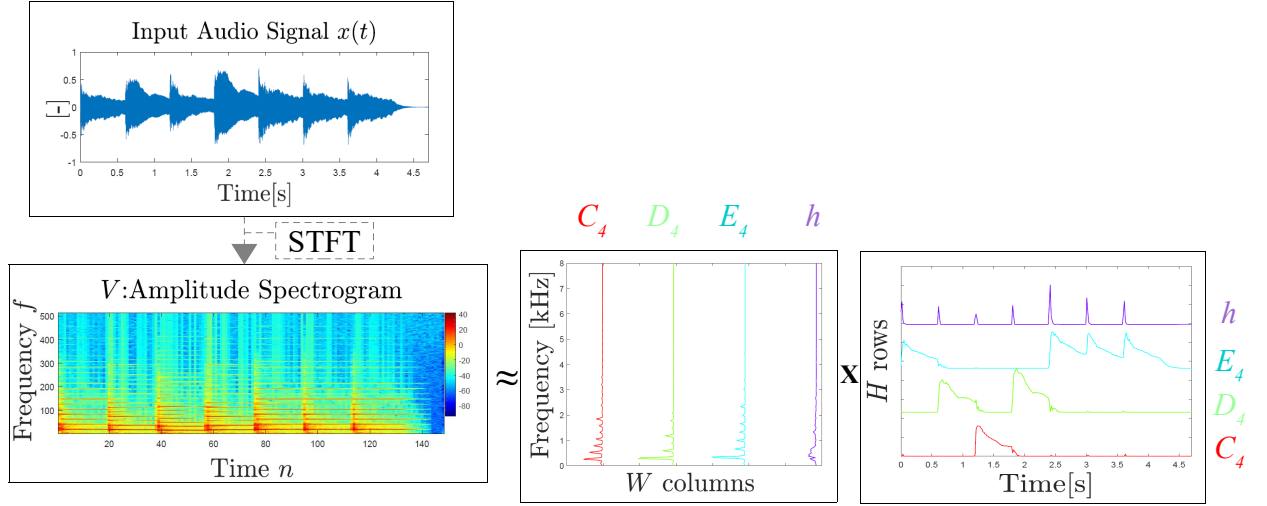
Under the assumptions (i) the amplitude spectrogram of the mixture  $V$  is a nonnegative linear combination of the spectrogram of the sources (no source cancelation) and (ii) the spectrograms have a low-rank structure, blind source separation is another NMF problem; given a non-negative matrix  $V \in \mathbb{R}_+^{F \times N}$  (the spectrogram) and a positive integer  $K \ll \min(F, N)$  (the number of sources, called the factorization rank), we have seen that NMF aims to compute two non-negative matrices  $W$  with  $K$  columns and  $H$  with  $K$  rows such that  $V \approx WH$ . When the matrix  $V$  corresponds to the amplitude spectrogram or the power spectrogram of an audio signal, we have that:

- $W$  is referred as the dictionary matrix and each column corresponds to the spectral content of a source, and,
- $H$  is the activation matrix specifying if a source is active at a certain time frame and in which intensity.

In other words, each rank-one factor  $W(:, k)H(k, :)$  will correspond to a source: the  $k$ th column  $W(:, k)$  of  $W$  is the spectral content of source  $k$ , and the  $k$ th row  $H(k, :)$  of  $H$  is its activation over time. To compute  $W$  and  $H$ , we need to solve Problem 1.4.1, we discuss in Section 1.9.1 the most appropriate choices for the metrics  $d(V_{fn} | [WH]_{fn})$  in the frame of audio source separation. Let us illustrate what would be the outputs of the NMF applied to a simple monophonic signal; namely a piano recording of “Mary had a little lamb” whose consists in a sequence of three notes as follows:  $E_4, D_4, C_4, D_4, E_4, E_4, E_4$ . For this data set, the three main sources are the piano notes whose spectrograms have rank-one: each column of  $W$  is the spectral content of each note while the entries of  $H$  indicate when a note is active. Figure 1.6 displays the input audio signal  $x(t)$ , the input matrix  $V$  (the amplitude spectrogram in this case) and the results for  $W$  and  $H$ . As we can observe, we identify three notes and a fourth note that corresponds to the hammer within the piano (a common trigger to each note) and the sequence of the activations is consistent with the sequence played. NMF is therefore able to blindly separate the different sources and identify which source is active at which moment in time. This ability led NMF to be used for automatic music transcription [16].

For comprehensive purposes, we summarize the full process of blind source separation with NMF-based methods considered in this thesis. The process consists in four steps detailed here-under:

1. Step 1: an input vector  $x(t)$  is loaded.
2. Step 2: computation of the STFT  $X = S(x(t))$  and the phase matrix  $\Phi$  such that  $X = |X| \odot \exp j\Phi$  where  $\Phi_{fn} = \tan^{-1} \left( \frac{\Im(X_{fn})}{\Re(X_{fn})} \right)$  and  $\exp$  is the component-wise exponential. In this illustration, we pose  $V = |X|$  (the amplitude spectrogram).
3. Step 3: Computation of a NMF  $(W, H)$  for the input matrix  $V$ .



**Fig. 1.6.** Application of NMF to blind audio source separation: decomposition of the piano recording "Mary had a little lamb" using NMF: (top-left) the input audio signal  $x(t)$ , (bottom-left) amplitude spectrogram  $V$  in dB, (middle-bottom) basis matrix  $W$  corresponding to the three notes  $C_4$ ,  $D_4$ ,  $E_4$  and  $h$  (the hammer noise), (bottom right) activation matrix  $H$  that indicates when each note is active.

4. Step 4: computation of the source estimates, denoted  $\hat{S}^k$ , based on the Wiener Filtering (or "Masking coefficients") as follows:  $\hat{s}^{(k)} = S^\dagger \left( \frac{[W(:,k)H(k,:)]}{[WH]} \odot V \odot \exp j\Phi \right)$ . This reconstruction technique for the sources estimates enables to reconstruct the input signal  $x(t)$ , indeed;

$$\begin{aligned}
 \sum_k \hat{s}^{(k)} &= \sum_k S^\dagger \left( \frac{[W(:,k)H(k,:)]}{[WH]} \odot V \odot \exp j\Phi \right) \\
 &= S^\dagger \left( \left( \sum_k \frac{[W(:,k)H(k,:)]}{[WH]} \right) \odot V \odot \exp j\Phi \right) \\
 &= S^\dagger(X) = x(t).
 \end{aligned}$$

NMF models have shown their potential to identify audio sources from more complex signals, such as polyphonic music where several notes and even several instruments are played at once. However, for such audio signals, it is advised to consider more elaborate NMF models (and associated optimization problems) such as sparse NMF [137] and more advanced mixture models such as convolutive NMF [125].

## 1.6 Geometry of NMF

*"Let none ignorant of geometry enter here"*

- Attributed to Plato, phrase over the entrance of his Academy [2]



### Motivations

In this section, we present two geometrical interpretations of NMF. These geometrical interpretation will be useful:

- to understand one of the main issues about NMF models; the NMF models are nonunique in general. In Section 1.8, this issue will be discussed and we will introduce the notion of identifiability of an NMF model with its associated optimization problem.
- to understand the main motivation for the NMF problems introduced in Chapters 2 and 3.

NMF has a nice underlying geometrical interpretation which is a key aspect. The understanding of the NMF geometry enables to formulate meaningful problems and design powerful algorithms. The geometrical interpretation is not recent and takes its origin in the fields of geoscience and remote sensing, see [56] and the references therein for a detailed historical review. In this section we describe the geometric interpretation of the exact NMF model. We will develop these aspects w.r.t. the so-called nested cones and nested convex hulls (or nested polytopes), the main reason is the ease to visualize it in the case we deal with nested convex hulls. Note that most of the geometric aspects developed here-under for exact NMF models are also useful for the approximate NMF models.

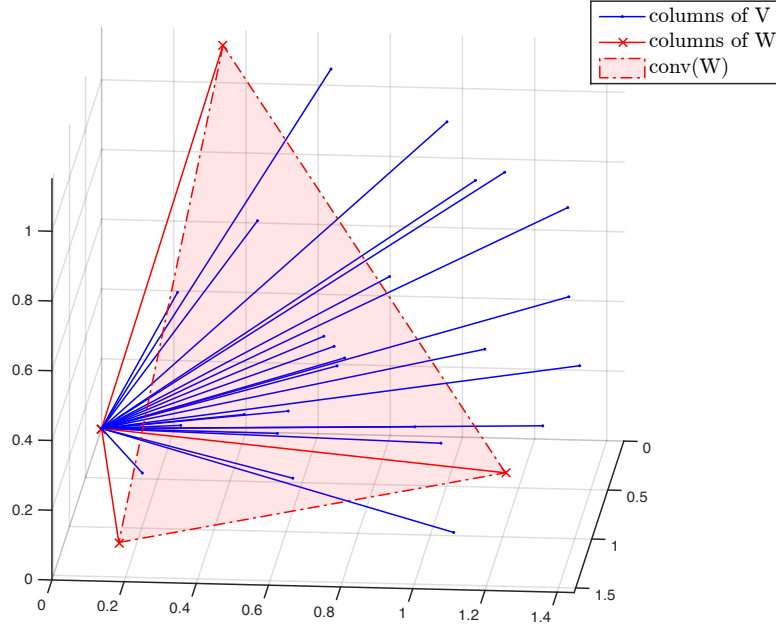
Let us first recall the mathematical definitions of a convex cone spanned by the columns of a matrix  $U$ . Given a matrix  $U$ : the convex cone spanned by the columns of  $U \in \mathbb{R}^{F \times K}$ , denoted  $\text{cone}(U)$ , is defined as follows:

$$\text{cone}(U) = \{x \in \mathbb{R}^F | x = U\theta, \theta \in \mathbb{R}_+^K\}.$$

The elements of  $\text{cone}(U)$  are conic combinations of the columns of  $U$ , that is, linear combinations with nonnegative weights. The dimension of  $\text{cone}(U)$  is the dimension of the linear subspace spanned by the columns of  $U$  and is therefore equal to the  $\text{rank}(U)$ . Now, we recall the exact NMF model  $V = WH$  where  $V \in \mathbb{R}_+^{F \times N}$ ,  $W \in \mathbb{R}_+^{F \times K}$ ,  $H \in \mathbb{R}_+^{K \times N}$ , that is each column of  $V$  is a nonnegative linear combination of the columns of  $W$  weighted by the components of the corresponding column of  $H$ . Therefore, one can easily see that the columns of  $V$  are contained in the convex cone generated by the columns of  $W$ . Mathematically  $V(:, n) = WH(:, n)$  and since  $W, H \geq 0$  we have  $V(:, n) \in \text{cone}(W) \subseteq \mathbb{R}_+^F$  for all  $n$ . Since this inclusion holds for all  $n$ , we can write the following chain of inclusions:

$$\text{cone}(V) \subseteq \text{cone}(W) \subseteq \mathbb{R}_+^F. \quad (1.5)$$

Based on equation (1.5), we can see why finding factors  $W \geq 0$  and  $H \geq 0$  such that  $V = WH$  is equivalent to find  $\text{cone}(W)$  nested between two cones:  $\text{cone}(V) \subseteq \mathbb{R}_+^F$ . In other words, we are looking for a set of  $K$  vectors within  $\mathbb{R}_+^F$  such that their convex cone contains the given cone spanned by the columns of the input data matrix  $V$ . Equation



**Fig. 1.7.** Geometric illustration of Exact NMF for  $K = F = 3$  and  $N = 25$ . We observe that  $\text{cone}(V) \subseteq \text{cone}(W) \subseteq \mathbb{R}_+^F$ .

(1.5) also gives a lower bound for the dimension of the linear subspace spanned by the columns of  $W$ , that is  $\text{rank}(W)$  has a lower bound equal to  $\text{rank}(V)$ . Hence, by definition of the rank of a matrix, we can build up the following interval for  $\text{rank}(W)$ :

$$\text{rank}(V) \leq \text{rank}(W) \leq \min(F, K) = K. \quad (1.6)$$

since  $K \ll \min(F, N)$  by hypothesis. Figure 1.7 provides such a geometric interpretation for  $K = F = 3$  and  $N = 25$ .

Another equivalent geometric interpretation of the exact NMF is reformulated by using the nested convex hulls, also referred to as nested polytopes. The reason is twofold (i) it is easier to visualize the problem and (ii) the intrinsic  $\ell_1$  normalizations for such representation will be used in the further chapters. Note that both geometric interpretations will be useful for Section 1.8.1 that deals with the uniqueness of NMF solutions. Let us first recall some basic useful notions from convex geometry.

**Definition 1.6.1.** A set  $S \subseteq \mathbb{R}^F$  is convex if for all  $a, b \in S$ , we have  $\lambda a + (1 - \lambda)b \in S$  for all  $\lambda \in [0, 1]$ .

**Definition 1.6.2.** A point  $v \in \mathbb{R}^F$  is a convex combination of vectors  $w_1, \dots, w_k, \dots, w_K \in \mathbb{R}^F$  if for some real nonnegative numbers  $\alpha_k$  which satisfy  $\sum_k^K \alpha_k = 1$  and  $\alpha_k \geq 0$  ( $1 \leq k \leq K$ ), we have  $v = \sum_k \alpha_k w_k$ .

**Definition 1.6.3.** The convex hull of a set  $S$  is the smallest convex set containing  $S$  or, equivalently, the set of convex combinations of points in  $S$ .

We define the convex hull spanned by the columns of a matrix  $U \in \mathbb{R}^{F \times K}$ :

$$\text{conv}(U) = \{x \in \mathbb{R}^F \mid x = U\theta, \theta \in \mathbb{R}_+^K, e^T \theta = 1\}, \quad (1.7)$$

where  $e$  is a all-one column vector of appropriate size. Therefore, each element of  $\text{conv}(U)$  is a convex combination of the columns of  $U$ . The vertices of the set  $\text{conv}(U)$  are the columns  $U(:, k)$  of  $U$ , in other words, no column  $U(:, k)$  of  $U$  can be expressed as a convex combination of the remaining columns of  $U$ . Figure 1.5 shows the geometric illustration of the convex hull of a matrix  $W$ . For the following, let us also define the unit simplex:

**Definition 1.6.4.** *The unit simplex in dimension  $F$ , denoted  $\Delta^F$ , is the subset of  $\mathbb{R}^F$  defined as:*

$$\Delta^F = \{x \in \mathbb{R}^F \mid x \geq 0, e^T x = 1\}$$

Based on expression (1.7),  $\Delta^F$  is equivalent to  $\text{conv}(I_F)$  where  $I_F$  is the identity matrix of dimension  $F \times F$ .

We have now everything in hand to introduce the geometric interpretation of Exact NMF in terms of nested convex hulls. Given an exact NMF  $V = WH$ , we can assume without loss of generality that the columns of  $V$  and  $W$  have a  $\ell_1$  norm equal to one, that is  $\|V(:, n)\|_1 = \sum_f |V_{fn}| = 1$  for all  $n$  and  $\|W(:, k)\|_1 = \sum_f |W_{fk}| = 1$  for all  $k$ . Indeed, for any exact NMF for which matrices  $V$  and  $W$  do not contain columns equal to the null vector, we have the equivalency:

$$V = WH \iff VQ_V = WHQ_V, \quad (1.8)$$

where matrix  $Q_V$  is the diagonal matrix whose entries are such that  $Q_V(i, j) = \delta_{ij} \frac{1}{\|V(:, i)\|_1}$  with  $\delta_{ij}$  the Kronecker delta. Further, note that for any invertible, diagonal and nonnegative  $Q_W$  matrix whose entries are such that  $Q_W(i, j) = \delta_{ij} \frac{1}{\|W(:, i)\|_1}$ , we have that:

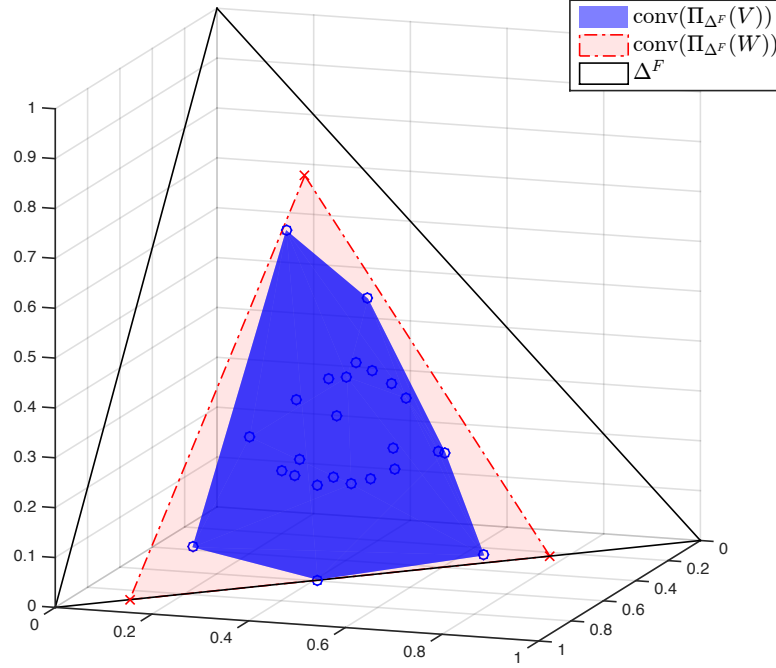
$$WH = (WQ_W)(Q_W^{-1}H), \quad (1.9)$$

where  $Q_W^{-1}(i, j) = \delta_{ij} \|W(:, i)\|_1$ . We finally insert (1.9) into (1.8) we get:

$$\begin{aligned} V = WH &\iff \underbrace{VQ_V}_{\Pi_{\Delta^F}(V)} = \underbrace{(WQ_W)}_{\Pi_{\Delta^F}(W)} \underbrace{(Q_W^{-1}H)}_{H'} \\ &\iff \Pi_{\Delta^F}(V) = \Pi_{\Delta^F}(W)H', \end{aligned} \quad (1.10)$$

where  $\Pi_{\Delta^F}(\cdot)$  denotes the projection operator of the columns of the input matrix onto the unit simplex in dimension  $F$ , hence the columns of the the resulting matrix are included in the subset  $\Delta^F$ . Let us first observe that  $H' \geq 0$  since each entry of  $Q_W^{-1}$  is nonnegative. Further, based on equation (1.10), we observe that the columns of matrix  $H'$  have unit  $\ell_1$  norm. Indeed by definition the columns of  $\Pi_{\Delta^F}(V)$  and  $\Pi_{\Delta^F}(W)$  have  $\ell_1$  norm (as they belong to the unit simplex), mathematically we have  $e^T = e^T \Pi_{\Delta^F}(V)$  and  $e^T = e^T \Pi_{\Delta^F}(W)$ . Therefore we can write the following chain:

$$e^T = e^T (\Pi_{\Delta^F}(V)) = e^T (\Pi_{\Delta^F}(W)H') = e^T H',$$



**Fig. 1.8.** Geometric illustration of Exact NMF w.r.t. the nested convex hulls on the data set from Figure 1.7. We observe that  $\text{conv}(\Pi_{\Delta^F}(V)) \subseteq \text{conv}(\Pi_{\Delta^F}(W)) \subseteq \Delta^F$ .

and hence  $e^T = e^T H'$ , that is, the entries of each column of  $H'$  sum to one.

In the case the columns of  $\Pi_{\Delta^F}(V)$  and  $\Pi_{\Delta^F}(W)$  are not equal to the null vector, one can easily see that the columns of  $\Pi_{\Delta^F}(V)$  are contained in the convex hull spanned by the columns of  $\Pi_{\Delta^F}(W)$ . Mathematically  $\Pi_{\Delta^F}(V)(:, n) = \Pi_{\Delta^F}(W)H'(:, n)$  and since  $H'(:, n) \geq 0$  and  $e^T H'(:, n) = 1$  for all  $n$ , we have  $\Pi_{\Delta^F}(V)(:, n) \in \text{conv}(\Pi_{\Delta^F}(W)) \subseteq \Delta^F$  for all  $n$ . Since this inclusion holds for all  $n$ , we have the following chain of inclusions:

$$\text{conv}(\Pi_{\Delta^F}(V)) \subseteq \text{conv}(\Pi_{\Delta^F}(W)) \subseteq \Delta^F. \quad (1.11)$$

Then, we see that computing an exact NMF  $V = WH$  with a factorization rank  $K$  is equivalent to finding a polytope,  $\text{conv}(\Pi_{\Delta^F}(W))$ , nested between two given polytopes,  $\text{conv}(\Pi_{\Delta^F}(V))$  and the unit simplex  $\Delta^F$ . Alternatively, it is equivalent to finding a set of  $K$  vectors (the columns of  $\Pi_{\Delta^F}(W)$ ) within  $\Delta^F$  such that their convex hull contains the columns of  $\Pi_{\Delta^F}(V)$ , in the case none of the columns of  $\Pi_{\Delta^F}(V)$  or  $\Pi_{\Delta^F}(W)$  equals the null vector. Figure 1.8 provides such a geometric interpretation for the data set used in Figure 1.7.

In summary, we have just showed that, without loss of generality, the exact NMF model has two equivalent geometric interpretations: a first one based on the inclusion of cones (1.5) and the second one based on the inclusions of convex hulls (1.11). In [56], the author shows that the geometric interpretation based on nested convex hulls is still valid in a lower dimensional linear subspace. The rationale begins by observing that a vector  $v \in \Delta^F$

contains redundant information as any single entry can be deduced from the others since  $\sum_f^F v_f = 1$  and then dimension of  $\Delta^F = F - 1$ . It is then possible to reduce the dimension of the problem by one and consider it in a lower dimensional subspace of dimension  $F - 1$  [56]. The key ingredients are:

1. the introduction of the subset  $\mathcal{S}^K = \{x \in \mathbb{R}^K | x \geq 0, e^T x \leq 1\}$  which is the convex hull of the unit simplex and the origin, that is  $\mathcal{S}^K = \text{conv}([I_K 0])$ ,
2. the abandon of the last coordinate of the matrices of interests without loss of generality. So for all matrices  $A \in \Delta^F$ , we use the reduced matrices  $\bar{A}(1 : F - 1, :)$  such that each column of  $\bar{A} \in \mathcal{S}^{F-1}$ .

Based on these ingredients, one can show that inclusions from (1.11) is equivalent to:

$$\text{conv}(\bar{\Pi}_{\Delta^F}(V)) \subseteq \text{conv}(\bar{\Pi}_{\Delta^F}(W)) \subseteq \mathcal{S}^{F-1} \quad (1.12)$$

Finally, based on the examples presented in Figures 1.7 and 1.8, three important remarks from [56] concerning the dimension of  $\text{conv}(\Pi_{\Delta^F}(V))$  and  $\text{conv}(\Pi_{\Delta^F}(W))$  must be done:

- in Figure 1.8, we see that dimension of  $\text{conv}(\Pi_{\Delta^F}(V))$  is equal to 2 which is also the dimension of  $\Delta^F$ . We must insist on the fact that the dimension of  $\text{conv}(\Pi_{\Delta^F}(V))$  is usually much smaller than  $F - 1$ , and therefore  $\text{rank}(V)$  is usually much smaller than  $F$ , otherwise performing NMF has no sense, indeed let us recall that NMF is a LDR technique and we expect to extract the meaningful properties of  $V$  with a factorization rank  $K$  such that  $K \ll \min(F, N)$ .
- One can show that the dimension of  $\text{conv}(\Pi_{\Delta^F}(V)) = \text{rank}(V) - 1$ , see [56, Lemma 2.5] for the detailed proof.
- The dimension of the nested polytope  $\text{conv}(\Pi_{\Delta^F}(W))$  is not known a priori but, since  $\text{rank}(W)$  belongs to interval  $[\text{rank}(V); K]$  (1.6), then dimension of  $\text{conv}(\Pi_{\Delta^F}(W))$  belongs to interval  $[\text{rank}(V) - 1; K - 1]$  per [56, Lemma 2.5]. When the three polytopes (inner, nested, and outer) have the same dimension, this problem is well known in computational geometry and is referred to as the nested polytope problem (NPP) [36].
- An important variant of the exact NMF model is the *restricted* exact NMF for which we impose  $\text{rank}(V) = \text{rank}(W)$ . We can prove that NPP and this restricted variant of exact NMF are equivalent, that is, they can be reduced to one another [61]. Indeed, if  $\text{rank}(V) = \text{rank}(W)$  then the linear subspace spanned by the columns of  $V$  and  $W$  coincide, one can show that the outer polytope can be restricted to  $\Delta^F \cap \text{col}(\Pi_{\Delta^F}(V))$  where  $\text{col}(A) = \{x \in \mathbb{R}^F | x = Ay, y \in \mathbb{R}^N\}$ . Since the dimension of  $\Delta^F \cap \text{col}(\Pi_{\Delta^F}(V)) = \text{rank}(V) - 1$  per [56, Lemma 2.5], then the inner, nested, and outer polytopes have the same dimension.

In the following Section 1.7, we introduce the major results about the complexity of NMF.

## 1.7 Computational complexity of NMF

In this section we briefly discuss the computational complexity of NMF. The standard algorithms proposed to tackle the problems associated to NMF models are generally based on local improvement heuristics. We can also cite another class of heuristics that are based on greedy rank-one downdating [12, 17, 19]. The most widespread optimization schemes to tackle the problems associated to NMF models are briefly presented in Section 1.10. To the best of our knowledge, there is no algorithm proposed in the literature that comes with a guarantee of optimality. This suggests that solving NMF to optimality is a difficult problem.

The early results about the computational complexity of NMF were introduced by Thomas [131] and concerns the relationship between the rank and nonnegative rank of a matrix in the case of exact NMF models. They are summarized as follows: if  $V$  is a nonnegative matrix with  $\text{rank}(V) \leq 2$ , then  $\text{rank}(V) = \text{rank}_+(V)$ . In this case, Tomas showed that Exact NMF is solvable in polynomial time, indeed it is easily solvable since  $W$  can be built up by picking two columns of  $V$ .

In the case  $\text{rank}(V) = 3$ , finding the minimal  $K$  such that an restricted exact NMF (RE-NMF) exists can be solved in polynomial time. This results follows:

- the fact that there are polynomial-time reductions from RE-NMF to the nested polytope problem (NPP) and from NPP to RE-NMF [135, 61, 28].
- the polynomial time algorithms from Silio [122] and Aggarwal et al. [3] for the 2-dimensional NPP.

In the case  $\text{rank}(X) \geq 4$ , finding the minimum  $K$  such that RE-NMF has a solution is NP-hard, this is a consequence of the NP-hardness results from [34, 35].

The main results for the complexity of exact NMF are introduced in the seminal paper of Vavasis [135]. Vavasis [135] provides the proof that

**Theorem 1.7.1.** [135, Theorem 4] *NMF is NP-hard.*

*Sketch of the proof.* Vavasis considers the exact NMF stated as follows: given a matrix  $V \in \mathbb{R}_+^{F \times N}$  such that its rank is equal to  $K$ , the input is the pair  $(V, K)$ . The output is a pair of nonnegative matrices  $(W, H)$ , where  $W \in \mathbb{R}^{F \times K}$  and  $H \in \mathbb{R}^{K \times N}$  such that  $V = WH$ . If no such  $(W, H)$  exist, then the output is a statement of nonexistence of a solution. The decision version of exact NMF takes the same input and gives as output yes if such a  $W$  and  $H$  exists else it outputs no. Vavasis considers the implicit statement that the rank of  $V$  is known, this assumption has no impact on the results as we can use polynomial-time algorithms to compute the rank via an echelon-form or a singular value decomposition. Let us remark that, in the case  $\text{rank}(V) = K$ , the exact NMF and restricted exact NMF models coincide; it can easily be proved by using inequalities (1.6):

$$K = \text{rank}(V) \leq \text{rank}(W) \leq \min(F, K) = K$$

since  $K \ll \min(F, K)$  in general. Therefore we have  $\text{rank}(V) = \text{rank}(W)$  which corresponds to the RE-NMF. The proof of NP-hardness has two parts: first Vavasis uses the corresponding NPP to RE-NMF, that he refers to as the the intermediate simplex problem, and secondly shows that there exists a polynomial-time reduction of the NP-complete problem 3-SAT to the intermediate simplex problem.  $\square$

Many important remarks are stated here-under:

- the exact NMF can be generalized by the problem of nonnegative rank determination. This generalization is due to Cohen and Rothblum, which asks: give an nonnegative matrix  $V$ , find the smallest integer  $K$  such that we find two matrices  $W \in \mathbb{R}^{F \times K}$  and  $H \in \mathbb{R}^{K \times N}$  such that  $V = WH$ . Cohen and Rothblum give a super-exponential time algorithm for this problem. Since nonnegative rank determination is a generalization of exact NMF, then the results in [135] shows that the nonnegative rank determination is also NP-hard.
- For any approximate NMF models for which  $\text{rank}(V)$  is not constrained and the exact equality is not required, an optimal algorithm when presented with an  $V$  whose rank is exactly  $K$  ought to solve the exact NMF problem. Hence, the standard NMF problem using any norm is a generalization of exact NMF. Therefore, any hardness result that applies to exact NMF apply to most approximated NMF models as well [135].
- In Theorem 1.7.1,  $K = \text{rank}(V)$  is part of the input meaning that, unless  $\mathbf{P}=\mathbf{NP}$ , there is no algorithm polynomial in  $K$  and in the size of  $V$  that solves Exact NMF. We can also cite [121] that gives different proof using algebraic arguments. Moreover, Arora et al. [11] showed that there is no algorithm to solve this problem that runs in time  $(FN)^{o(r)}$  unless 3-SAT<sup>3</sup> can be solved in time  $2^{o(n)}$  on instances with  $n$  variables. However in practice,  $K$  is small and it makes senses to wonder what becomes the complexity if we assume instead that  $K$  is a fixed constant. It turns out that we can actually solve RE-NMF in polynomial time in  $F$  and  $N$ , namely in time  $\mathcal{O}((FN)^{cr^2})$  for some constant  $c$  [11, Lemma 2.2]. The argument is based on the quantifier elimination theory (in particular by using the seminal result by Basu, Pollack and Roy [14]). Unfortunately, this cannot be used in practice even for small size matrix because of its high computational cost: although the term  $\mathcal{O}((FN)^{cr^2})$  is a polynomial in  $F$  and  $N$  for  $K$  fixed, it grows extremely fast. Let us illustrate with a 4-by-4 matrix with  $K = 3$ , we have a complexity of order  $16^9$  and for a 5-by-5 matrix with  $K = 4$ , the complexity raises up to  $25^{16}$ . Therefore developing an effective code for exact NMF for small matrices is an important direction for further research. Some heuristics have been recently developed that allows solving exact

---

<sup>3</sup>3-SAT, or 3-satisfiability, is an instrumental problem in computational complexity to prove NP-completeness results. 3-SAT is the problem of deciding whether a set of clauses containing 3 Boolean variables or their negation can be satisfied. A clause is for example "x = 1 or y = 0 or z = 1".

NMF for matrices up to a few dozen rows and columns but these heuristics come with no global optimality guarantee [133]. This observation led us to develop new formulation for problems associated to the exact NMF models, these problems are presented in Chapter 6.

- Recently, Shitov showed that
  - NMF remains NP-hard when restricted to Boolean matrices [146].
  - The nonnegative rank over the reals might be different from the nonnegative rank over the rationals. Shitov [121, 145] gives an explicit example of a  $21 \times 21$  matrix with integral entries that can be written as a sum of 19 nonnegative rank-one matrices but not as a sum of 19 rational nonnegative rank-one matrices. This gives a solution of the Cohen–Rothblum problem, which has been open until [145, 121]. Let mention that same results have been independently obtained by Chistikov et al. [27] with a 6-by-11 rational matrix whose nonnegative rank is 5 while over the rational it is 6. This results also imply the nonnegative rank computation is not in NP since the size of the output is not bounded by the size of the input [55].
- Recent NMF models are proposed with additional assumptions such that the associated problems are easier to solve or come with some theoretical guarantees on the nature of the solutions. In [11] for instance, Arora et al. identify a class of models for which the problem is much easier. This class of models is the so-called Separable NMF that will be detailed in Section 1.8.3.



## 1.8 From model uniqueness to identifiable NMF

*"The investigation of the truth is in one way hard, in another easy. An indication of this is found in the fact that no one is able to attain the truth adequately, while, on the other hand, no one fails entirely, but everyone says something true about the nature of all things, and while individually they contribute little or nothing to the truth, by the union of all a considerable amount is amassed."*

- Aristotle, Metaphysics

### Motivations

The problems associated to NMF models like the standard problem presented in Section 1.4 have one common goal in the end; identify the "good"  $W$  and  $H$ , by good we mean: the ground-truth (or latent) factors, denoted  $W^\#$  and  $H^\#$ , that generated the data  $V$ . The main limitation of such parameter identification problems is the consequence of intrinsic limitations of NMF models: the NMF models without additional constraints are nonunique. This section introduces the basics for Chapters 2 and 3 and is organized as follows:

- in Section 1.8.1 we formally define the concept of identifiability for NMF models. The idea is to find conditions on  $W$  and  $H$  under which the optimal solutions  $(W^*, H^*)$  of an optimization problem associated to an NMF model are just permuted and scaled versions of the true factors  $W^\#$  and  $H^\#$ .
- In Section 1.8.2, we introduce two sufficient conditions under which recent NMF models and their associated optimization problems have been proved to be identifiable.
- In Section 1.8.4, we show how to relax these sufficient conditions, we present an exact NMF model dubbed as "simplex-structured NMF model" and its associated optimization problems that have been proved to be identifiable.

### 1.8.1 Identifiability for NMF

The notion of identifiability is closely related to the notion of uniqueness, which is known in the signal processing community since it is one of the goals when it comes to parameter estimation. For NMF, identifiability refers to the ability to identify the data generating factors  $W^\# \geq 0$  and  $H^\# \geq 0$  that give rise to  $V = W^\# H^\#$ . A nonnegative matrix factorization model is nonunique: given a factorization  $V = WH$  with  $W \geq 0$  and  $H \geq 0$ ; then any invertible matrix  $Q$  such that:

- $WQ \geq 0$

- $Q^{-1}H \geq 0$

provides an alternative solution  $V = \tilde{W}\tilde{H} = (WQ)(Q^{-1}H)$ .

Remark that  $Q$  is not restricted to generalised permutation matrices but can be anything else as soon as the alternative solutions  $(WQ, Q^{-1}H)$  satisfy the nonnegativity constraints. The study of NMF identifiability tends to elaborate models, associated optimization problems and conditions under which most  $Q$  can be removed [48]. In particular, we ignore the cases  $Q$  is a composition operator of permutation and scaling of the columns of  $W$  and  $H$ . These degrees of freedom are unavoidable and, most importantly, inconsequential for the applications at hand. Let us now formally define the identifiability for NMF models:

**Definition 1.8.1.** *Given a data matrix  $V = W^\#H^\#$  where  $W^\#$  and  $H^\#$  are the ground-truth (or latent) factors. Suppose that  $W^\#$  and  $H^\#$  satisfy a certain condition. Let  $(W^*, H^*)$  be the optimal solution of the optimization problem associated to the NMF model or an output from a procedure. If, under the aforementioned condition of  $W^\#$  and  $H^\#$ , we have:*

$$W^* = W^\#Q, H^* = Q^{-1}H^\#, \quad (1.13)$$

where  $Q = \Pi D$  with  $\Pi$  a permutation matrix and  $D$  is a full-rank diagonal matrix, then we say that the NMF model is identifiable under that condition.

Let us illustrate the rationale to show that a model is identifiable; let us consider the most popular optimization problem associated to the approximate NMF model:

$$\begin{aligned} \min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} \quad & \|V - WH\|_F^2 \\ \text{subject to} \quad & H \geq 0, W \geq 0, \end{aligned} \quad (1.14)$$

where  $\|V - WH\|_F^2$  is the squared Forbenius norm of the residual matrix  $V - WH$ . One can observe that (1.14) is a particular case of the standard problem presented in Section 1.4 in which the metric  $d(V_{fn}|[WH]_{fn}) = ([V - WH]_{fn})^2$ . To demonstrate the identifiability of an NMF model and its associated optimization problem, one needs to determine the conditions for  $W$  and  $H$  to be satisfied such that equations (1.13) hold. Many important remarks from [50] are listed here-under:

- in linear algebra, identifiability of a factorization model such the matrix factorization or the tensor factorization, commonly designates the ability of the model to identify the latent factors independently from any associated optimization problems (also referred to as "identification criteria"). In our case, we define the concept of identifiability of a model with its associated optimization problem. This seems to be unusual in the linear algebra community but it turns out that the choice of the associated optimization problem is crucial to establish identifiability of NMF models. Actually, an NMF model with a particular associated optimization problem may have be identifiable but the same model with a different optimization problem may not be. However, we must mitigate these assertions as under strong conditions on  $W$

and  $H$ , identifiability for exact NMF models holds independently from the associated problem. Further, we will relax these strong assumptions and then it will become necessary to build an optimization problem associated to a regularized exact NMF model to guarantee identifiability, see Section 1.8.4.

- A natural question should arrive: Should we care about identifiability ? It turns out that yes, it is important as the notion of identifiability is closely related to meaningful results when we have engineering applications at hand. The authors of [50] illustrates this remark based on a widespread application of NMF; topic mining model. In a nutshell, the task of topic mining model is to discover prominent topics (represented by sets of words) from a large set of documents. The results obtained for the analysis of real news articles with two methods, one identifiable and the second one not, are compared. We observe that the different topics mined from the non-identifiable method are mixed leading to weird associations such as "Lewinsky-white-star-president". Of course, we could think to others applications in which inconsistent results can have detrimental impact, let us imagine a case where NMF models are used to un-mix signals recorded on an aerospace structure in which there are different sources of mechanical vibrations. Some mechanical vibrations indicate the near failure of a system, therefore the identification of the ground truth signals is crucial for safety purposes. This illustrates the pivotal role of identifiability in many real life applications [50].

In the following sections we present some key conditions under which recent NMF models and their associated optimization problems have been proved to be identifiable. Note that we consider only the exact NMF models for which the factorization rank  $K$  is equal to the rank of the input matrix  $V$ . Most theoretical results on identifiability focus on Exact NMF. The reason is mainly due to the fact that it is easier to study identifiability for noiseless cases.

### 1.8.2 Sufficient conditions

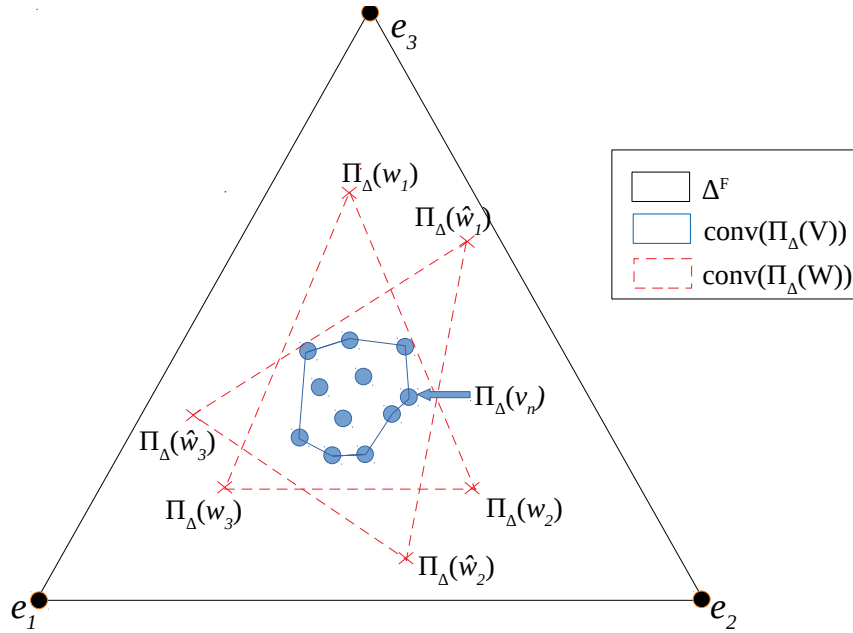
Most results on identifiability of Exact NMF have been derived under the hypothesis  $K = \text{rank}(V) = \text{rank}_+(V)$  (which implies  $K = \text{rank}(W) = \text{rank}(H)$ ). The reason comes from the fact that identifiability is intensively considered in applications for which we typically consider that  $\text{rank}(V) = \text{rank}_+(V)$  [56]. We introduce two sufficient conditions for  $W$  and  $H$  to obtain unique exact NMF  $V = WH$ , namely

- the separability condition, and
- the sufficiently scattered condition.

Note that we speak about "unique exact NMF", but we spoke about identifiability in the previous section. As mentioned earlier, the notions of identifiability and uniqueness of a solution for an exact NMF model are closely related, they are equivalent in the case we define uniqueness up to permutation and scaling ambiguities, that are inconsequential for

applications. Therefore, if we prove that an exact NMF  $V = WH$  is unique under mild conditions on the factors, then we can conclude to the identifiability of the exact NMF model or the identifiability of the exact NMF model and its associated optimization problem.

Before we formally define the separability and the sufficiently scattered conditions, we give the geometrical intuition that led researchers to come up with such conditions. We use in particular the geometric interpretation of exact NMF model in terms of nested convex cones: we showed in Section 1.6 that computing an exact NMF  $V = WH$  with a factorization rank  $K$  is equivalent to finding a polytope,  $\text{conv}(\Pi_{\Delta^F}(W))$ , nested between two given polytopes,  $\text{conv}(\Pi_{\Delta^F}(V))$  and the unit simplex  $\Delta^F$ , which can be very ill-posed as many solutions may exist. Any  $\hat{W}$  that satisfies  $\text{conv}(\Pi_{\Delta^F}(V)) \subseteq \text{conv}(\Pi_{\Delta^F}(W)) \subseteq \Delta^F$  also satisfies the data model  $\Pi_{\Delta^F}(V) = \Pi_{\Delta^F}(W)H' = \Pi_{\Delta^F}(\hat{W})\hat{H}'$ , see Figure 1.9.



**Fig. 1.9.** An illustration of the ill posedness of NMF for  $K = 3 = F$  based on the nested convex hulls interpretation (we look at the unit simplex from "above", in the direction  $(-1,-1,-1)$ ). The dashed lines represent the boundaries of  $\text{conv}(\Pi_{\Delta^F}(W))$  and  $\text{conv}(\Pi_{\Delta^F}(\hat{W}))$  that enclose the convex hull of the data showing there are many solutions that satisfy the inclusions of convex hulls. Vectors  $\Pi_{\Delta^F}(w_k)$  and  $\Pi_{\Delta^F}(\hat{w}_k)$  ( $1 \leq k \leq K$ ) correspond to the columns of  $\Pi_{\Delta^F}(W)$  and  $\Pi_{\Delta^F}(\hat{W})$ . Vectors  $e_i$  with  $1 \leq i \leq 3$  correspond to the vectors of the canonical basis of  $\mathbb{R}^3$ .

By closely looking at Figure 1.9, intuitively if the data points, namely the columns of  $\Pi_{\Delta^F}(V)$ , are well spread in  $\Delta^F$  such that  $\Pi_{\Delta^F}(V) \approx \Delta^F$ , then it would be hard to find a  $\hat{W}$  such that the above inclusion holds and then, an unique factorization model is probably guaranteed. This intuition suggests that, to understand the way we establish identifiability

for NMF models, it is important to characterize the distribution of columns of  $\Pi_{\Delta^F}(V)$  within the unit simplex, or equivalently the distribution of the data vectors  $V(:, n)$  within the nonnegative orthant. Indeed we can build up equivalent conclusions based on the nested cones interpretation of NMF as we showed the equivalence between the two geometric interpretations of NMF. In the literature, we characterize the distributions in terms of the geometric properties of  $H$  or equivalently  $W$  by symmetry of the factorization. The dispersion, or more precisely the *scattering*, of the data points (columns of  $V$ ) is translated to the scattering of the columns  $H(:, n)$  of  $H$  as  $V$  and  $H$  are linked by a full-column-rank matrix  $W$  through the exact NMF model. In others words, we can characterize the scattering of the data points whether in  $\mathbb{R}_+^F$  (and equivalently in  $\Delta^F$ ) or in  $\mathbb{R}_+^K$ .

In the following section, we define the first sufficient condition that is the separability condition.

### Separability

Let us now define a separable matrix.

**Definition 1.8.2.** *A nonnegative matrix  $H \in \mathbb{R}_+^{K \times N}$  is said to satisfy the separability condition if*

$$\text{cone}(H) = \text{cone}(e_1, \dots, e_K) = \mathbb{R}_+^K. \quad (1.15)$$

where  $e_k$  is the  $k$ th canonical basis vector in  $\mathbb{R}^K$ .

We now define the notion of a dual cone; given a cone  $\mathcal{S} \subseteq \mathbb{R}^K$ , its dual cone denoted  $\mathcal{S}^*$  is defined as follows

$$\mathcal{S}^* = \{y \in \mathbb{R}^K \mid y^T x \geq 0 \text{ for all } x \in \mathcal{S}\}. \quad (1.16)$$

**Lemma 1.8.1.** *Given a matrix  $H \in \mathbb{R}^{K \times N}$ , the dual cone of  $\text{cone}(H)$  is given by:*

$$\text{cone}^*(H) = \{y \in \mathbb{R}^K \mid y^T H \geq 0\}. \quad (1.17)$$

Two natural consequences of lemma 1.8.1 are:

1. Given two matrices  $W \in \mathbb{R}^{F \times K}$  and  $H \in \mathbb{R}^{K \times N}$ ,  $WH \geq 0$  if and only if  $\text{cone}(W^T) \subseteq \text{cone}^*(H)$ .
2. The nonnegative orthant is a self-dual cone, that is,  $(\mathbb{R}_+^F)^* = \mathbb{R}_+^F$ .

We have the theoretical background to introduce equivalent conditions for a matrix  $H$  to satisfy the separability condition.

**Lemma 1.8.2.** *Let  $H \in \mathbb{R}_+^{K \times N}$ . The following conditions are equivalent:*

- $H$  satisfy the separability condition, that is,  $\text{cone}(H) = \mathbb{R}_+^K$ ,
- $\text{cone}^*(H) = \mathbb{R}_+^K$ ,

- $H$  contains an  $K$ -by- $K$  submatrix which is a permutation of a diagonal matrix with positive diagonal entries [Lemma 4.11, [56]]. In others words, For every  $k = 1, \dots, K$ , there exists a column index  $n_k$  such that  $H(:, n_k) = \alpha_k e_k$  where  $\alpha_k > 0$  is a scalar.

Before we provide the conditions of the uniqueness of Exact NMF, let us provide a seminal result from [84]:

**Theorem 1.8.1.** [84, Theorem 1] *The Exact NMF  $V = WH$  of  $V$  of size  $K = \text{rank}(V)$  is unique if and only if the only simplicial<sup>4</sup> cone  $\mathcal{T}$  of order  $K$  such that*

$$\text{cone}(W^T) \subseteq \mathcal{T} \subseteq \text{cone}^*(H), \quad (1.18)$$

*is the nonnegative orthant  $\mathbb{R}_+^K$ .*

We will see in the following that Theorem 1.8.1 is particularly useful to derive sufficient conditions on  $W$  and  $H$  to ensure uniqueness of the exact NMF model. Indeed, let us combine the definition of a separable matrix from 1.8.2 and Theorem 1.8.1 we have the following sufficient condition for a solution of an exact NMF model to be unique.

**Theorem 1.8.2.** [56, Theorem 4.12] *If matrix  $V$  admits an exact NMF  $V = WH$  of size  $K = \text{rank}(V)$  with  $W^T$  and  $H$  that satisfy the separability condition, then the solution is unique.*

*Proof.* Since  $W^T$  and  $H$  satisfy the separability condition, by Lemma 1.8.2, we have

$$\text{cone}(W^T) = \mathbb{R}_+^K = \text{cone}^*(H).$$

Therefore, the only simplicial cone  $\mathcal{T}$  nested between  $\text{cone}(W^T)$  and  $\text{cone}^*(H)$  is the non-negative orthant  $\mathbb{R}_+^K$ . We conclude the proof by using the sufficient condition from Theorem 1.8.1.  $\square$

Then, under the condition that  $W^T$  and  $H$  are separable, we have the following equivalent conclusions:

- the solution  $(W, H)$  is unique up to permutation and scaling ambiguities,
- the exact NMF model  $V = WH$  with  $W \geq 0$  and  $H \geq 0$  is identifiable,

and these conclusions are valid independently from the procedure used or from the associated optimization problem that has been solved to compute  $(W, H)$ . The reason is due to the strongness of the conditions on both  $W^T$  and  $H$ . However, it is considered a relatively restrictive condition. Indeed the requirement that  $W^T$  and  $H$  both satisfy the separability condition is unlikely to be satisfied in real-world settings. Indeed, for many real-life applications, we can hope for one factor to be separable at most. In Section 1.8.3, we will show how to ensure uniqueness of the solution with only one factor that satisfies

<sup>4</sup>A simplicial cone is a polyhedral cone of the form  $\text{cone}(W)$  where  $W$  has full column rank, such a cone is of order  $K$  if  $W$  has  $K$  columns.

the separability condition by adding constraints to the model, namely by imposing that  $W = V(:, \mathcal{K})$  with some set of indices  $\mathcal{K}$  of size  $K$ .

Another condition that can effectively model the scattering of the data vectors in the nonnegative orthant but much relaxed than the previous ones is the so-called sufficiently-scattered conditions, this is discussed in the following section.

### Sufficiently scattered

Let us define the sufficiently scattered condition:

**Definition 1.8.3.** *A nonnegative matrix  $H \in \mathbb{R}_+^{K \times N}$  is said to be sufficiently scattered if the following two conditions SSC1 and SSC2 are satisfied:*

1. *SSC1: the columns of  $H$  are spread enough so that  $\mathcal{C} \subseteq \text{cone}(H)$  where  $\mathcal{C}$  is the second-order cone defined as follows:  $\mathcal{C} = \{x \in \mathbb{R}_+^K \mid \|x\|_1 \geq \sqrt{K-1} \|x\|_2\}$ .*
2. *SSC2: There does not exist any orthogonal matrix  $Q$  such that  $\text{cone}(H) \subseteq \text{cone}(Q)$  except for permutation matrices. Note that an orthogonal matrix is a square matrix such that  $Q^T Q = I_K$ .*

The sufficiently scattered condition is a milder condition than separability, but not easy to picture. To understand conditions SSC1 and SSC2, one key aspect is the second-order cone  $\mathcal{C}$ . One can show that this second-order cone is tangent to every facet of the nonnegative orthant, see [Lemma 4.18, [56]], see Figure 1.10 for a closer look at  $\mathcal{C}$  for  $K = 3$ . Figure 1.10 also displays the set  $\mathcal{C} \cap \Delta^K$  which corresponds to a  $(K-1)$ -dimensional sphere. This sphere is centered at  $e/K$  ( $e$  is a all-one column vector of size  $K$ ) of radius  $\frac{1}{\sqrt{K(K-1)}}$  and pass through the points  $\bar{e}_k = \frac{1}{K-1}(e - e_k)$  ( $1 \leq k \leq K$ ). Hence, if  $H$  satisfies condition SSC1, namely  $\mathcal{C} \subseteq \text{cone}(H)$ , it means that the columns are spread enough in the nonnegative orthant; at least every facet of the nonnegative orthant is touched by some columns of  $H$  [50]. It then imply some sparsity (zero elements) in the columns of  $H$ .

The condition SSC2 is kind of regularity condition that geometrically means that  $\text{cone}(H)$  has to be slightly larger than  $\mathcal{C}$ .

Before illustrating the different scenarios for  $\text{cone}(H)$ , we need to introduce the following Lemmas:

**Lemma 1.8.3.** *Let  $\mathcal{S}$  and  $\mathcal{Q}$  be two convex cones. If  $\mathcal{S} \subseteq \mathcal{Q}$  then  $\mathcal{Q}^* \subseteq \mathcal{S}^*$ .*

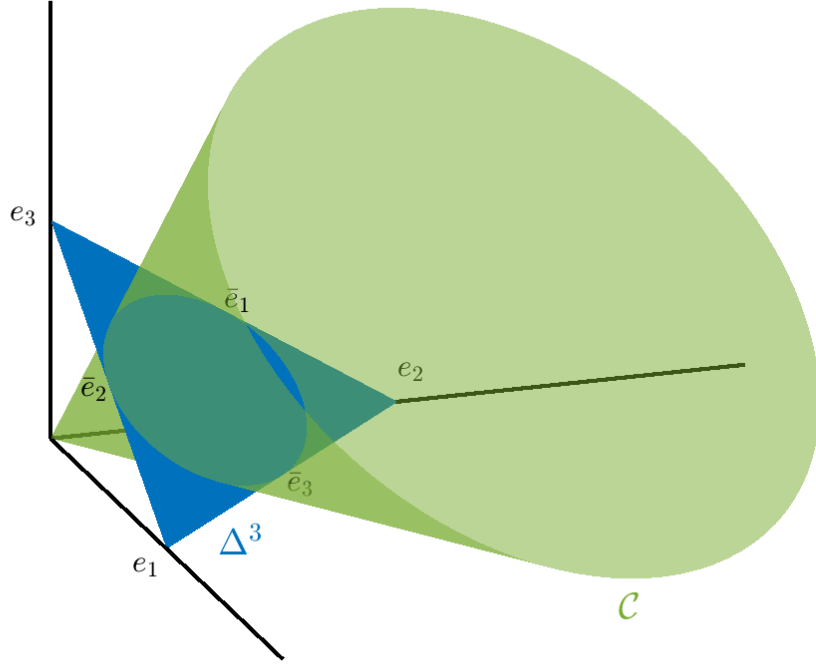
**Lemma 1.8.4.** *The dual cone of  $\mathcal{C}$  is given by*

$$\mathcal{C}^* = \{x \in \mathbb{R}^K \mid \|x\|_1 \geq \|x\|_2\} \quad (1.19)$$

**Lemma 1.8.5.** [75, Lemma 1] *Let  $Q \in \mathbb{R}^{K \times K}$  such that  $\|Q(:, k)\|_2 = 1$  for all  $k$ , if*

$$\mathcal{C} \subseteq \text{cone}(Q) \subseteq \mathcal{C}^* \quad (1.20)$$

*then:*



**Fig. 1.10.** Illustration of the second-order cone  $\mathcal{C}$  for  $K = 3$  and the 2-dimensional sphere (disk) centered at  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  of radius  $\frac{1}{\sqrt{6}}$  that goes through the points  $\bar{e}_1 = (0, \frac{1}{2}, \frac{1}{2})$ ,  $\bar{e}_2 = (\frac{1}{2}, 0, \frac{1}{2})$  and  $\bar{e}_3 = (\frac{1}{2}, \frac{1}{2}, 0)$ . This disk is the intersection between  $\mathcal{C}$  and the unit simplex  $\Delta^3$ . This figure has been reproduced from [56].

- $Q$  is orthogonal and
- $e^T Q = e^T$ .

A first consequence of Lemmas 1.8.4 and 1.8.5 concerns the orthogonal matrices  $Q \in \mathbb{R}^{K \times K}$  that satisfies  $\mathcal{C} \subseteq \text{cone}(Q) \subseteq \mathcal{C}^*$ . For  $e^T Q = e$  and  $Q^T Q = I$ , it implies that the columns of  $Q$  belong to the boundary of  $\mathcal{C}^*$ , denoted  $\mathbf{bDC}^* = \{x \in \mathbb{R}^K \mid \|x\|_1 = \|x\|_2\}$ .

We finally introduce two Lemmas that will be useful for Theorem 3.2.1 in Section 3.2.2.

**Lemma 1.8.6.** [56, Lemma 4.8] *Let  $Q \in \mathbb{R}^{K \times K}$  be an invertible matrix, then  $\text{cone}^*(Q^T) = \text{cone}(Q^{-1})$ .*

By posing  $A = Q^{-1}$ , Lemma 1.8.6 implies that  $\text{cone}^*(A^{-T}) = \text{cone}(A)$ .

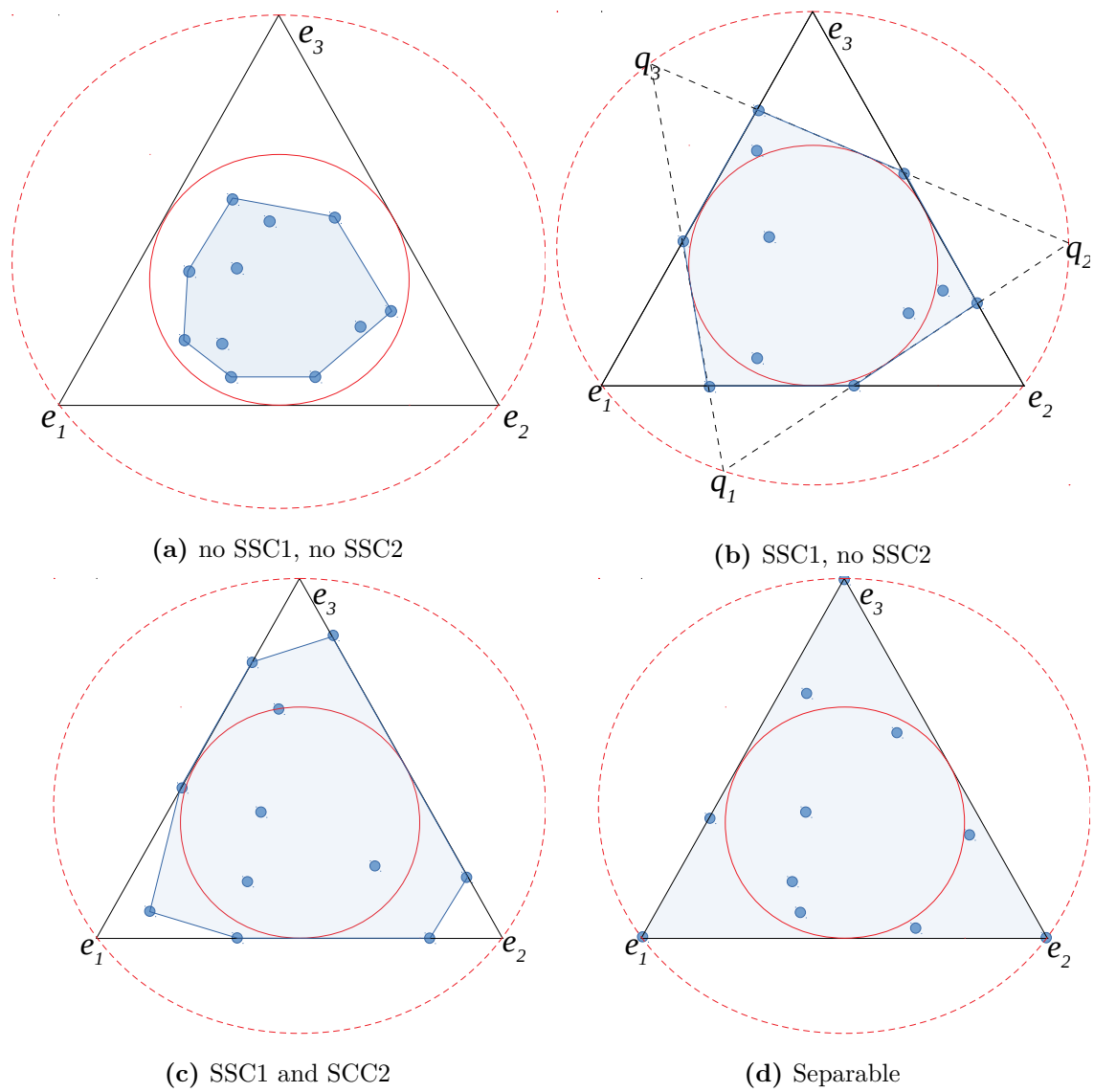
**Lemma 1.8.7.** *Let  $Q$  be an orthogonal matrix of size  $K \times K$ , then  $\text{cone}^*(Q) = \text{cone}(Q)$ , hence  $\text{cone}(Q)$  is self-dual.*

*Proof.* Since  $Q$  is orthogonal by hypothesis, then  $Q^T = Q^{-1}$  and  $Q^{-T} = Q$ , and per Lemma 1.8.6 we can write the following chain  $\text{cone}^*(Q) = \text{cone}^*(Q^{-T}) = \text{cone}(Q)$ .  $\square$

Since  $Q$  is orthogonal, Lemma 1.8.6 implies that  $\text{cone}^*(Q^T) = \text{cone}(Q^T)$ .

Figure 1.11 illustrates the different scenarios for  $\text{cone}(H)$  for  $K = 3$  projected onto the subset  $\{x \in \mathbb{R}^K \mid e^T x = 1\}$ , that are:





**Fig. 1.11.** Illustration of the SSC and Separable conditions after projection onto  $\{x \in \mathbb{R}^K | e^T x = 1\}$ . The inner circle corresponds to the boundary of  $\mathcal{C}$ , the shadowed polytope corresponds to  $\text{cone}(H)$ , the outer dashed circle corresponds to the boundary of  $\mathcal{C}^*$ , the triangle corresponds to the boundary of the unit simplex and the dashed triangle correspond to boundary of  $\text{cone}(Q)$  with  $Q$  orthogonal and whose columns belong to the boundary of  $\mathcal{C}^*$ , the blue dots correspond to the columns of  $H$ . (Figure similar to [53, Figure 8])

- in Figure 1.11 (a): matrix  $H$  does not satisfy SSC1 ( $\mathcal{C} \not\subseteq \text{cone}(H)$ ) nor SSC2 as we can easily find an orthogonal matrix  $Q$  which is not a permutation matrix such that  $\text{cone}(H) \subseteq \text{cone}(Q)$ , for example a triangle within  $\mathcal{C}^*$ .
- In Figure 1.11 (b): matrix  $H$  satisfies SSC1 but not SSC2. The dot triangle corresponds to the boundary of  $\text{cone}(Q)$  with  $Q$  an orthogonal matrix. In this example,  $Q$  is the rotation matrix of angle  $\theta = \pi/6$  around axis  $v = (1, 1, 1)$  defined as follows:

$$Q = \begin{pmatrix} 1 - 2s^2 + 2x^2s^2 & 2xys^2 - 2zsc & 2xzs^2 + 2yyc \\ 2xys^2 + 2zsc & 1 - 2s^2 + 2y^2s^2 & 2yys^2 - 2xsc \\ 2xzs^2 - 2yyc & 2yys^2 + 2xsc & 1 - 2s^2 + 2z^2s^2 \end{pmatrix},$$

where  $(x, y, z) = \frac{v}{\|v\|_2}$ ,  $c = \cos\left(\frac{\theta}{2}\right)$  and  $s = \sin\left(\frac{\theta}{2}\right)$ . We can easily check that the columns of  $Q$  belong to the boundary of  $\mathcal{C}^*$ .

- In Figure 1.11 (c): matrix  $H$  satisfies SSC1 ( $\mathcal{C} \subseteq \text{cone}(H)$ ) and SSC2 as no triangle within  $\mathcal{C}^*$  contains  $\text{cone}(H)$ , except the unit simplex.
- In Figure 1.11 (d): matrix  $H$  satisfies the separability condition as  $H$  contains a 3-by-3 submatrix which is a permutation of a diagonal matrix with positive diagonal entries, see Lemma 1.8.2.

Let us now state the main results of this section about the uniqueness of an exact NMF.

**Theorem 1.8.3.** [75, Theorem 4] *If  $W^T$  and  $H$  satisfy the sufficiently scattered conditions, then the exact NMF  $V = WH$  of size  $K = \text{rank}(V)$  is unique.*

*Proof.* As per Theorem 1.8.1, we need to prove that the the only simplicial cone  $\mathcal{T}$  nested between  $\text{cone}(W^T)$  and  $\text{cone}^*(H)$  is the nonnegative orthant  $\mathbb{R}_+^K$ . Since  $W^T$  and  $H$  satisfy the sufficiently scattered conditions, we have:

- SSC1 for  $W$ :  $\mathcal{C} \subseteq \text{cone}(W^T)$ ,
- SSC1 for  $H$ :  $\mathcal{C} \subseteq \text{cone}(H)$  and then  $\text{cone}^*(H) \subseteq \mathcal{C}^*$  by Lemma 1.8.3.

Hence, based on the inclusion  $\text{cone}(W^T) \subseteq \mathcal{T} \subseteq \text{cone}^*(H)$  we have  $\mathcal{C} \subseteq \mathcal{T} \subseteq \mathcal{C}^*$ . By Lemma 1.8.5, we have  $\mathcal{T} = \text{cone}(Q)$  for  $Q$  orthogonal satisfying  $e^T Q = e$  and  $Q^T Q = I$ . Finally, since  $H$  satisfies SSC2, then the only orthogonal matrix  $Q$  such that  $\text{cone}(H) \subseteq \text{cone}(Q)$  is a permutation matrix, then  $\mathcal{T} = \text{cone}(Q) = \mathbb{R}_+^K$  which concludes the proof.  $\square$

Then, under the condition that  $W^T$  and  $H$  are sufficiently scattered, we have the uniqueness of the exact NMF and therefore the exact NMF model is identifiable independently from any procedure used or associated optimization problem solved to compute  $(W, H)$ . As we pointed out for the separability conditions, both SSC conditions on  $W^T$  and  $H$  are unlikely to be satisfied in real-world applications. We show in Section 1.8.4 how to relax the requirement such that only one factor, usually  $H$ , satisfies the SSC condition.

### 1.8.3 Separable NMF

Separable NMF, in noiseless scenario, is a particular case of exact NMF where matrix  $H$  satisfies the separability condition, see Section 1.8.2. It means that  $H$  contains a  $K$ -by- $K$  submatrix which is a permutation of a diagonal matrix with positive diagonal entries. It implies that the columns of  $W$  are scaled versions of a subset  $\mathcal{K}$  of columns of  $V$  of size  $K$ . Mathematically, the separable model is expressed as follows:

$$V = WH = (V(:, \mathcal{K})\Lambda) \left( \begin{bmatrix} I_K & H' \end{bmatrix} \Pi_n \right), \quad (1.21)$$

for some  $H' \geq 0$ ,  $\Pi_n$  is a permutation matrix and  $\Lambda$  is a diagonal matrix such that the  $k$ -th diagonal element denoted  $\Lambda_{k,k}$  is  $\Lambda_{k,k} = \frac{\|W(:,k)\|}{\|V(:,k)\|}$  allowing the columns of  $W$  be scaled versions of  $V(:, \mathcal{K})$ .

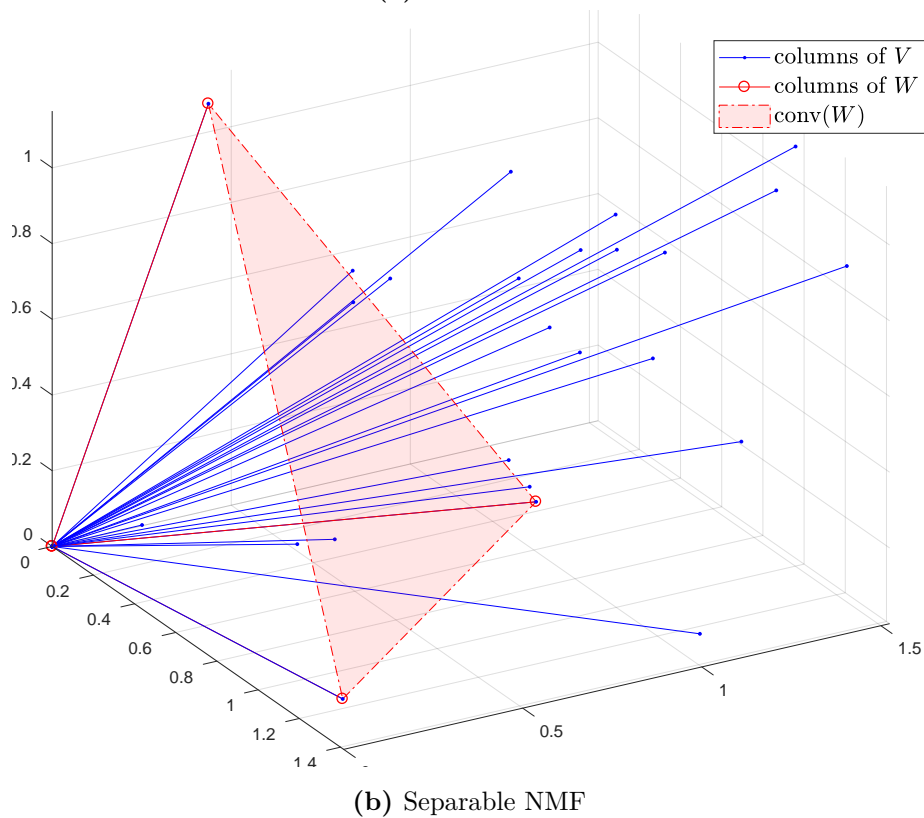
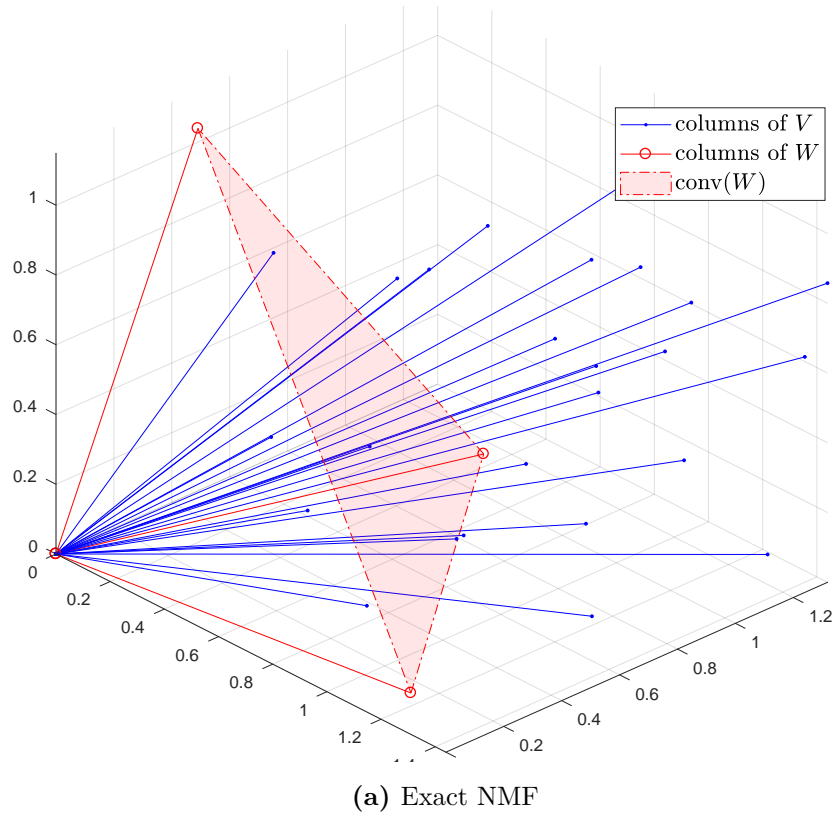
Geometrically, the conical hull of  $V(:, \mathcal{K})$  contains all the columns of  $V$  and  $\text{cone}(V) = \text{cone}(W)$ . Indeed, the "remaining" data points can be expressed as  $V(:, \bar{\mathcal{K}}) = V(:, \mathcal{K})\Lambda H' \Pi_n$  where  $\bar{\mathcal{K}}$  designates the complement of  $\mathcal{K}$ , so it means that these data points are some nonnegative linear combinations of data points  $V(:, \mathcal{K})$ . This is why Separable NMF is sometimes qualified as a self-dictionary model; indeed the  $K$  basis vectors are within the data set. Figure 1.12 illustrates the geometrical interpretations respectively for exact NMF and separable NMF for  $K = 3 = F$  and  $N = 25$ .

The main advantages of separable NMF are briefly enunciated here-under, we refer the reader to references for more details on the topic:

- The separable NMF is identifiable, see [56, Theorem 4.36].
- For Separable NMF it is possible to derive polynomial-time algorithms that provably recover  $(W, H)$ , even in the presence of noise. An example is the Successive Projection Algorithm (SPA) [10]. Moreover, SPA is provably robust to bounded additive noise [66]. However, a main drawback of SPA (and any algorithm relying on orthogonal projections) is that it requires  $\text{rank}(W) = K$ . In some real scenarios, this assumption is not satisfied, see Chapter 2 for more details. To get rid of this assumption, the author [57] proposes a projection onto the convex hull of the extracted columns and the origin. Moreover, Gillis [57] shows that, in the full-column rank case, that is  $\text{rank}(W) = K$ , SNPA is more robust to bounded noise than SPA algorithm. We refer the reader to [57] for more details about SNPA.

### 1.8.4 Minimum-volume NMF

We noticed earlier an important issue; the exact NMF model is identifiable under the condition that  $W^T$  and  $H$  both satisfy the sufficiently scattered condition. We mentioned that for real-life applications, these requirements on  $W$  and  $H$  are unlikely to be satisfied. For example, in image processing, the matrix  $W$  is usually dense. Actually, dense  $W$  frequently arise [50]. For such cases, how can we guarantee the identifiability? To handle this, let us consider the regularized Exact NMF model  $V = WH$  where the columns of



**Fig. 1.12.** Geometric illustration of separable NMF and NMF for  $F = K = 3$  and  $N = 25$ . The blue rays are data points  $V$ , the red rays correspond to the basis vectors (columns of  $W$ ) whose are a subset of columns of  $V$ , then we can see that  $\text{cone}(V) = \text{cone}(W)$ .

$H$  belong to the unit simplex, that is,  $H \geq 0$  and  $e^T H = e^T$ . This assumption is also called column-stochasticity. For such a model, it means that each data point is a convex combination of the columns of  $W$  and therefore we have  $V(:, n) \in \text{conv}(W)$  for all  $n$ . Since it is valid for all  $n$ , then we have  $\text{conv}(V) \subseteq \text{conv}(W)$ . In the literature, this model is referred to as simplex-structured NMF model (SSNMF). For comprehensive purposes, the geometry of exact NMF (based on nested cones) and exact NMF with  $H$  column-stochastic (SSNMF) models is shown on Figure 1.13 for  $K = F = 3$ .

Back in 1994, Craig [32] proposed an innovative way to recover  $W^\#$  (the ground truth basis matrix). He formulated the conjecture known as "Craig's belief" as follows: if the data points  $V(:, n)$  are sufficiently spread within  $\text{conv}(W)$ , then finding the matrix  $W$  whose convex hull has minimum volume identifies  $W^\#$  in the sense of Definition 1.8.1, or equivalently  $W$  is unique (up to permutation and scaling of the columns of  $W$  and the rows of  $H$ ). The corresponding problem is referred to as minimum-volume NMF (min-vol NMF). Intuitively, min-vol NMF looks for basis vectors as close as possible to the data points [50, 56]. The use of min-vol NMF has led to a new class of NMF methods that outperforms existing ones in many applications such as document analysis and blind hyperspectral unmixing; see the recent survey [50]. Note that min-vol NMF implicitly enhances the factor  $H$  to be sparse: the fact that  $W$  has a small volume implies that many data points will be located on the facets of the  $\text{conv}(W)$  hence  $H$  will be sparse.

Figure 1.14 illustrates the intuition of min-vol NMF; shrinking a data-enclosing convex hull to have minimum "volume".

Craig did not provide an optimization problem associated to his conjecture. In 2015, Fu et al. [53] came up with the first identifiability results for min-vol NMF. They formulated the following optimization problem:

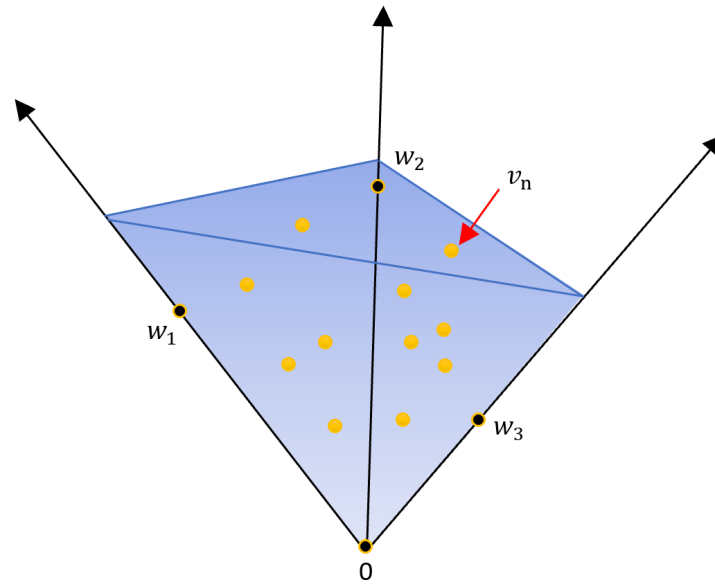
$$\begin{aligned} & \min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} \det(W^T W) \\ & \text{subject to} \quad V = WH, \\ & \quad \quad \quad H \geq 0, e^T H = e^T, \end{aligned} \tag{1.22}$$

where  $\det(W^T W)$  is a surrogate of the volume of  $\text{conv}(W)$ . Let us give more insights about this "volume" measure. One can easily show that the volume of  $\text{conv}(W)$  for a matrix  $W \in \mathbb{R}^{F \times K}$  with  $F \geq K$  is null (see Figure 1.13-(b),  $\text{conv}(W)$  is an open surface that obviously encloses no volume in  $\mathbb{R}^F$ ). In [101], the authors show how to measure the volume of  $\text{conv}(W)$  in a  $(K - 1)$ -dimensional linear subspace in the case  $W$  is full-column rank by using PCA. Let us introduce another volume measure, namely the volume of  $\text{conv}([0, W])$ :

**Lemma 1.8.8.** [56, Lemma 4.39] Given  $W \in \mathbb{R}^{F \times K}$  and  $\text{rank}(W) = K$ ,

$$\frac{1}{K!} \sqrt{\det(W^T W)}$$

is the volume of the convex hull of the columns of  $W$  and the origin in the linear subspace spanned by the columns of  $W$ .

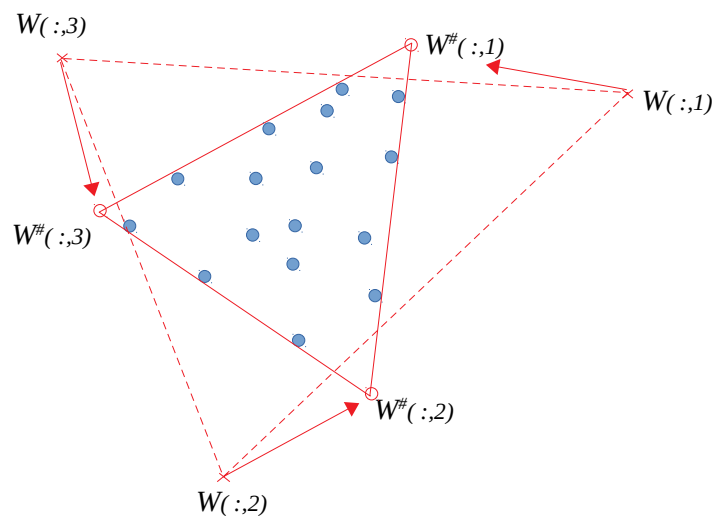


(a) standard exact NMF model(1.1)



(b) SSNMF model

**Fig. 1.13.** Geometric illustration of the standard exact NMF model (a) and exact NMF model with  $H$  being column-stochastic (b), referred to as SSNMF, for  $F = K = 3$ . The orange dots are the data vectors (columns of  $V$ ) and  $w_k$  ( $1 \leq k \leq 3$ ) correspond to the basis vectors (columns of  $W$ ). The shadowed polytope designates a part of  $\text{cone}(W)$  in (a) and  $\text{conv}(W)$  in (b). Figure reproduced from [90, Figures 1 and 2].



**Fig. 1.14.** Geometric intuition of min-vol NMF; finding the minimum-volume enclosing simplex will recover the ground-truth  $\text{conv}(W^\#)$ . The dots correspond to the data points. The red triangle designates  $\text{conv}(W^\#)$ , the dashed triangle designates the current  $\text{conv}(W)$ . This figure looks from "above" the convex hulls. Figure similar to [90, Figure 7].

Therefore, minimizing  $\det(W^T W)$  makes sense if the ultimate goal is to minimize the "volume" of  $\text{conv}(W)$ .

Craig's belief was a conjecture without theoretical proof, however it has been supported by many empirical results over the years. In [53], Fu et al. showed the following:

**Theorem 1.8.4.** [53, Theorem 1] *If  $\text{rank}(W^\#) = \text{rank}(H^\#) = K$  and  $H^\#$  satisfies the sufficiently scattered condition (Definition 1.8.3), then any optimal solution  $(W^*, H^*)$  of (1.22) is such that  $W^* = W^\# \Pi$  and  $H^* = \Pi^T H^\#$  where  $\Pi$  is a permutation matrix.*

*Proof.* We refer the reader to [53, Appendix A] for the detailed proof.  $\square$

Note that although min-vol NMF guarantees identifiability, there is a price to pay; the corresponding optimization problem (1.22) is still hard to solve in general, as for the original NMF problem [135]. Despite this nice result, the constraint  $H^T e = e$  makes the NMF model less general and does not apply to all data sets. In the case where the data does not naturally belong to a convex hull, one needs to normalize the data points so that their entries sum to one so that  $H^T e = e$  can be assumed without loss of generality (in the noiseless case). This normalization can sometimes increase the noise and might greatly influence the solution, hence are usually not recommended in practice; consider the case some columns of  $V$  have small norm, they typically contain less information and are more

easily affected by noise. The  $\ell_1$  normalization gives the same importance compared to columns with large norms which is not desirable; see the discussion in [48]. In order to fix this issue, Fu et al. [49] recently proposed the following optimization problem:

$$\begin{aligned} & \min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} \det(W^T W) \\ & \text{subject to} \quad V = WH, \\ & \quad \quad \quad H \geq 0, He = \rho e, \end{aligned} \tag{1.23}$$

where  $\rho > 0$  is any positive real number. One can see that both problems look very similar. However, the sum-to-one constraint is now on the rows of  $H$  (when  $\rho = 1$ ). As opposed to column stochasticity, row stochasticity of  $H$  can be assumed without loss of generality, even in the presence of noise, since any factorization  $WH$  can be properly normalized so that this assumption holds. In fact,  $WH = \sum_{k=1}^K (a_k W(:, k))(H(k, :)/a_k)$  for any  $a_k > 0$  for  $k = 1, \dots, K$ . In other terms, letting  $A$  be the diagonal matrix with  $A(k, k) = a_k = \sum_{j=1}^n H(k, j)$  for  $k = 1, \dots, K$ , we have  $WH = (WA)(A^{-1}H) = W'H'$  where  $H' = A^{-1}H$  is row stochastic.

Fu et al. [49] proved:

**Theorem 1.8.5.** [49, Theorem 1] *If  $\text{rank}(W^\#) = \text{rank}(H^\#) = K$  and  $H^\#$  satisfies the sufficiently scattered condition (Definition 1.8.3), then any optimal solution  $(W^*, H^*)$  of (1.23) is such that  $W^* = W^\#Q$  and  $H^* = Q^{-1}H^\#$  where  $Q = \Pi D$  with  $\Pi$  a permutation matrix and  $D$  is a full-rank diagonal matrix*

*Proof.* We refer the reader to [49] for the detailed proof.  $\square$

Although the two results above are equivalent in noiseless settings (after normalizing the input matrix), they might behave rather differently in noisy scenarios [56].

In Chapter 3 we propose a new optimization problem in which we use the condition  $W^T e = e$  that normalizes the columns of  $W$  instead of the rows of  $H$ . We prove in that chapter that requiring  $W$  to be column stochastic (which can also be made without loss of generality) also leads to identifiability. We also show that this normalization balances the importance of the columns of  $W$  leading to a better conditioned  $W$  and better performance in terms of recovery for noisy settings.

## 1.9 Approximate factorization

In the previous sections, we have focused on exact NMF models. This section presents the main concepts involved when we need to find solution  $(W, H)$  for approximate NMF models. As explained in Section 1.4, for an approximate NMF model the exactness is not required and we are searching for an approximate decomposition, that is,  $V \approx WH$ . The reason is the presence of noise, and the linear model being in most cases only an approximate model. As already mentioned in Section 1.8, in order to find  $(W, H)$ , we need to built



up optimization problems associated to NMF models. When it comes to optimization problems, we need to define the optimization variables, the objective function and the constraints. Over the years, many optimization problems associated to approximate NMF models have been introduced. The goal of this section is to present the most widespread optimization problems for NMF. In particular, we will discuss the choice of the error measure  $D(V|WH)$  in Problem 1.4.1 which is related to the statistic of the noise that we assume present in the data set. Then we will show that the optimization problem for NMF models are maximum-likelihood in disguise. We finally present a distributionally robust NMF problem.

### 1.9.1 The metrics: the $\beta$ -divergences

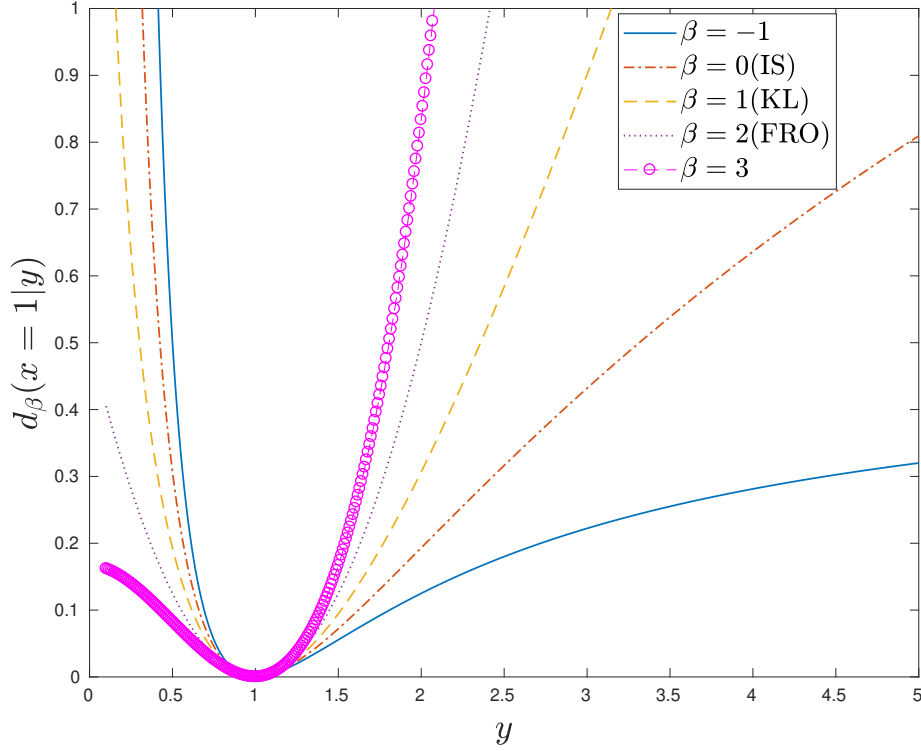
As explained in Section 1.4, the choice of a particular error measure  $D(V|WH)$  boils down to the choice for the scalar metric  $d(V_{fn}|[WH]_{fn})$ , also referred to as the scalar *divergence*. In this thesis, we mainly focus on the  $\beta$ -divergences: given two nonnegative scalars  $x$  and  $y$ , the  $\beta$ -divergence (for  $\beta \geq -1$ ) between  $x$  and  $y$  denoted  $d_\beta(x|y)$  is defined as follows:

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta(\beta-1)} (x^\beta + (\beta-1)y^\beta - \beta xy^{\beta-1}) & \text{for } \beta \in \mathbb{R} \setminus \{0, 1\}, \\ x \log \frac{x}{y} - x + y & \text{for } \beta = 1, \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \text{for } \beta = 0. \end{cases}$$

For  $\beta = 2$ , this is the standard squared Euclidean distance and  $D(V|WH)$  boils down to the squared Frobenius norm of  $V - WH$ , that is,  $\frac{1}{2} \|V - WH\|_F^2$ . For  $\beta = 1$  and  $\beta = 0$ , the  $\beta$ -divergence corresponds to the Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence, respectively. Figure 1.15 displays the function  $d_\beta(x|y)$  for  $x = 1$  and for different values of the parameter  $\beta$ , namely  $\beta = [-1, 0, 1, 2, 3]$ . We observe that, for  $\beta \leq 1$ ,  $d_\beta(x|y)$  goes to infinity as  $y$  goes to zero (because of the term  $y^{\beta-1}$ ). Hence a positive entry (here  $x = 1$ ) cannot be approximated by zero in the case for  $\beta \leq 1$ . This implies that  $\beta$ -divergences for  $\beta \leq 1$  tend to overapproximate the input matrix. On the opposite, as  $\beta$  increases, the values of  $\beta$ -divergences for  $y \leq x$  decrease, then  $\beta$ -divergences tend to underapproximate the input matrix for  $\beta \geq 1$ .

Let us mention important properties of the  $\beta$ -divergences:

- The  $\beta$ -divergence  $d_\beta(x|y)$  is homogeneous of degree  $\beta$ :  $d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y)$ . It implies that factorizations obtained with  $\beta > 0$  (such as the Euclidean distance or the KL divergence) will rely more heavily on the largest data values and less precision is to be expected in the estimation of the low-power components. The IS divergence ( $\beta = 0$ ) is scale-invariant that is  $d_{IS}(\lambda x|\lambda y) = d_{IS}(x|y)$ . The IS divergence is the only one in the  $\beta$ -divergences family to possess this property. It implies that entries of low power are as important in the divergence computation as the areas of high power. This property is interesting in audio source separation as low-power frequency bands can perceptually contribute as much as high-power frequency bands. Note that both



**Fig. 1.15.** Graph of the  $\beta$ -divergences  $d_\beta(x = 1|y)$  for  $\beta = [-1, 0, 1, 2, 3]$ .

KL and IS divergences are more adapted to audio source separation than Euclidean distance as it is built on logarithmic scale as human perception; see [88] and [45].

- The function  $d_\beta(x|y)$  is convex in the second argument  $y$  for  $\beta \in [1, 2]$ . Otherwise, the objective function is non-convex. This implies that, for  $\beta < 1$ , even the problem of inferring  $H$  with  $W$  fixed is non-convex.
- $d_\beta(x|y)$  for  $x = 0$  is not defined for all values of  $\beta$ , indeed:

$$d_\beta(0|y) = \begin{cases} \text{not defined} & \text{for } \beta \leq 0, \\ \frac{1}{\beta}y^\beta & \text{for } \beta > 0, \end{cases}$$

This means that, for NMF, one should use  $\beta$ -divergences with  $\beta \leq 0$  only when the input matrix is positive.

**Lemma 1.9.1.** *Given two nonnegative scalars  $x$  and  $y$  with  $x$  fixed, the KL-divergence  $d_{KL}(x|y) = x \log \frac{x}{y} - x + y$  is not  $L$ -smooth in  $y$  on its active domain  $\mathbb{R}_+$*

*Proof.* The gradient of the function w.r.t.  $y$  (for  $x$  fixed) is:

$$\nabla_y f(y) = 1 - x/y.$$

Let us consider  $y_1$  and  $y_2$  that belong to the domain of the function and  $x$  fixed, we write:

$$\begin{aligned} \|\nabla_y f(y_1) - \nabla_y f(y_2)\| &= \left\| \left(1 - \frac{x}{y_1}\right) - \left(1 - \frac{x}{y_2}\right) \right\| \\ &= \left\| x \left( \frac{1}{y_2} - \frac{1}{y_1} \right) \right\| \\ &\leq \|x\| \left\| \frac{1}{y_2} - \frac{1}{y_1} \right\| \\ &= \left\| \frac{x}{y_1 y_2} \right\| \|y_1 - y_2\|. \end{aligned}$$

It is then clear that the coefficient

$$L(y_1, y_2) = \left\| \frac{x}{y_1 y_2} \right\|$$

goes to infinite as  $y_1$  and/or  $y_2$  approaches to zero, therefore the gradient of function  $f$  is not Lipschitz-continuous in  $y$  in  $\text{int}(\text{dom} f)$ .

Note. Similar rationale can be followed for Itakura - Saito divergence ( $\beta = 0$ ).

□

For more properties of the  $\beta$ -divergences, we refer the reader to [45, 56].

## 1.9.2 The probabilistic view of NMF

In the NMF literature, the choice for the metric is typically dictated by the noise statistics assumed in the data sets. In this section we explain the equivalency between the error measures and maximum likelihood estimators of  $WH$  when we consider particular statistical distribution for the noise. We focus on  $\beta$ -divergences which are the most widely used in the NMF literature. First let us recall briefly what is a maximum-likelihood estimator; given a simple random set<sup>5</sup>  $x_i$  ( $1 \leq i \leq n$ ) for a random variable  $X$ , we assume that  $X$  follows a probability density function (p.d.f.) for which some parameters are unknown. The maximum-likelihood estimator consists in building the probability to observe the sample random set as a function of the unknown parameters, referred to as likelihood function, and finally estimate the values for these parameters such that the probability function is maximized; in others terms, we search for the parameters of the p.d.f. such that the observed random sample set has the maximum-likelihood. For NMF, we assume that the random simple set corresponds to the entries of input matrix  $V$  (the "observations") and the parameters of the p.d.f. are  $(W, H)$ . Let us give an example; we assume that we have noise in the data sets. For instance, we assume we have additive Gaussian noise of mean 0 and standard deviation  $\sigma$ , mathematically we write:  $\tilde{V} = WH + N$  where  $N_{fn} \sim \mathcal{N}(0, \sigma)$ . Remark that  $\tilde{V}$  and  $V$  respectively denote the random variables and the observations or the realizations of the random variable. Let us denote by  $p(\tilde{V}_{fn} | [WH]_{fn})$  the probability

<sup>5</sup>A simple random set is a set of observations of a random variable such that they are identically distributed and statistically independent.

to observe  $\tilde{V}_{fn}$  given  $[WH]_{fn}$ . Under the Gaussian assumption on the noise, the model  $\tilde{V} = WH + N$  and  $WH$  given, then  $\tilde{V}_{fn} \sim \mathcal{N}([WH]_{fn}, \sigma)$ , indeed:

- $\mathbb{E}(\tilde{V}_{fn}) = \mathbb{E}([WH]_{fn} + N_{fn}) = [WH]_{fn}$  per the properties of the expected value operator  $\mathbb{E}(\cdot)$  and,
- $\text{Var}(\tilde{V}_{fn}) = \text{Var}([WH]_{fn} + N_{fn}) = \text{Var}(N_{fn}) = \sigma^2$  per the properties of the variance operator  $\text{Var}(\cdot)$ .

Then the p.d.f.  $p(\tilde{V}_{fn}|[WH]_{fn})$  becomes:

$$p(\tilde{V}_{fn}|[WH]_{fn}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\tilde{V}_{fn}-[WH]_{fn})^2}. \quad (1.24)$$

Due to the i.i.d. assumption for a random simple set, we can build the likelihood function denoted  $L$ , that is the probability to observe  $V$ :

$$L(V|WH) = \prod_{fn}^{FN} p(V_{fn}|[WH]_{fn}) = \frac{1}{\sigma^{(FN)}(\sqrt{2\pi})^{(FN)}} e^{-\frac{1}{2\sigma^2} \sum_{fn} (V_{fn}-[WH]_{fn})^2}. \quad (1.25)$$

Given  $V$ , computing  $(W, H)$  that maximizes  $L(V|WH)$  under the nonnegativity constraints on  $(W, H)$  is the so-called maximum likelihood estimator; we compute  $(W, H)$  such as it satisfies the Karush–Kuhn–Tucker conditions. In practice, as  $L$  is a product, it is easier to cancel the derivative of its logarithm; the maximum of  $l = \log(L) = -(FN) \log(\sqrt{2\pi}) - (FN) \log(\sigma) - \frac{1}{2\sigma^2} \sum_{fn} (V_{fn} - [WH]_{fn})^2$  is reached for the same value of  $WH$ . It does not change the result as the logarithm is a monotone increasing function. Let us finally note that maximizing a function  $l$  is strictly equivalent to minimizing  $-l$ .

Now here comes the link with optimization problem associated to NMF. Let us consider the optimization problem associated to NMF models for which  $d(V_{fn}|[WH]_{fn}) = (V_{fn} - [WH]_{fn})^2$ , then we minimize the error measure  $D(V|WH) = \sum_{fn} (V_{fn} - [WH]_{fn})^2$ . Hence the divergence minimization becomes equivalent to a maximum likelihood estimator since

$$D(V|WH) = (-\log(L) - b) a$$

with  $b = (FN) \log(\sqrt{2\pi}) + (FN) \log(\sigma)$  and  $a = 2\sigma^2 > 0$ . Here-under we list a few important examples (where we assume i.i.d. noise) of equivalences between divergences and maximum-likelihood estimator:

- Poisson distribution [87]: The entries  $V_{fn}$  of  $V$  are distributed as a Poisson distribution of parameter  $[WH]_{fn}$ , that is  $\mathbb{P}(V_{fn} = k) = \frac{[WH]_{fn}^k e^{-[WH]_{fn}}}{k!}$  for  $k = 0, 1, \dots$ . Here the noise is not an additive noise. The Poisson distribution assumption is reasonable when we deal with count data such as a vector of word counts used in text mining topic or in image processing, as the acquisition process of a picture can be seen as a photo counter [91]. One can show that the maximum likelihood estimator for the Poisson distribution is the Kullback-Leibler (KL) divergence ( $\beta = 1$ ).

- Exponential distribution [43]: in this case we consider a multiplicative noise model  $V = [WH] \odot N$  such that  $N$  follows an exponential distribution of parameter  $\lambda = 1$  ( $N \sim \text{Exp}(1)$ ). Then, the entries  $V_{fn} \sim \text{Exp}(\frac{1}{[WH]_{fn}})$  as  $\mathbb{E}(V_{fn}) = \mathbb{E}([WH]_{fn}N_{fn}) = [WH]_{fn}\mathbb{E}(N_{fn}) = [WH]_{fn}\frac{1}{1} = [WH]_{fn}$ . One can show that the maximum likelihood estimator for the Exponential distribution is the Itakura-Saito (IS) divergence ( $\beta = 0$ ).

For others equivalences, we refer the reader to

- [78]: for an additive noise that follows a Laplacian distribution, the maximum-likelihood estimator is the  $\ell_1$  norm of the residual matrix  $V - WH$ ,
- [64]: for an additive noise following a uniform distribution, the maximum-likelihood estimator is the infinite norm of the residual matrix  $V - WH$ .
- [45] shows that the Tweedie distributions correspond to the maximum likelihood estimators for problems including the generalized  $\beta$ -divergence.

### 1.9.3 How to choose the metric in practice ?

Choosing a suitable objective function for the optimization problems associated to NMF models can be crucial. NMF divergence choice depends on the data and on the application and one can choose the divergence as follows:

- by intuition or from some prior knowledge of the application goal. For instance, if we use NMF for predicting the unseen data while minimizing the mean squared error then the Frobenius norm of the residual matrix  $V - WH$  is well suited. We can also cite the case we are looking for some invariances properties of the error measure, for instance the scale invariance of Itakura-Saito is well appreciated for audio signals analysis [43].
- For computational reasons; the Frobenius norm is L-smooth and therefore we can use efficient numerical scheme to solve it such as accelerated projected gradients methods [107].
- By cross-validation: the objective function is automatically selected using cross validation. In the case we know the ground-truth factors  $(W^\#, H^\#)$ , then they can be used to assess the quality of solutions computed with different error measures.
- From some probabilistic considerations: for some applications, particular distributions for the noise can be reasonably assumed and therefore we can choose the divergence accordingly as explained in Section 1.9.2. In some cases, we would like to compute an NMF solution that is robust to different types of noise distributions, this reason led us to develop a new NMF problem based on the minimization of the maximum value among several  $\beta$ -divergences. The optimization problem is detailed in Section 1.9.4.

### 1.9.4 Distributionally Robust and Multi-Objective Nonnegative Matrix Factorization

The content of this section is extracted from: [62] N. Gillis, Le Thi Khanh Hien, V. Leplat, and Vincent Y. F. Tan. *Distributionally robust and multi-objective nonnegative matrix factorization*. 2019. arXiv:1901.10757.

This section briefly summarizes research output [62]: a key aspect of optimization problems associated to NMF models is the choice of the objective function that depends, for instance, on the noise model (or statistics of the noise) assumed on the data. In many applications, the noise model is unknown and difficult to estimate. If the choice of the objective function is wrong, the NMF solution provided could be far from the desired solution. In [62] we compute an NMF solution that is robust to different types of noise distributions; this is referred to as distributionally robust. In mathematical terms, we will consider the problem:

$$\min_{W, H \geq 0} \max_{\beta \in \Omega} D_\beta(V, WH). \quad (1.26)$$

where  $\Omega$  is a subset of  $\beta$ 's of interest. However, the beta-divergences are homogeneous functions of degree  $\beta$ , therefore, depending on the scaling of the input matrix, the minimization of (1.26) amounts, in most cases, to minimizing a single objective corresponding to the  $\beta$ -divergence with the largest value. To tackle this issue we will scale the different objective functions as follows: first, we compute a solution  $(W_\beta, H_\beta)$  for  $\min_{W, H \geq 0} D_\beta(V, WH)$  to obtain the error  $e_\beta = D_\beta(V, W_\beta H_\beta)$ . Note that we can only compute this minimization in an approximate fashion because the NMF problem is NP-hard [135] as explained in Section 1.7. Then, we define  $V_\beta = \alpha_\beta V$  with  $\alpha_\beta = (1/e_\beta)^{1/\beta}$  so that

$$\bar{D}_\beta(V, WH) = \frac{D_\beta(V, WH)}{e_\beta},$$

so that  $\bar{D}_\beta(V, W_\beta H_\beta) = 1$ .

Finally, if the noise model on the data is unknown but corresponds to a distribution associated with a  $\beta$ -divergence with  $\beta \in \Omega$ , it makes sense to consider the following distributionally robust NMF (DR-NMF) problem

$$\min_{W, H \geq 0} \max_{\beta \in \Omega} \bar{D}_\beta(V, WH).$$

Note that we use  $\bar{D}_\beta(\cdot, \cdot)$ , not  $D_\beta(\cdot, \cdot)$ .

We then propose to use Lagrange duality to judiciously optimize for a set of weights to be used within the framework of the weighted-sum approach, that is, we minimize a single objective function which is a weighted sum of the all objective functions. We design a simple algorithm using multiplicative updates to minimize this weighted sum. We show how this can be used to find distributionally robust NMF (DR-NMF) solutions, that is, solutions that minimize the largest error among all objectives. We illustrate the effectiveness of this approach on synthetic, document and audio datasets. The results show that DR-NMF is robust to our in-cognizance of the noise model of the NMF problem.

## 1.10 Standard optimization schemes

In this section we present the state-of-the-art optimization schemes to tackle the standard optimization problems associated to NMF models that we introduced in Section 1.4. To the best of our knowledge, most algorithms developed to solve the NMF optimization problems are based on iterative local optimization schemes converging to local solutions. Some elementary notions of non-linear programming are used to characterize these algorithms and we refer the reader to Appendix 3 for a brief introduction to these concepts starting with the first-order optimality conditions for Problem 1.4.1.

For optimization problems associated to NMF models, it is usually easier to optimize over one matrix (say  $H$ ) given the other matrix (say  $W$ ) is known and fixed. Indeed, for several scalar metrics  $d(V_{fn}|[WH]_{fn})$ , the resulting error measure  $D(V|WH)$  (the objective function in Problem 1.4.1) is even convex separately w.r.t.  $H$  and w.r.t.  $W$  for  $\beta \in [1, 2]$ , but not w.r.t.  $\{W, H\}$  jointly. For this reason many state-of-the-art NMF optimization algorithms rely on a two-block coordinate descent (BCD) scheme by optimizing alternatively over  $W$  for  $H$  fixed and vice versa; see Algorithm (1). In this thesis, we mostly propose algorithms that rely on the two-block coordinate descent scheme, therefore at each iteration we successively solve two sub-problems; one in  $W$  and the other in  $H$ . Note that in Chapter 6, we propose an algorithm that is able to iteratively solve NMF problems on  $W$  and  $H$  jointly.

---

### Algorithm 1 BCD framework to tackle NMF optimization problems

---

**Require:** Input matrix  $V \in \mathbb{R}_+^{F \times N}$ , the factorization rank  $K$  and number of iterations maxiter.

**Ensure:**  $(W, H)$  is an approximate solution of 1.4.1.

- 1: Initialize  $(W, H)$ .
  - 2: **for**  $k = 1, 2, \dots$ , maxiter **do**
  - 3:     % Update  $W$
  - 4:      $W^k \leftarrow \text{update}(V, W^{k-1}, H^{k-1})$  such that  $D(W^k, H^{k-1}) \leq D(W^{k-1}, H^{k-1})$
  - 5:     % Update  $H$
  - 6:      $H^k \leftarrow \text{update}(V, W^k, H^{k-1})$  such that  $D(W^k, H^k) \leq D(W^k, H^{k-1})$
  - 7: **end for**
- 

In the case we have no additional constraints and penalty terms in the objective functions, then the updates of  $W$  and  $H$  are performed in a similar fashion. The reason is the symmetry of the NMF models since  $V \approx WH$  is equivalent to  $V^T \approx H^T W^T$  and  $D(V|WH) = D(V^T|H^T W^T)$ .

The BCD-based algorithms used to tackle NMF optimization problems mainly differ from the strategy followed to tackle the two sub-problems. In the case  $\beta \in [1, 2]$ , the subproblems are convex w.r.t. variables and therefore, the subproblems can be solved by a large variety of off-the-shelf convex optimization algorithms. Many such BCD algorithms

are summarized in [152, 143, 55, 73]. We list here-under the most popular ones:

- The multiplicative updates (MU); they were proposed by Lee and Seung [86] in 1999 and 2001, and since then they have become the standard update rules for many NMF-based applications. Their main advantage is a low computational complexity. However, it has several drawbacks: it converges slowly compared to accelerated gradient descent methods in the case  $\beta = 2$ ; it cannot modify zero entries ("zero-locking phenomena"); and it is not guaranteed to converge to a stationary point. Note that it can be interpreted as a rescaled gradient descent [56]. Févotte and Idier [45] proposed an efficient general framework (based on the "Majorization-Minimization" framework [129]) to design multiplicative updates to tackle NMF optimization with  $\beta$ -divergences. These multiplicative updates are guaranteed to monotonically decrease the objective function and will be the baseline for algorithms developed in Chapters 3 and 4. To the best of our knowledge, multiplicative updates show good rates of convergence to tackle NMF optimization problems for  $\beta < 2$ .
- Active set methods; these methods are used to solve exactly the two subproblems in the case  $\beta = 2$ , see [79] and the references therein. The state-of-the art algorithm using active-set is the so-called "Alternating Nonnegative Least Squares" (ANLS), see [58]. ANLS is guaranteed to converge to a stationary point [68].
- The exact coordinate descent method in the case  $\beta = 2$ ; the most popular algorithm for NMF is the so-called "Hierarchical alternating least squares" that updates one column  $w$  of  $W$  (resp.  $H$ ) at the time. The optimal solutions of the corresponding subproblems can be written in closed form. HALS was proposed in [[23], pp.161-170] and rediscovered in [29]. HALS was observed to converge much faster than the MU while having similar computational cost. Furthermore, HALS is, under some mild assumptions, guaranteed to converge to a stationary point; see [58] and the references therein.
- The projected gradient method [94]; these methods tackle sub-problems using projected gradient steps. Gradient methods belong to the family of line-search methods (the gradient being one possible direction of descent). Such methods recently regained interests, the reason is due to the advances in first-order optimization (exploit information on values and gradients/subgradients (but not Hessians) of the objective function) [15, 108].

Finally, one of the recently developed algorithm, referred to as "Alternating Optimization-Alternating Direction for Multiplier" (AO-ADMM), employs ADMM methods for solving the subproblems in  $W$  and  $H$ . The main advantage lies in the fact ADMM can easily handle a large variety of regularizations and constraints simultaneously by the judicious introduction of slack variables [74].



## 2 Minimum-Volume Rank-Deficient Nonnegative Matrix Factorizations

In recent years, nonnegative matrix factorization (NMF) with volume regularization has been shown to be a powerful approach to identify the latent factors that generated the data; for example for hyperspectral unmixing, document classification, community detection and hidden Markov models. In this chapter, we show that minimum-volume NMF can also be used when the basis matrix is rank deficient, which is a reasonable scenario for some real-world NMF problems (e.g., for unmixing multispectral images). We propose an alternating fast projected gradient method for minimum-volume NMF and illustrate its use on rank-deficient NMF problems; namely a synthetic data set and a multispectral image.

The content of this chapter is extracted from: [89] V. Leplat, A.M.S. Ang, and N. Gillis. Minimum-volume rank-deficient nonnegative matrix factorizations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3402-3406. IEEE, 2019.

### 2.1 Introduction

As introduced in Section 1.4, computing an NMF requires to find two nonnegative matrices  $W \in \mathbb{R}_+^{F \times K}$  and  $H \in \mathbb{R}_+^{K \times N}$  for a given nonnegative matrix  $V \in \mathbb{R}_+^{F \times N}$  and a factorization rank  $K$  such that  $V \approx WH$ . In this chapter;

- we mostly deal with noisy scenarios and therefore we are searching for an approximate decomposition.
- We will use the Frobenius norm of the residual matrix  $V - WH$  for the error measurement  $D(V|WH)$  in Problem 1.4.1. Due to its simplicity and its L-smooth property, Frobenius norm is arguably the most widely used to assess the error of an NMF solution.

Problem 1.4.1 now becomes:

$$\min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} \|V - WH\|_F^2 \text{ s.t. } W \geq 0 \text{ and } H \geq 0.$$

In Section 1.8, it has been shown that NMF is in most cases ill-posed because the optimal solution is not unique. In order to make the solution of the above problem unique (up to permutation and scaling of the columns of  $W$  and rows of  $H$ ) hence making the problem well-posed and the parameters  $(W, H)$  of the problem identifiable, a key idea is to look for

a solution  $W$  with minimum volume; see Section 1.8.4 for more details. A possible problem formulation for minimum-volume NMF is as follows

$$\min_{W \geq 0, H(:,n) \in \mathcal{S}^K \forall n} \|V - WH\|_F^2 + \lambda \text{vol}(W), \quad (2.1)$$

where  $\mathcal{S}^K = \{x \in \mathbb{R}_+^K \mid \sum_k x_k \leq 1\}$ ,  $\lambda$  is a penalty parameter, and  $\text{vol}(W)$  is a function that measures the volume of the columns of  $W$ . Let us recall that  $H$  needs to be normalized otherwise  $W$  would go to zero since  $WH = (cW)(H/c)$  for any  $c > 0$ . In this chapter, we will use  $\text{vol}(W) = \log \det(W^T W + \delta I)$ , where  $I$  is the identity matrix of appropriate dimensions.

The reason for using such a measure is twofold (i)  $\sqrt{\det(W^T W)}/K!$  is the volume of the convex hull of the columns of  $W$  and the origin, see Section 1.8.4 and (ii) it has been shown to be one of the most efficient volume regularization [7]. Under some appropriate conditions on  $V = WH$ , this model will provably recover the true underlying  $(W, H)$  that generated  $V$ . These recovery conditions require that the columns of  $V$  are sufficiently well spread in the convex hull generated by the columns of  $W$ ; this is the so-called sufficiently scattered condition. In particular, data points need to be located on the facets of this convex hull hence  $H$  needs to be sufficiently sparse, see Section 1.8.2 for more details. Let us remark that, as far as we know, these theoretical results only apply in noiseless conditions hence robustness to noise of problem (2.1) associated to approximate NMF models still needs to be rigorously analyzed (this is a very promising but difficult direction of further research).

Another key assumption that is used in minimum-volume NMF is that the basis matrix  $W$  is full rank, that is,  $\text{rank}(W) = K$ ; otherwise  $\det(W^T W) = 0$ . However, there are situations when the matrix  $W$  is not full rank: this happens in particular when  $\text{rank}(V) \neq \text{rank}_+(V)$ . Here is a simple example:

$$V = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad (2.2)$$

for which  $\text{rank}(V) = 3 < \text{rank}_+(V) = 4$ . Indeed, given an exact NMF decomposition  $V = WH$  under the assumption  $K = \text{rank}_+(V)$ , as per equation (1.6) we have the following inequalities for the ranks:  $\text{rank}(V) \leq \text{rank}(W) \leq \min(F, K) = K = \text{rank}_+(V)$  in the case  $F \geq K$ .

It is then clear that if  $\text{rank}(V) \neq \text{rank}_+(V)$ , i.e.  $\text{rank}(V) < \text{rank}_+(V)$ ,  $W$  may not be full column rank. Furthermore, if  $\text{rank}(V) = \text{rank}_+(V)$ , then we have  $\text{rank}(W) = K$ ;  $W$  is therefore full column rank. The columns of the matrix  $V$  defined in (2.2) are the vertices of a square in a 2-dimensional subspace; see Figure 2.2 for an illustration.

A practical situation where this could also happen is in multispectral imaging. Let us construct the matrix  $V$  such that each column  $V(:, n) \geq 0$  is the spectral signature of a pixel. As explained in Section 1.5.1, under the linear mixing model, each column of  $V$  is the nonnegative linear combination of the spectral signatures of the constitutive materials

present in the image, referred to as endmembers: we have  $V(:, n) = \sum_{k=1}^K W(:, k)H(k, n)$ , where  $W(:, k)$  is the spectral signature of the  $k$ th endmember, and  $H(k, n)$  is the abundance of the  $k$ th endmember in the  $n$ th pixel. For multispectral images, the number of materials within the scene being imaged can be larger than the number of spectral bands meaning that  $K > F$  hence  $\text{rank}(W) \leq F < K$ .

In this chapter, we focus on the rank-deficient scenario for minimum-volume NMF, that is, when  $\text{rank}(W) < K$ . The main contribution of this chapter is three-fold: (i) We explain why optimization problem (2.1) can be used meaningfully when the basis matrix  $W$  is not full rank. This is, as far as we know, the first time this observation is made in the literature. (ii) We propose an algorithm based on alternating projected fast gradient method to tackle this problem. (iii) We illustrate our results on a noisy synthetic data set and a real-life multispectral image.

## 2.2 Minimum-volume NMF in the rank-deficient case

Let us discuss in more details the optimization problem that we consider in this chapter, namely,

$$\min_{W \geq 0, H(:, n) \in \mathcal{S}^K \forall n} \|V - WH\|_F^2 + \lambda \log \det(W^T W + \delta I), \quad (2.3)$$

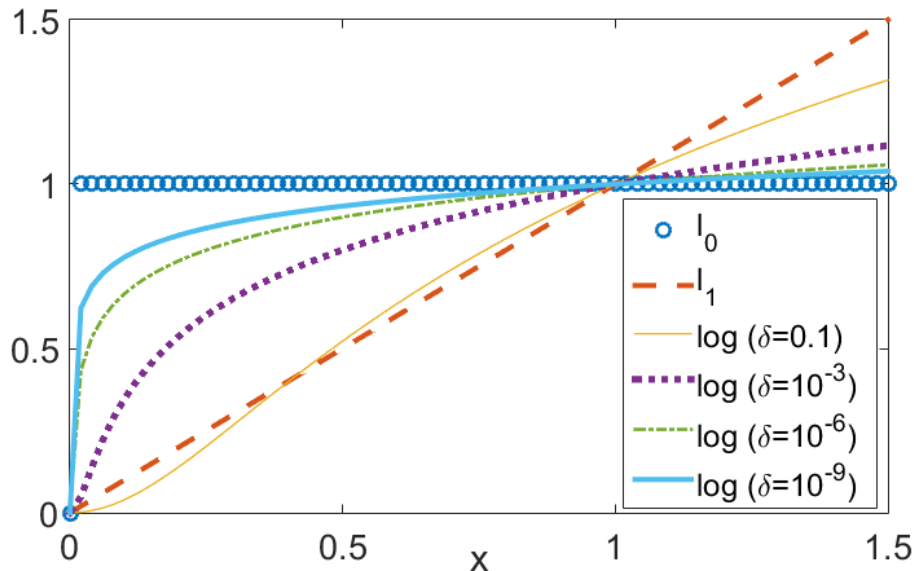
which has three key ingredients: the choice of the volume regularizer, that is,  $\log \det(W^T W + \delta I)$ , the parameters  $\delta$  and  $\lambda$ . They are discussed in the next three paragraphs.

**Choice of the volume regularizer** Most functions used to minimize the volume of the columns of  $W$  are based on the Gram matrix  $W^T W$ ; in particular,  $\det(W^T W)$  and  $\log \det(W^T W + \delta I)$  for some  $\delta > 0$  are the most widely used measures; see, e.g., [102, 52]. Note that  $\det(W^T W) = \prod_{k=1}^K \sigma_k^2(W)$ , hence the log term allows to weight down large singular values and has been observed to work better in practice; see, e.g., [9]. When  $W$  is rank deficient (that is,  $\text{rank}(W) < K$ ), some singular values of  $W$  are equal to zero hence  $\det(W^T W) = 0$ . Therefore, the function  $\det(W^T W)$  cannot distinguish between different rank-deficient solutions<sup>1</sup>. However, we have  $\log \det(W^T W + \delta I) = \sum_{k=1}^K \log(\sigma_k^2(W) + \delta)$ . Hence if  $W$  has one (or more) singular value equal to zero, this measure still makes sense: among two rank-deficient solutions belonging to the same low-dimensional subspace, minimizing  $\log \det(W^T W + \delta I)$  will favor a solution whose convex hull has a smaller volume within that subspace since decreasing the non-zero singular values of  $(W^T W + \delta I)$  will decrease  $\log \det(W^T W + \delta I)$ . In mathematical terms, let  $W \in \mathbb{R}^{F \times K}$  belong to a  $r$ -dimensional subspace with  $r < K$  so that  $W = US$  where  $U \in \mathbb{R}^{F \times r}$  is an orthogonal basis of that subspace and  $S \in \mathbb{R}^{r \times K}$  are the coordinates of the columns of  $W$  in that subspace. Then,  $\log \det(W^T W + \delta I) = \sum_{k=1}^r \log(\sigma_k^2(S) + \delta) + (K - r) \log(\delta)$ . The minimum-volume criterion  $\log \det(W^T W + \delta I)$  with  $\delta > 0$  is therefore meaningful even when  $W$  does not have rank  $K$ .

<sup>1</sup>Of course, one could also use the measure  $\det(W^T W + \delta I)$  meaningfully in the rank-deficient case.

However, it would be numerically more challenging since for each singular value of  $W$  equal to zero, the objective is multiplied by  $\delta$  which should be chosen relatively small.

**Choice of  $\delta$**  The function  $\log\det(W^T W + \delta I)$  which is equal to  $\sum_{k=1}^K \log(\sigma_k^2(W) + \delta)$  is a non-convex surrogate for the  $\ell_0$  norm of the vector of singular values of  $W$  (up to constants factors), that is, of  $\text{rank}(W)$  [41, 42]. It is sharper than the  $\ell_1$  norm of the vector of singular values (that is, the nuclear norm) for  $\delta$  sufficiently small; see Fig. 2.1. Therefore, if one wants to promote rank-deficient solutions,  $\delta$  should not be chosen too large, say  $\delta \leq 0.1$ . Moreover,  $\delta$  should not be chosen too small otherwise  $W W^T + \delta I$



**Fig. 2.1.** Function  $\frac{\log(x^2 + \delta) - \log(\delta)}{\log(1 + \delta) - \log(\delta)}$  for different values of  $\delta$ ,  $\ell_1$  norm ( $= |x|$ ) and  $\ell_0$  norm ( $= 0$  for  $x = 0$ ,  $= 1$  otherwise).

might be badly conditioned which makes the optimization problem harder to solve (see Section 2.3) –also, this could give too much importance to zero singular values which might not be desirable. Therefore, in practice, we recommend to use a value of  $\delta$  between 0.1 and  $10^{-3}$ . We will use  $\delta = 0.1$  in this chapter. Note that in previous works,  $\delta$  was chosen very small (e.g.,  $10^{-8}$  in [52]) which, as explained above, is not a desirable choice, at least in the rank-deficient case. Even in the full-rank case, we argue that choosing  $\delta$  too small is also not desirable since it promotes rank-deficient solutions.

**Choice of  $\lambda$**  The choice of  $\delta$  will influence the choice of  $\lambda$ . In fact, the smaller  $\delta$ , the larger  $|\log\det(\delta)|$ , hence to balance the two terms in the objective (2.3),  $\lambda$  should be smaller. For the practical implementation, we will initialize  $W^{(0)} = V(:, \mathcal{K})$  where  $\mathcal{K}$  is computed with the successive nonnegative projection algorithm (SNPA) that can handle the rank-deficient separable NMF problem [57]. Note that SNPA also provides the matrix  $H^{(0)}$  so as to minimize  $\|V - W^{(0)} H^{(0)}\|_F^2$  while  $H^{(0)}(:, n) \in \mathcal{S}^K$  for all  $n$ . Finally, we will choose

$$\lambda = \tilde{\lambda} \frac{\|V - W^{(0)} H^{(0)}\|_F^2}{|\log\det(W^{(0)T} W^{(0)} + \delta I)|},$$

where we recommend to choose  $\tilde{\lambda}$  between 1 and  $10^{-3}$  depending on the noise level (the noisier the input matrix, the larger  $\lambda$  should be).

### 2.3 Algorithm for minimum-volume NMF

Most algorithms for NMF optimize alternatively over  $W$  and  $H$ , and we adopt this strategy in this chapter. For the update of  $H$ , we will use the projected fast gradient method (PFGM) from [57]. Note that, as opposed to previously proposed methods for minimum-volume NMF, we assume that the sum of the entries of each column of  $H$  is smaller or equal to one, not equal to one, which is more general. For the update of  $W$ , we use a PFGM applied on an strongly convex upper approximation of the objective function; similarly as done in [52]—although in that paper, authors did not consider explicitly the case  $W \geq 0$  ( $W$  is unconstrained in their model) and did not write down explicitly a PFGM taking advantage of strong convexity. For the sake of completeness, we briefly recall this approach. The following upper bound for the logdet term holds: for any  $Q > 0$  and  $S > 0$ , we have

$$\begin{aligned} \log\det(Q) &\leq g(Q, S) = \log\det(S) + \text{Tr}(S^{-1}(Q - S)) \\ &= \text{Tr}(S^{-1}Q) + \log\det(S) - K. \end{aligned}$$

This follows from the concavity of  $\log\det(\cdot)$  as  $g(Q, S)$  is the first-order Taylor approximation of  $\log\det(Q)$  around  $S$ —it has also been used for example in [150]. This gives  $\log\det(W^T W + \delta I) \leq \text{Tr}(Y W^T W) + \log\det(Y^{-1}) - K$  for any  $W$  and any  $Y = (Z^T Z + \delta I)^{-1}$  with  $\delta > 0$ . Plugging this in the original objective function, and denoting  $w_i^T$  the  $i$ th row of matrix  $W$  and  $\langle \cdot, \cdot \rangle$  is the Frobenius inner product of two matrices, we obtain

$$\begin{aligned} \ell(W) &= \|V - WH\|_F^2 + \lambda \log\det(W^T W + \delta I) \\ &= \|V\|_F^2 - 2\langle V H^T, W \rangle + \langle W^T W, H H^T \rangle \\ &\quad + \lambda \log\det(W^T W + \delta I) \\ &\leq \langle W^T W, H H^T + \lambda Y \rangle - 2\langle C, W \rangle + b \\ &= 2 \sum_{i=1}^n \left( \frac{1}{2} w_i^T A w_i - c_i^T w_i \right) + b = \bar{\ell}(W), \end{aligned}$$

where  $Y = (Z^T Z + \delta I)^{-1}$  and  $A = H H^T + \lambda Y$  are positive definite for  $\delta, \lambda > 0$ ,  $C = V H^T$ , and  $b$  is a constant independent of  $W$ . Note that  $\bar{\ell}(W) = \ell(W)$  for  $Z = W$ . Minimizing the upper bound  $\bar{\ell}(W)$  of  $\ell(W)$  requires to solve  $m$  independent strongly convex optimization problems with Hessian matrix  $A$ . Using PFGM on this problem, we obtain a linear convergence method with rate  $1 - \sqrt{\kappa^{-1}}$  where  $\kappa$  is the condition number of  $A$  [107]. Note that the subproblem in variable  $H$  is not strongly convex when  $W$  is rank deficient in which case PFGM converges sublinearly, in  $O(1/k^2)$  where  $k$  is the iteration number. In any case, PFGM is an optimal first-order method in both cases [107], that is, no first-order method can have a faster convergence rate. When  $W$  is rank deficient, we have  $\frac{\lambda}{\delta} \leq L = \lambda_{\max}(A) \leq \|H\|_2^2 + \frac{\lambda}{\delta}$ , where  $L$  is the largest eigenvalue of  $A$ . This shows the importance of not choosing  $\delta$  too small, since the smaller  $\delta$ , the larger the conditioning of  $A$  hence the slower will be the PFGM. Note that  $L$  is the Lipschitz constant of the gradient of the objective function and controls the stepsize which is equal to  $1/L$ . Our proposed algorithm is summarized in Algorithm 2. We will use 10 inner iterations for the PFGM on  $W$  and  $H$ .

---

**Algorithm 2** Min-vol NMF using alternating PFGM

---

**Require:** Input matrix  $V \in \mathbb{R}_+^{F \times N}$ , the factorization rank  $K$ ,  $\delta > 0$ ,  $\tilde{\lambda} > 0$ , number of iterations maxiter.

**Ensure:**  $(W, H)$  is an approximate solution of (2.3).

- 1: Initialize  $(W, H)$  using SNPA [57].
  - 2: Let  $\lambda = \tilde{\lambda} \frac{\|V - WH\|_F^2}{\log \det(W^T W + \delta I)}$ .
  - 3: **for**  $k = 1, 2, \dots$ , maxiter **do**
  - 4:     % Update  $W$
  - 5:     Let  $A = HH^T + \lambda(W^T W + \delta I)^{-1}$  and  $C = VH^T$ .
  - 6:     Perform a few steps of PFGM on the problem  $\min_{U \geq 0} \frac{1}{2} \langle U^T U, A \rangle - \langle U, C \rangle$ , with initialization  $U = W$ . Set  $W$  as last iterate.
  - 7:     % Update  $H$
  - 8:     Perform a few steps of PFGM on the problem  $\min_{H(:,n) \in \mathcal{S}^K \forall n} \|V - WH\|_F^2$  as in [57].
  - 9: **end for**
- 

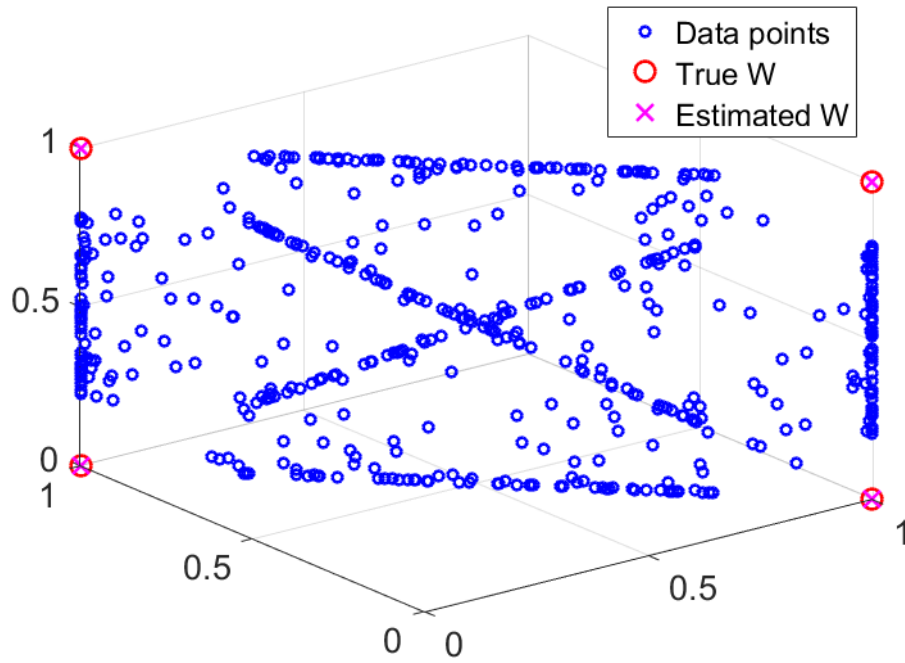
## 2.4 Numerical Experiments

We now apply our method on a synthetic and a real-world data set. All tests are performed using Matlab R2015a on a laptop Intel CORE i7-7500U CPU @2.9GHz 24GB RAM. The code is available from <http://bit.ly/minvolNMF>.

**Synthetic data set.** Let us construct the matrix  $V \in \mathbb{R}^{4 \times 500}$  as follows:  $W$  is taken as the matrix from (2.2):

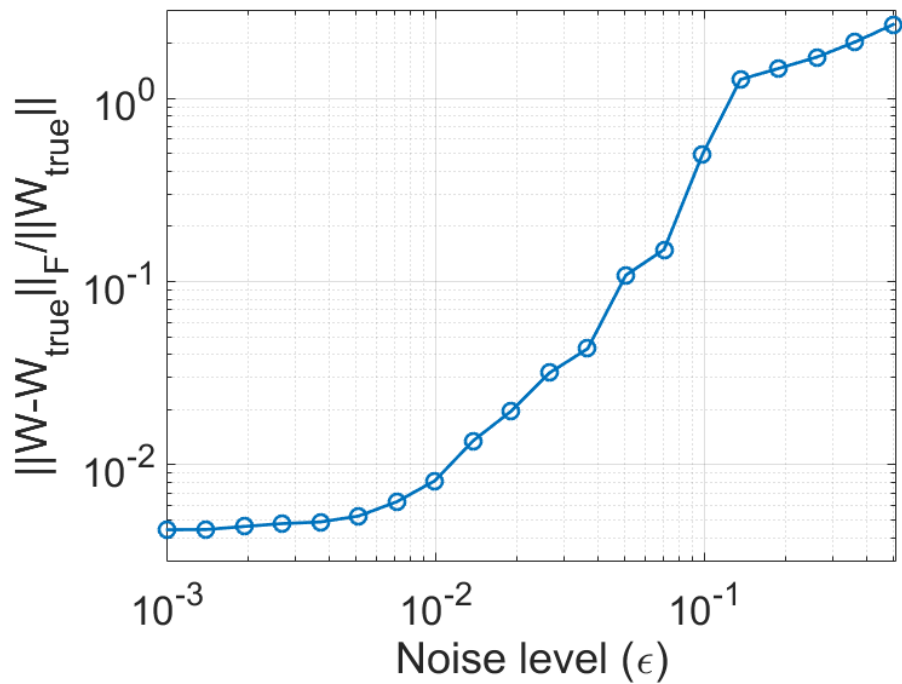
$$W = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

so that  $\text{rank}(W) = 3 < K = 4$ , and each column of  $H$  is distributed using the Dirichlet distribution of parameter  $(0.1, \dots, 0.1)$ . Each column of  $H$  with an entry larger 0.8 is resampled as long as this condition does not hold. This guarantees that no data point is close to a column of  $W$  (this is sometimes referred to as the purity index). Figure 2.2 illustrates this geometric problem. As observed on Figure 2.2, Algorithm 2 is able to perfectly recover the true columns of  $W$ . For this experiment, we use  $\tilde{\lambda} = 0.01$ . Figure 2.3 illustrates the same experiment where noise is added to  $V = \max(0, WH + N)$  where  $N = \epsilon \text{randn}(F, N)$  in Matlab notation (i.i.d. Gaussian distribution of mean zero and standard deviation  $\epsilon$ ). Note that the average of the entries of  $V$  is 0.5 (each column is a linear combination of the columns of  $W$ , with weights summing to one). Figure 2.3 displays the average over 20 randomly generated matrices  $V$  of the relative error  $d(W, \tilde{W}) = \frac{\|W - \tilde{W}\|_F}{\|W\|_F}$  where  $\tilde{W}$  is the solution computed by Algorithm 2 depending on the noise level  $\epsilon$ . This



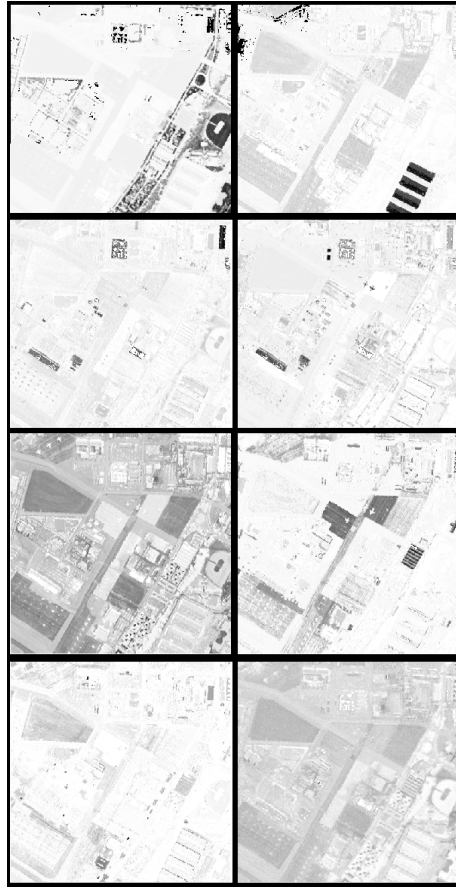
**Fig. 2.2.** Synthetic data set and recovery. (Only the first three entries of each four-dimensional vector are displayed.)

illustrates that problem (2.1) is robust against noise since the  $d(W, \tilde{W})$  is smaller than 1% for  $\epsilon \leq 1\%$ .



**Fig. 2.3.** Evolution of the recovery of the true  $W$  depending on the noise  $N = \epsilon \text{ rand}(F, N)$  using Algorithm 2 ( $\tilde{\lambda} = 0.01$ ,  $\delta = 0.1$ ,  $\text{maxiter} = 100$ ).

**Multispectral image.** The San Diego airport is a HYDICE hyperspectral image (HSI) containing 158 clean bands, and  $400 \times 400$  pixels for each spectral image; see, e.g., [63]. There are mainly three types of materials: road surfaces, roofs and vegetation (trees and grass). The image can be well approximated using  $K=8$ . Since we are interested in the case  $\text{rank}(W) < K$ , we select  $F=5$  spectral band using the successive projection algorithm [65] (this is essentially Gram-Schmidt with column pivoting) applied on  $V^T$ . This provides bands that are representative: the selected bands are 4, 32, 116, 128, 150. Hence, we are factoring a 5-by-160000 matrix using a  $F=8$ . Note that we have removed outlying pixels (some spectra contain large negative entries while others have a norm order of magnitude larger than most pixels). Figure 2.4 displays the abundance maps extracted (that is, the rows of matrix  $H$ ): they correspond to meaningful locations of materials. Here we have used  $\tilde{\lambda}=0.1$  and 1000 iterations. From the initial solution provided by SNPA, Algorithm 2



**Fig. 2.4.** Abundance maps extract by Algorithm 2 using only five bands of the San Diego airport HSI. From left to right, top to bottom: vegetation (grass and trees), three different types of roof tops, four different types of road surfaces.

is able to reduce the error  $\|V - WH\|_F$  by a factor of 11.7 while the term  $\log \det(W^T W + \delta I)$  only increases by a factor of 1.06. The final relative error is  $\frac{\|V - WH\|_F}{\|X\|_F} = 0.2\%$ .



## 2.5 Faster Algorithm for minimum-volume NMF

This section gives preliminary results on the development and the test of a new algorithm to solve the following minimum-volume NMF problem:

$$\min_{W \geq 0, H(:,n) \in \mathcal{S}^K \forall n} \|V - WH\|_F^2 + \lambda \log \det(W^T W + \delta I). \quad (2.4)$$

In Section 2.3, we use the alternating approach to compute a numerical solution  $(W, H)$  for the problem (2.4). In particular, for the update of  $W$ , we used PFGM applied on a strongly convex upper approximation of the objective function; the logdet term is replaced by its first-order Taylor approximation. For the new algorithm, referred to as fast min-vol NMF, we directly apply PFGM on the objective function of the minimization sub-problem in  $W$ . For the update of  $H$ , we still apply the optimal fast gradient method using strong convexity of the objective function.

The acceleration in PFGM scheme is made by adding extrapolation step after the projected gradient descent step. For instance, the optimization over  $W$  is performed as follows:

$$\begin{aligned} \text{Gradient step} \quad W^{k+1} &= \left[ Y^k - \alpha_k^Y \nabla_Y f(Y^k, H) \right]_+ \\ \text{Extrapolation step} \quad Y^{k+1} &= W^{k+1} + \beta_k (W^{k+1} - W^k) \end{aligned} \quad (2.5)$$

where  $Y$  is the pairing variable of  $W$ ,  $\alpha_k^Y$  is the step size and  $[\cdot]_+ = \max(\cdot, 0)$  is the projection operator onto the feasible set (the non-negative orthant). The expression of the gradient of the objective function w.r.t.  $W$  has been determined as follows: we first focus on computing the gradient of  $\log \det(W^T W + \delta I)$  w.r.t.  $W$  by using the following formulas:

$$\begin{aligned} \frac{\partial}{\partial W} \text{trace}(AW) &= A^T, \quad \frac{\partial}{\partial W} \log \det(W) = W \\ \frac{\partial}{\partial W} \log \det(F(W)) &= \frac{\partial}{\partial W} \text{trace}(F^{-1}(Z)F(W))|_{Z=W} \\ \frac{\partial}{\partial W} \text{trace}(G(W)H(W)) &= \frac{\partial}{\partial W} \text{trace}(H(Z)G(W) + H(W)G(Z))|_{Z=W} \end{aligned}$$

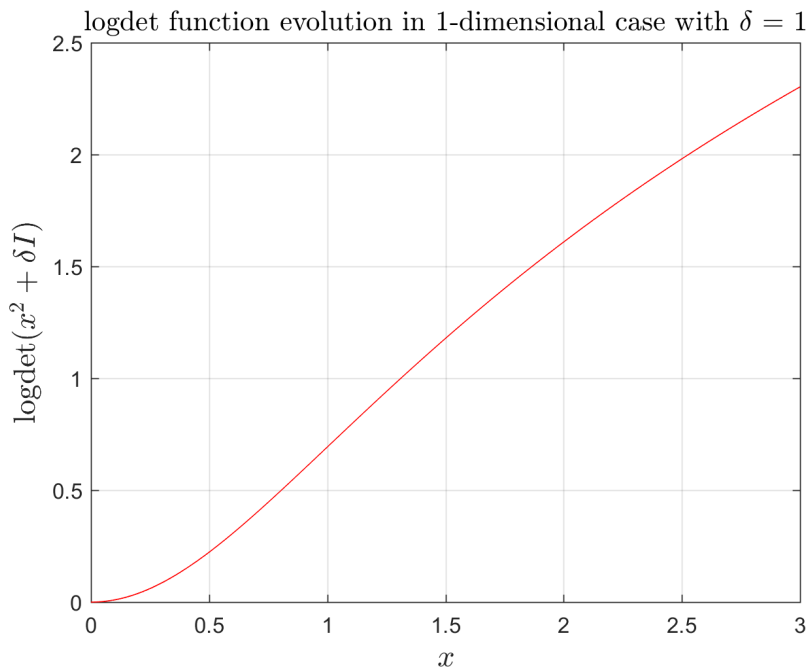
Therefore we obtain:

$$\begin{aligned} \frac{\partial}{\partial W} \log \det \left( \underbrace{W^T W + \delta I}_{F(W)} \right) &= \frac{\partial}{\partial W} \text{trace}(F^{-1}(Z)F(W))|_{Z=W} \\ &= \frac{\partial}{\partial W} \text{trace} \left( (Z^T Z + \delta I)^{-1} (W^T W + \delta I) \right) |_{Z=W} \\ &= \frac{\partial}{\partial W} \text{trace} \left( \underbrace{(Z^T Z + \delta I)^{-1} W^T}_{G(W)} \underbrace{W}_{H(W)} \right) |_{Z=W} \\ &= \frac{\partial}{\partial W} \text{trace} \left( \left[ Z (Z^T Z + \delta I)^{-1} + Z (Z^T Z + \delta I)^{-T} \right] W^T \right) |_{Z=W} \\ &= \left[ Z (Z^T Z + \delta I)^{-1} + Z (Z^T Z + \delta I)^{-T} \right]_{Z=W} \\ &= W \left[ (W^T W + \delta I)^{-1} + (W^T W + \delta I)^{-T} \right]. \end{aligned}$$

As  $(W^T W + \delta I)^{-1}$  is symmetric, then  $\frac{\partial}{\partial W} \log \det (W^T W + \delta I) = 2W (W^T W + \delta I)^{-1}$ . We finally derive the gradient of the objective function denoted  $f(W)$  (for the sub-problem in  $W$ ) w.r.t.  $W$  as follows:

$$\nabla_W f(W, H) = 2(WH - V)H^T + 2\lambda W (W^T W + \delta I)^{-1}.$$

The critical parts of this scheme are the step size and the extrapolation parameter  $\beta_k$ . The parameter  $\beta_k$  is adapted over iterations of the scheme and is always within the range  $[0, 1]$ . In the case  $\beta_k = 0$ , the scheme (2.5) reduces to the plain projected gradient scheme. In convex non-linear programming,  $\beta$  has a closed-form expression and the resulting scheme has optimal rate of convergence [107]. However, if the problem to solve is non-convex, it is not known how to determine  $\beta$ . For problem (2.4), let us first note that the logdet function is nor convex nor concave as it can be observed on Figure 2.5 in the 1-dimensional case. It is then clear that the objective function for the sub-problem in  $W$  is also nor convex

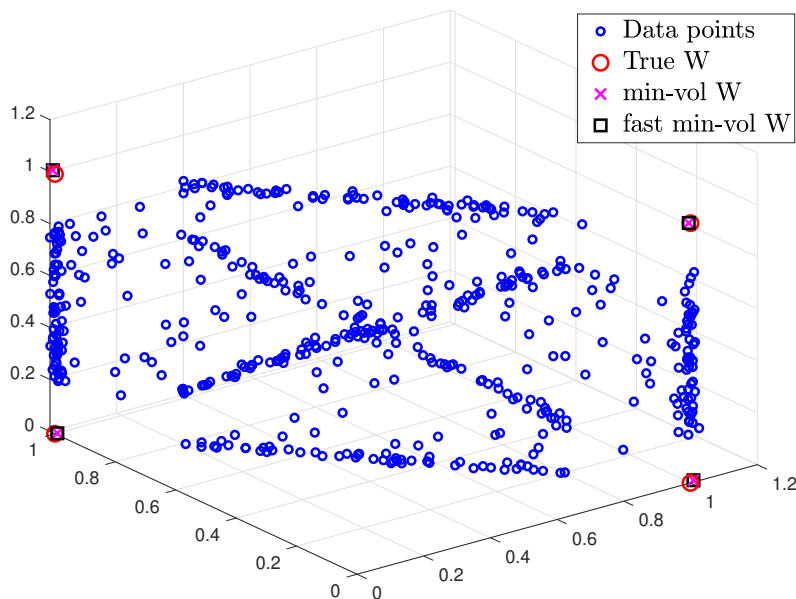


**Fig. 2.5.** logdet function evolution in the 1-dimensional case with  $\delta = 1$

nor concave when  $\lambda > 0$ . Therefore, we do not have theoretically grounded closed-form formula to tune  $\beta_k$ , we have followed the approach developed in [8] to numerically tune  $\beta_k$ . The step size  $\alpha_k$  is also tuned numerically with a similar approach than  $\beta_k$ . In short, the idea is to change the value of  $\alpha_k$  based on the increase or decrease of the objective function. When the objective function of the sub-problem in  $W$  increases,  $\alpha_k$  is decreased, otherwise  $\alpha_k$  is slightly increased.

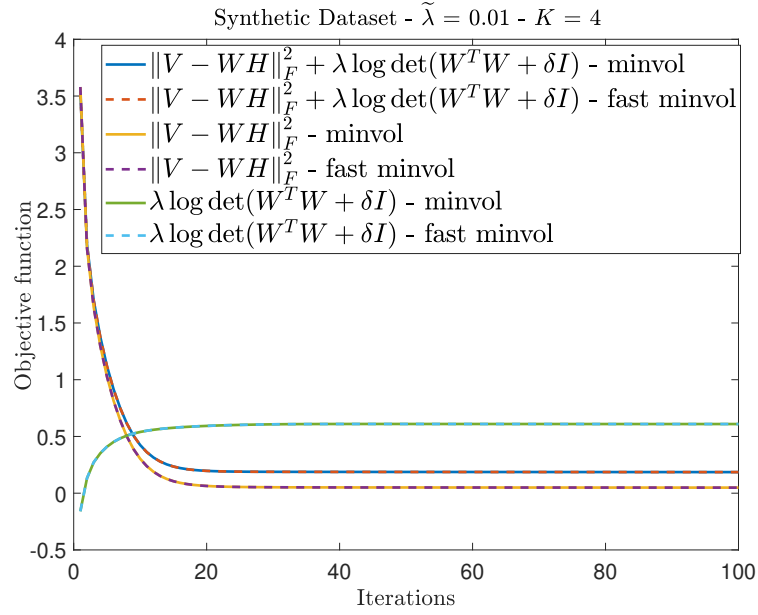
We have tested this scheme, denoted fast-min vol, on the data sets considered in Section 2.4.

**Synthetic data set.** Figure 2.6 illustrates the problem and the solutions obtained with two algorithms. The first algorithm, simply denoted "min-vol", corresponds to Algorithm 2 presented in Section 2.3. The second algorithm is the fast-min vol algorithm presented above. Figure 2.7 displays the evolution of the objective functions as the Frobenius norm and the  $\lambda \log \det$  term over iterations for both algorithms. Figure 2.6 depicts the estimated  $W$  for both algorithms as the ground truth  $W$ . As it can be observed, both algorithms are able to perfectly recover the true columns of  $W$  and the rates of convergence of the objective function are almost identical.



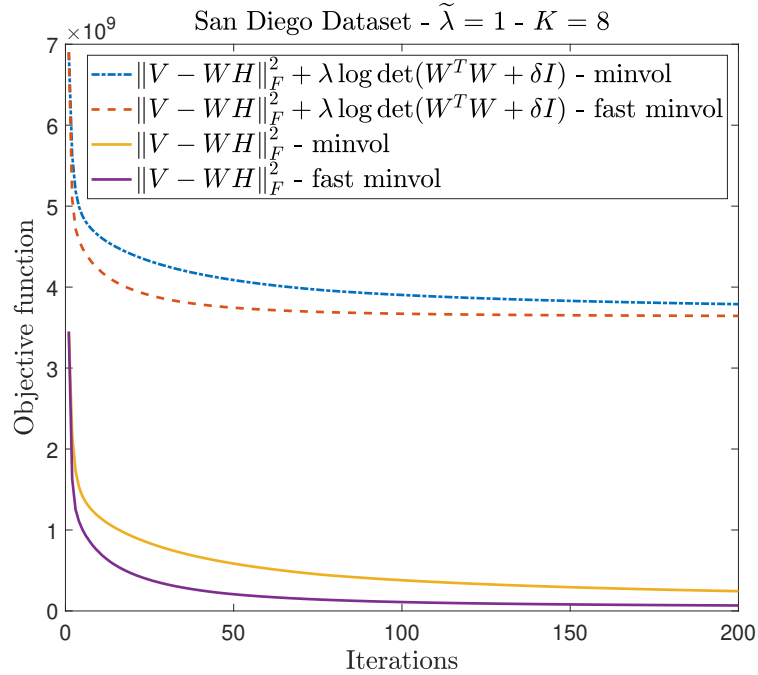
**Fig. 2.6.** Synthetic data set and recovery for "min-vol" and "fast-min-vol" algorithms. (Only the first three entries of each four-dimensional vector are displayed.)

**Multispectral image.** We selected  $F = 5$  spectral bands from the San Diego airport hyperspectral image (HSI) using the successive projection algorithm [65] applied to  $V^T$  and we have chosen a factorization rank  $K = 8$ ; this corresponds to a rank-deficient scenario as  $\text{rank}(W) < K$ . Note that we have removed outlying pixels. Figure 2.8 displays the evolution of the objective functions as the Frobenius norm and the  $\log \det$  term over iterations for both algorithms. Here we have used  $\tilde{\lambda} = 1$  and 200 iterations. From the initial solution provided by SNPA, min-vol algorithm is able to reduce the error  $\|V - WH\|_F^2$  by a factor of 14.1 while the term  $\log \det(W^T W + \delta I)$  only increases by a factor of 1.03. Fast-min-vol algorithm is able to reduce this initial error by a factor of 51.2 while the term  $\log \det(W^T W + \delta I)$  only increases by a factor of 1.04. Furthermore, the final error for the objective function obtained with the fast-min-vol algorithm is approximately half the one obtained with min-vol algorithm. For this scenario, fast-min vol algorithm clearly outperforms min-vol algorithm. Similar conclusions have been derived when both algorithms are



**Fig. 2.7.** Rates of convergence for "min-vol" and "fast-min-vol" algorithms on synthetic data set

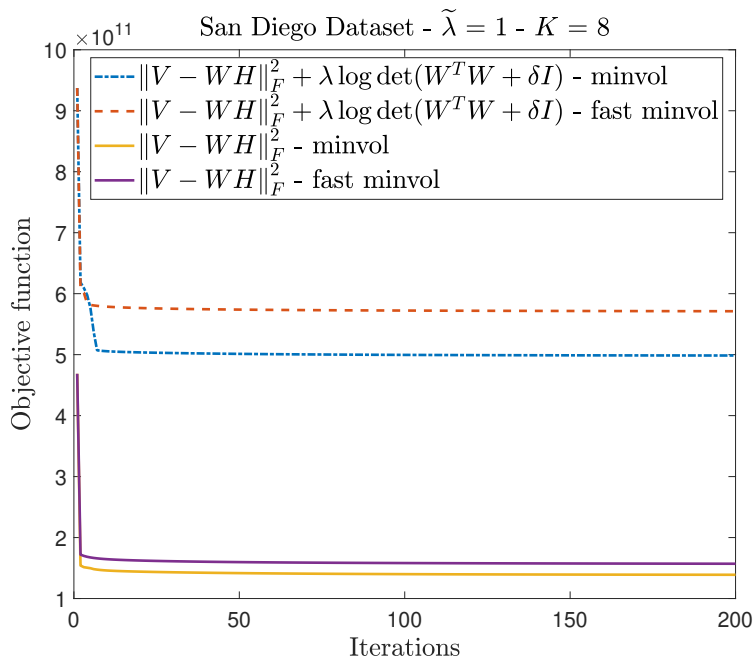
compared on other famous data sets for HSI such as "Urban" and "Jasper".



**Fig. 2.8.** Rates of convergence for the objective function of problem (2.4) and the Frobenius norm along iterations for fast-min vol and min-vol algorithms in the rank deficient case for San Diego Dataset.

**Hyperspectral image.** We repeated the previous experiment without selecting a subset of spectral bands, i.e. the data matrix to factorize is the original San Diego HSI. In this context, the solution for  $W$  is most likely not rank-deficient and we are dealing with a

higher-dimensional problem. As it can be observed in Figure 2.9, min vol algorithm gives a better solution than the fast-min vol algorithm in terms of final errors. It appears that fast-min vol algorithm is more quickly stuck in a stationary point after few iterations.



**Fig. 2.9.** Rates of convergence for the objective function of problem (2.4) and the Frobenius norm along iterations for fast-min vol and min-vol algorithms in the high-dimensional case for San Diego Dataset.

Our claim is that this is due to the explosion of the number of saddle points for a non-convex objective function when the dimension of the problem increases. Fast-min vol uses the gradient of the objective function of (2.4) while min-vol algorithm minimizes an auxiliary function built at each iteration (first-order Taylor approximation which boils down to a trace function of  $W$ ), fast-min vol seems to be more sensitive to the saddle points. For some data sets, we can show that the min-vol algorithm with a fast-min vol initialization can escape from the current solution which tends to support our claim. Similar conclusions have been derived with random initialization for  $W$  and  $H$  instead of using SNPA.

Further works will try to tackle this issue observed for the high dimensional case and allows fast-min vol to escape from saddle points.

## 2.6 Conclusions

In this chapter, we have shown that minimum-volume NMF can be used meaningfully for the rank-deficient scenario. We have provided a simple algorithm to tackle this problem and have illustrated the behaviour of the method on synthetic and real-world data sets. We have developed and tested a faster algorithm referred to as fast-min vol. We have shown that fast-min vol outperforms min-vol algorithm (Algorithm 2) in low-dimensional

setting but tends to be more easily stuck in saddle points as the dimension of the problem increases. This work is only preliminary and many important questions remain open; in particular

- Under which conditions can we prove the identifiability of models and associated optimization problems for minimum-volume NMF in the rank-deficient case (as done in [95, 53] for the full-rank case)? Intuitively, it seems that a condition similar to the sufficiently-scattered condition would be sufficient but this has to be analysed thoroughly.
- Can we prove robustness to noise of such techniques? (The question is also open for the full-rank case.)
- Can we design more robust algorithms? And algorithms taking advantage of the fact that the solution is rank-deficient?

### 3 Minimum-Volume $\beta$ -NMF for blind audio source separation

Considering a mixed signal composed of various audio sources and recorded with a single microphone, we consider in this chapter the blind audio source separation problem which consists in isolating and extracting each of the sources, see Section 1.5.2. To perform this task, nonnegative matrix factorization (NMF) based on the Kullback-Leibler and Itakura-Saito  $\beta$ -divergences is a standard and state-of-the-art technique that uses the time-frequency representation of the signal. We present a new NMF model and its associated (optimization) problem, referred to as minimum-volume  $\beta$ -NMF (min-vol  $\beta$ -NMF), better suited for this task. It is based on the minimization of  $\beta$ -divergences along with a penalty term that promotes the columns of the dictionary matrix  $W$  to have a small volume with  $W$  being column stochastic. To the best of our knowledge, this problem is novel in two aspects: (1) it is the first time a minimum-volume penalty is associated with a  $\beta$ -divergence for  $\beta \neq 2$  and it is the first time such problems are used in the context of audio source separation, and (2) as opposed to most previously proposed minimum-volume NMF problems, our problem imposes a normalization constraints on the factor  $W$  instead of  $H$ . As far as we know, the only other references that used a normalization of  $W$  is [153] but the authors did not justify this choice compared to the normalization of  $H$  (the choice seems arbitrary, motivated by the ‘elimination of the norm indeterminacy’), nor provided theoretical guarantees. In this chapter, we explain why normalization of  $W$  is a better choice in practice, and we prove that, under some mild assumptions and in the noiseless case, this problem provably identify the sources; see Theorem 3.2.1. To the best of our knowledge, this is the first result of this type in the audio source separation literature. In Section 3.3 we propose an algorithm to tackle min-vol  $\beta$ -NMF, focusing on the KL and IS divergences. The algorithm is based on multiplicative updates (MU) that are derived using the standard majorization-minimization framework, and that monotonically decrease the objective function. In Section 3.4, we present several numerical experiments, comparing min-vol  $\beta$ -NMF with standard NMF and sparse NMF. The two mains conclusions are that (1) minimum-volume  $\beta$ -NMF performs consistently better to identify the sources, and (2) as opposed to NMF and sparse NMF, min-vol  $\beta$ -NMF is able to detect when the factorization rank is overestimated by automatically setting sources to zero hence performs model order selection automatically.

The content of this chapter is extracted from:

- [90] V. Leplat, N. Gillis, and A.M.S. Ang. Blind Audio Source Separation With Minimum-Volume Beta-Divergence NMF. In *IEEE Transactions on Signal Processing* 68(2020), pp. 3400-3410.
- [93] V. Leplat, N. Gillis, X. Siebert and A.M.S. Ang. Séparation aveugle de sources sonores par factorisation en matrices positives avec pénalité sur le volume du dictionnaire. In *XXVIIeme Colloque francophone de traitement du signal et des images*. GRESTI. 2019

### 3.1 Introduction: NMF for audio source separation

Given an input matrix  $V \in \mathbb{R}_+^{F \times N}$  that correspond whether to an amplitude audio spectrogram whether a power spectrogram (see Section 1.5.2), we are searching for two nonnegative matrices  $W \in \mathbb{R}_+^{F \times K}$  and  $H \in \mathbb{R}_+^{K \times N}$  where  $K$  is the factorization rank such that  $V \approx WH$ . Because we deal in this section with real-life audio signals, that are noisy settings, we consider the approximate NMF models. When the matrix  $V$  corresponds to the amplitude spectrogram or the power spectrogram of an audio signal, let us briefly recall that:

- $W$  is referred as the dictionary matrix and each column corresponds to the spectral content of a source, and
- $H$  is the activation matrix specifying if a source is active at a certain time frame and in which intensity.

In other words, each rank-one factor  $W(:,k)H(k,:)$  will correspond to a source: the  $k$ th column  $W(:,k)$  of  $W$  is the spectral content of source  $k$ , and the  $k$ th row  $H(k,:)$  of  $H$  is its activation over time. To compute  $W$  and  $H$ , NMF requires to solve the following optimization problem associated to the approximate NMF models:

$$\min_{W \geq 0, H \geq 0} D(V|WH) = \sum_{f,n} d(V_{fn}|[WH]_{fn}),$$

where  $A \geq 0$  means that  $A$  is component-wise nonnegative, and  $d(x|y)$  is an appropriate measure of fit. In audio source separation, a common measure of fit is the discrete  $\beta$ -divergence denoted  $d_\beta(x|y)$ , see Section 1.9.1.

### 3.2 Minimum-volume $\beta$ -NMF

In this section, we present a new optimization problem for the separation based on the minimization of  $\beta$ -divergences including a penalty term promoting solutions with minimum volume spanned by the columns of the dictionary matrix  $W$ . Section 3.2.1 briefly recalls the geometric interpretation of NMF which motivated the use of a minimum volume penalty on the dictionary  $W$ . Section 3.2.2 discusses the new proposed normalization



compared to previous minimum volume NMF problems, and proves that min-vol  $\beta$ -NMF provably recovers the true factors  $(W, H)$  under mild conditions and in the noiseless case; see Theorem 3.2.1.

### 3.2.1 Geometry and min-vol $\beta$ -NMF problem

As mentioned earlier,  $V = WH$  means that each column of  $V$  is a linear combination of the columns of  $W$  weighted by the components of the corresponding column of  $H$ . As mentioned in previous chapters, NMF decompositions are not unique because there exists several (often, infinitely many) sets of columns of  $W$  that span the convex cone generated by the data points. In order to make the problem well-posed and the parameters  $(W, H)$  of the problem identifiable, a key idea is to look for a solution  $W$  with minimum volume, see Section 1.8.4. The use of minimum-volume NMF has led to a new class of NMF methods that outperforms existing ones in many applications such as document analysis and BHU; see the recent survey [90]. Note that minimum-volume NMF implicitly enhances the factor  $H$  to be sparse: the fact that  $W$  has a small volume implies that many data points will be located on the facets of the cone( $W$ ) hence  $H$  will be sparse.

Hence, in this chapter, we consider the following optimization problem, referred to as min-vol  $\beta$ -NMF:

$$\min_{W(:,j) \in \Delta^F \forall j, H \geq 0} D_\beta(V|WH) + \lambda \text{vol}(W), \quad (3.1)$$

where  $\Delta^F = \{x \in \mathbb{R}_+^F \mid \sum_{i=1}^F x_i = 1\}$  is the unit simplex,  $\lambda$  is a penalty parameter and  $\text{vol}(W)$  is a function that measures the volume spanned by the columns of  $W$ . In this thesis, we use  $\text{vol}(W) = \log \det(W^T W + \delta I)$ , where  $\delta$  is a small positive constant that prevents  $\log \det(W^T W)$  to go to  $-\infty$  when  $W$  tends to a rank-deficient matrix (that is, when  $r = \text{rank}(W) < K$ ). The justification for using such volume measurement has been discussed earlier in Chapter 2.

### 3.2.2 Normalization and identifiability

The first identifiability results for minimum-volume NMF problems assumed that the entries in each column of  $H$  sum to one, that is, that  $H^T e = e$  where  $e$  is the all-one column vector whose dimension is clear from the context, meaning that  $H$  is column stochastic, see Section 1.8.4. Under this condition, each column of  $V$  lies in the *convex hull* of the columns of  $W$ . Under the three assumptions that (1)  $H$  is column stochastic, (2)  $W$  is full column rank, and (3)  $H$  satisfies the sufficiently scattered condition, as per Theorem 1.8.4; minimizing the volume of  $\text{conv}(W)$  such that  $V = WH$  recover the true underlying factors, up to permutation and scaling. The sufficiently scattered condition makes sense for most audio source data sets as it is reasonable to assume that, for most time points, only a few sources are active hence  $H$  is sparse. As mentioned earlier, the SSC condition is a relaxation of the separability that is a much stronger assumption in this context since

it requires that, for each sources, there exists a time point where only that source is active [90].

Similarly as done in [49], we prove that requiring  $W$  to be column stochastic (which can also be made without loss of generality) also leads to identifiability. Geometrically, the columns of  $W$  are constrained to be on the unit simplex. Minimizing the volume still makes a lot of sense: we want the columns of  $W$  to be as close as possible to one another within the unit simplex. Here-under, we prove the following theorem.

**Theorem 3.2.1.** *Assume  $V = W^\# H^\#$  with  $\text{rank}(V) = K$ ,  $W^\# \geq 0$  and  $H^\#$  satisfies the sufficiently scattered condition (Definition 1.8.3 in section 1.8.2). Then the optimal solution of*

$$\min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} \text{logdet}(W^T W) \quad (3.2)$$

*such that  $V = WH, W^T e = e, H \geq 0$ ,*

*recovers  $(W^\#, H^\#)$  up to permutation and scaling.*

*Proof.* Recall that  $W^\#$  and  $H^\#$  are the true latent factors that generated  $V$ , with  $\text{rank}(V) = K$  and  $H^\#$  is sufficiently scattered. Let us consider  $\hat{W}$  and  $\hat{H}$  a feasible solution of (3.2). Since  $\text{rank}(V) = K$  and  $V = \hat{W}\hat{H}$ , we must have  $\text{rank}(\hat{W}) = \text{rank}(\hat{H}) = K$ . Hence there exists an invertible matrix  $A \in \mathbb{R}^{K \times K}$  such that  $\hat{W} = W^\# A^{-1}$  and  $\hat{H} = AH^\#$ . Since  $\hat{W}$  is a feasible solution of problem (3.2), we have

$$e^T \hat{W} = e^T W^\# A^{-1} = e^T A^{-1} = e^T,$$

where we assumed  $e^T W^\# = e^T$  without loss of generality since  $W^\# \geq 0$  and  $\text{rank}(W^\#) = K$ . Note that  $e^T A^{-1} = e^T$  is equivalent to  $e^T A = e^T$ . This means that matrix  $A$  is column stochastic. Therefore we have that  $e^T A e = K$ . Since  $\hat{H}$  is a feasible solution, we also have  $\hat{H} = AH^\# \geq 0$ . Let us denote by  $a_j$  the  $j$ th row of  $A$ , and by  $a_k^T$  the  $k$ th column of  $A^T$ . By the definition of the a dual cone,  $AH^\# \geq 0$  means that the rows  $a_j \in \text{cone}^*(H^\#)$  for  $j = 1, \dots, K$ , hence  $\text{cone}(A^T) \subseteq \text{cone}^*(H^\#)$  as per consequence 1 or Lemma 1.8.1. Since  $H^\#$  is sufficiently scattered,  $\text{cone}^*(H^\#) \subseteq \mathcal{C}^*$  (by Definition 1.8.3 and Lemma 1.8.3) hence  $a_j \in \mathcal{C}^*$ . Therefore we have  $\|a_j\|_2 \leq a_j e$  by definition of  $\mathcal{C}$  and Lemma 1.8.4. This leads to

the following:

$$\begin{aligned}
 |\det(A)| &= |\det(A^T)| \leq \prod_{k=1}^K \|a_k^T\|_2 \\
 &= \prod_{j=1}^K \|a_j\|_2 \\
 &\leq \prod_{j=1}^K a_j e \\
 &\leq \left( \frac{\sum_{j=1}^K a_j e}{K} \right)^K \\
 &= \left( \frac{e^T A e}{K} \right)^K \\
 &= 1.
 \end{aligned} \tag{3.3}$$

The first inequality is the Hadamard inequality, the second inequality is due to  $a_j \in \mathcal{C}^*$ , the third inequality is the arithmetic-geometric mean inequality.

Let us now consider two possible situations:

1. If  $|\det(A)| = 1$ , all inequalities above hold as equality's and specifically, we have that :

$$a_j e = \|a_j\|_1 = \|a_j\|_2 \text{ for all } j=1, \dots, K.$$

Therefore matrix  $A^T$  is orthogonal (this is a standard linear algebra result). Since  $\text{cone}(A^T) \subseteq \text{cone}^*(H^\#)$ , we can write  $\text{cone}(H^\#) \subseteq \text{cone}^*(A^T)$  by Lemma 1.8.3 (and since the dual of the dual of a convex cone  $\mathcal{S}$  is the cone  $\mathcal{S}$ ). Since  $A^T$  is orthogonal, we know that  $\text{cone}^*(A^T) = \text{cone}(A^T)$  per Lemma 1.8.7, then  $\text{cone}(H^\#) \subseteq \text{cone}(A^T)$ . Per hypothesis,  $H^\#$  satisfies SSC2 (see Definition 1.8.3), then the only matrix  $A^T$  such that  $\text{cone}(H^\#) \subseteq \text{cone}(A^T)$  must be a permutation matrix. Hence,  $A$  can only be permutation matrix as well and therefore identifiability for problem (3.2) holds per Definition 1.8.1.

2. Let us consider an optimal solution  $(W^*, H^*)$  to problem (3.2) where  $H^*$  is not a row permutation of  $H^\#$  (note that the scaling ambiguity is absent due to the constraint  $W^T e = e$ ). This assumption implies that  $A$  is not a permutation matrix and per the chain rule in (3.3) we have  $|\det(A)| < 1$ . Further, by optimality, we have that  $\log \det((W^*)^T W^*) \leq \log \det((W^\#)^T W^\#)$  which implies  $\det((W^*)^T W^*) \leq \det((W^\#)^T W^\#)$  since  $\log$  is an increasing monotone function. Since the optimal solution  $(W^*, H^*)$  is also feasible, it means that  $W^* = W^\# A^{-1}$  and  $H^* = A H^\#$  for a certain invertible matrix  $A \in \mathbb{R}^{K \times K}$ . Let us insert these relations in the volume function:

$$\begin{aligned}
 \det((W^*)^T W^*) &= \det \left( A^{-T} (W^\#)^T W^\# A^{-1} \right) \\
 &= \det \left( (W^\#)^T W^\# \right) |\det(A)|^{-2} \\
 &> \det \left( (W^\#)^T W^\# \right)
 \end{aligned} \tag{3.4}$$

since  $\det(A^T) = \det(A)$ ,  $\det(A^{-1}) = (\det(A))^{-1}$  and  $|\det(A)| < 1$  per hypothesis. This result then contradicts our assumption for the optimality of  $(W^*, H^*)$ . Therefore, matrix  $A$  can only be a permutation matrix for an optimal solution  $(W^*, H^*)$  of (3.2), and therefore identifiability for problem (3.2) holds which concludes the proof.  $\square$

In noiseless conditions, replacing  $W^T e = e$  with  $He = e$  in (3.2) leads to the same identifiability result; see [48, Theorem 1]. Therefore, in noiseless conditions and under the conditions of Theorem 3.2.1, both problems return the same solution up to permutation and scaling. However, in the presence of noise, we have observed that the two problems may behave very differently. In fact, we advocate that the constraint  $W^T e = e$  is better suited for noisy real-world problems, which we have observed on many numerical examples. In fact, we have observed that the normalization  $W^T e = e$  is much less sensitive to noise and returns much better solutions. The reason is mostly twofold:

(i) As described above, using the normalization  $He = e$  amounts to multiply  $W$  by a diagonal matrix whose entries are the  $\ell_1$  norms of the rows of  $H$ . Therefore, the columns of  $W$  that correspond to dominating (resp. dominated) sources, that is, sources with much more (resp. less) power and/or active at many (resp. few) time points, will have much higher (resp. lower) norm. Therefore, the term  $\log\det(W^T W + \delta I)$  is much more influenced by the dominating sources and will have difficulties to penalize the dominated sources. In other terms, the use of the term  $\log\det(W^T W + \delta I)$  with the normalization  $He = e$  implicitly requires that the rank-one factors  $W(:, k)H(k, :)$  for  $k = 1, \dots, K$  are well balanced, that is, have similar norms. This is not the case for many real (audio) signals.

(ii) As it will be explained in Section 3.3, the update of  $W$  needs the computation of the matrix  $Y$  which is the inverse of  $W^T W + \delta I$ —this terms appears in the gradient with respect to  $W$  of the objective function. The numerical stability for such operations is related to the condition number of  $W^T W + \delta I$ . For a  $\ell_1$  normalization on the columns of  $W$ , the condition number is bounded above as follows:

$$\begin{aligned} \text{cond}(W^T W + \delta I) &= \frac{\sigma_{\max}(W^T W + \delta I)}{\sigma_{\min}(W^T W + \delta I)} \\ &= \frac{\sigma_{\max}(W)^2 + \delta}{\sigma_{\min}(W)^2 + \delta} \\ &\leq \frac{(\sqrt{K} \max_k \|W(:, k)\|_2)^2 + \delta}{\delta} \\ &\leq 1 + \frac{K}{\delta} \end{aligned}$$

where  $\sigma_{\min}(W)$  and  $\sigma_{\max}(W)$  are the smallest and largest singular values of  $W$ , respectively. In the numerical experiments, we use  $\delta = 1$ . On the other hand, the normalization  $He = e$  may lead to arbitrarily large values for the condition number of  $W^T W + \delta I$ , which we have observed numerically on several examples. This issue can be mitigated with the

use of the normalization  $He = \rho e$  for some  $\rho > 0$  sufficiently large for which identifiability still holds [48]. However, it still performs worse because of the first reason explained above.

For these reasons, we believe that our problem would also be better suited (compared to the normalization on  $H$ ) in other contexts; for example for document classification [51].

### 3.3 Algorithm for min-vol $\beta$ -NMF

As explained in Section 1.10 most NMF algorithms alternatively update  $H$  for  $W$  fixed and vice versa, and we adopt this strategy in this chapter. For  $W$  fixed, (3.1) is equivalent to subproblem in  $H$  of standard NMF problem 1.4.1 and we will use the MU that have already been derived in the literature [86, 45].

To tackle (3.1) for  $H$  fixed, let us consider

$$\min_{W \geq 0} F(W) = D_{\beta}(V|WH) + \lambda \log \det(W^T W + \delta I). \quad (3.5)$$

Note that, for now, we have discarded the normalization on the columns of  $W$ . In our algorithm, we will use the update for  $W$  obtained by solving (3.5) as a descent direction along with a line search procedure to integrate the constraint on  $W$ . This will ensure that the objective function  $F$  is non-increasing at each iteration. In the following sections we derive MU for  $W$  that decrease the objective in (3.5). We follow the standard majorization-minimization framework [130]. First, an auxiliary function, which we denote  $\bar{F}$ , is constructed so that it majorizes the objective. An auxiliary function for  $F$  at point  $\tilde{W}$  is defined as follows.

**Definition 3.3.1.** *The function  $\bar{F}(W|\tilde{W}) : \Omega \times \Omega \rightarrow \mathbb{R}$  is an auxiliary function for  $F(W) : \Omega \rightarrow \mathbb{R}$  at  $\tilde{W} \in \Omega$  if the conditions  $\bar{F}(W|\tilde{W}) \geq F(W)$  for all  $W \in \Omega$  and  $\bar{F}(\tilde{W}|\tilde{W}) = F(\tilde{W})$  are satisfied.*

Then, the optimization of  $F$  can be replaced by an iterative process that minimizes  $\bar{F}$ . More precisely, the new iterate  $W^{(i+1)}$  is computed by minimizing exactly the auxiliary function at the previous iterate  $W^{(i)}$ . This guarantees  $F$  to decrease at each iteration.

**Lemma 3.3.1.** *Let  $W, W^{(i)} \geq 0$ , and let  $\bar{F}$  be an auxiliary function for  $F$  at  $W^{(i)}$ . Then  $F$  is non-increasing under the update  $W^{(i+1)} = \arg \min_{W \geq 0} \bar{F}(W|W^{(i)})$ .*

*Proof.* In fact, we have by definition that  $F(W^{(i)}) = \bar{F}(W^{(i)}|W^{(i)}) \geq \min_W \bar{F}(W|W^{(i)}) = \bar{F}(W^{(i+1)}|W^{(i)}) \geq F(W^{(i+1)})$ .  $\square$

The most difficult part in using the majorization-minimization framework is to design an auxiliary function that is easy to optimize. Usually such auxiliary functions are separable (that is, there is no interaction between the variables so that each entry of  $W$  can be updated independently) and convex.

### 3.3.1 Separable auxiliary functions for $\beta$ -divergences

For the sake of completeness, we briefly recall the auxiliary function proposed in [45] for the data fitting term. It consists in majorizing the convex part of the  $\beta$ -divergence using Jensen's inequality and majorizing the concave part by its tangent (first-order Taylor approximation). We have

$$d_\beta(x|y) = \check{d}_\beta(x|y) + \hat{d}_\beta(x|y) + \bar{d}_\beta(x|y), \quad (3.6)$$

where  $\check{d}$  is convex function of  $y$ ,  $\hat{d}$  is a concave function of  $y$  and  $\bar{d}$  is a constant of  $y$ ; see Table 3.1.

Table 3.1: Differentiable convex-concave-constant decomposition of the  $\beta$ -divergence under the form (3.6) [45].

	$\check{d}(x y)$	$\hat{d}(x y)$	$\bar{d}(x)$
$\beta = 0$	$xy^{-1}$	$\log(y)$	$x(\log(x) - 1)$
$\beta \in [1, 2]$	$d_\beta(x y)$	0	0

The function  $D_\beta(V|WH)$  can be written as  $\sum_f D_\beta(v_f|w_fH)$  where  $v_f$  and  $w_f$  are respectively the  $f$ th row of  $V$  and  $W$ . Therefore we only consider the optimization over one specific row of  $W$ . To simplify notation, we denote iterates  $w^{(i+1)}$  (next iterate) and  $w^{(i)}$  (current iterate) as  $w$  and  $\tilde{w}$ , respectively.

**Lemma 3.3.2** ([45]). *Let  $\tilde{v} = \tilde{w}H$  and  $\tilde{w}$  be such that  $\tilde{v}_n > 0$  for all  $n$  and  $\tilde{w}_k > 0$  for all  $k$ . Then the function*

$$G(w|\tilde{w}) = \sum_n \left[ \sum_k \frac{\tilde{w}_k h_{kn}}{\tilde{v}_n} \check{d}(v_n | \tilde{v}_n \frac{w_k}{\tilde{w}_k}) \right] + \bar{d}(v_n) + \left[ \check{d}(v_n | \tilde{v}_n) \sum_k (w_k - \tilde{w}_k) h_{kn} + \hat{d}(v_n | \tilde{v}_n) \right] \quad (3.7)$$

is an auxiliary function for  $\sum_n d(v_n | [wH]_n)$  at  $\tilde{w}$ .

### 3.3.2 A separable auxiliary function for the minimum-volume regularizer

The minimum-volume regularizer  $\log\det(W^T W + \delta I)$  is a non-convex function. However, it can be upper-bounded using the fact that  $\log\det(\cdot)$  is a concave function so that its first-order Taylor approximation provides an upper bound; see for example [52]. For any positive-definite matrices  $A$  and  $B \in \mathbb{R}^{K \times K}$ , we have:

$$\begin{aligned} \log\det(B) &\leq \log\det(A) + \text{Tr}(A^{-1}(B - A)) \\ &= \text{Tr}(A^{-1}B) + \log\det(A) - K. \end{aligned}$$

This implies that for any  $W, Z \in \mathbb{R}^{F \times K}$ , we have

$$\log \det(W^T W + \delta I) \leq l(W, Z), \quad (3.8)$$

where  $l(W, Z) = \text{Tr}(Y W^T W) + \log \det(Y^{-1}) - K$ ,  $Y = (Z^T Z + \delta I)^{-1}$  with  $\delta > 0$ . Note that  $Z^T Z + \delta I$  is positive definite hence is invertible and its inverse  $Y$  is also positive definite. Finally  $l(W, Z)$  is an auxiliary function for  $\log \det(W^T W + \delta I)$  at  $Z$ . However, it is quadratic and not separable hence non-trivial to optimize over the nonnegative orthant. The non-constant part of  $l(W, Z)$  can be written as  $\sum_f w_f Y w_f^T$  where  $w_f$  is the  $f$ th row of  $W$ . Henceforth we will focus on one particular row vector  $w$  with  $l(w) = w^T Y w$  which will be further considered as a column vector of size  $K \times 1$ .

**Lemma 3.3.3.** *Let  $w, \tilde{w} \in \mathbb{R}_+^K$  be such that  $\tilde{w}_k > 0$  for all  $k$ ,  $Y = Y^+ - Y^-$  with  $Y^+ = \max(Y, 0)$  and  $Y^- = \max(-Y, 0)$ , and  $\Phi(\tilde{w})$  be the diagonal matrix  $\Phi(\tilde{w}) = \text{diag}\left(2 \frac{[Y^+ \tilde{w} + Y^- \tilde{w}]}{[\tilde{w}]}\right)$  where  $\frac{[A]}{[B]}$  is the component-wise division between  $A$  and  $B$ , and  $\Delta w = w - \tilde{w}$ . Then*

$$\bar{l}(w|\tilde{w}) = l(\tilde{w}) + \Delta w^T \nabla l(\tilde{w}) + \frac{1}{2} \Delta w^T \Phi(\tilde{w}) \Delta w, \quad (3.9)$$

is a separable auxiliary function for  $l(w) = w^T Y w$  at  $\tilde{w}$ .

*Proof.* Separability of  $\bar{l}(w|\tilde{w})$  holds since  $\Phi(\tilde{w})$  is diagonal. The condition  $\bar{l}(\tilde{w}|\tilde{w}) = l(\tilde{w})$  from Definition 3.3.1 can be checked easily. It remains to prove that  $\bar{l}(w|\tilde{w}) \geq l(w)$  for all  $w$ . Let first rewrite the quadratic function  $l(w)$  using its Taylor expansion at  $w = \tilde{w}$ :  $l(w) = l(\tilde{w}) + (w - \tilde{w})^T \nabla l(\tilde{w}) + \frac{1}{2} (w - \tilde{w})^T \nabla^2 l(\tilde{w}) (w - \tilde{w}) = l(\tilde{w}) + (w - \tilde{w})^T 2Y \tilde{w} + \frac{1}{2} (w - \tilde{w})^T 2Y (w - \tilde{w})$ . Proving that  $\bar{l}(w|\tilde{w}) \geq l(w)$  is equivalent to proving that  $\frac{1}{2} (w - \tilde{w})^T [\Phi(\tilde{w}) - 2Y] (w - \tilde{w}) \geq 0$ , which boils down to proving that the matrix  $[\Phi(\tilde{w}) - 2Y]$  is positive semi-definite. We have  $\Phi_{ij}(\tilde{w}) = 2\delta_{ij} \frac{(Y^+ \tilde{w})_i + (Y^- \tilde{w})_i}{\tilde{w}_i}$ , where  $\delta_{ij}$  is the Kronecker symbol. Let us consider the following matrix:  $M_{ij}(\tilde{w}) = \tilde{w}_i [\Phi(\tilde{w}) - 2Y]_{ij} \tilde{w}_j$ , which is a rescaling of  $[\Phi(\tilde{w}) - 2Y]$ . Therefore,  $[\Phi(\tilde{w}) - 2Y]$  is positive semi-definite if

and only if  $M$  is positive semi-definite if and only if for all  $\nu$  we have  $\nu^T M \nu \geq 0$ . We have:

$$\begin{aligned}
 \nu^T M \nu &= \sum_{ij} M_{ij} \nu_i \nu_j = \sum_{ij} \left[ \tilde{w}_i [\Phi(\tilde{w}) - 2Y]_{ij} \tilde{w}_j \right] \nu_i \nu_j \\
 &= 2 \sum_{ij} \left[ \tilde{w}_i \left[ \delta_{ij} \frac{(Y^+ \tilde{w})_i + (Y^- \tilde{w})_i}{\tilde{w}_i} - Y_{ij} \right] \tilde{w}_j \right] \nu_i \nu_j \\
 &= 2 \left[ \sum_{ij} \delta_{ij} [(Y^+ \tilde{w})_i + (Y^- \tilde{w})_i] \tilde{w}_j \nu_i \nu_j \right. \\
 &\quad \left. - \sum_{ij} \tilde{w}_i Y_{ij}^+ \tilde{w}_j \nu_i \nu_j + \sum_{ij} \tilde{w}_i Y_{ij}^- \tilde{w}_j \nu_i \nu_j \right] \\
 &= 2 \left[ \sum_{ij} Y_{ij}^+ \tilde{w}_j \tilde{w}_i \nu_i^2 - \sum_{ij} Y_{ij}^+ \tilde{w}_j \tilde{w}_i \nu_i \nu_j \right. \\
 &\quad \left. + \sum_{ij} Y_{ij}^- \tilde{w}_j \tilde{w}_i \nu_i^2 + \sum_{ij} Y_{ij}^- \tilde{w}_j \tilde{w}_i \nu_i \nu_j \right] \\
 &= \left[ \sum_{ij} Y_{ij}^+ \tilde{w}_j \tilde{w}_i [\nu_i^2 + \nu_j^2 - 2\nu_i \nu_j] \right. \\
 &\quad \left. + \sum_{ij} Y_{ij}^- \tilde{w}_j \tilde{w}_i [\nu_i^2 + \nu_j^2 + 2\nu_i \nu_j] \right] \\
 &= \left[ \sum_{ij} Y_{ij}^+ \tilde{w}_j \tilde{w}_i [\nu_i - \nu_j]^2 + \sum_{ij} Y_{ij}^- \tilde{w}_j \tilde{w}_i [\nu_i + \nu_j]^2 \right],
 \end{aligned}$$

which is nonnegative hence concludes the proof.

Alternatively, we can show that  $M$  is positive semi-definite<sup>1</sup> as follows: since  $M$  is symmetric and its diagonal entries are non-negative, it is sufficient to show that  $M$  is diagonally dominant [71, Proposition 7.2.3], that is,

$$|M_{ii}| \geq \sum_{j \neq i} |M_{ij}| \quad \text{for all } i.$$

We have for all  $i$  that

$$M_{ii} = 2w_i \sum_j (Y_{ij}^+ + Y_{ij}^-) w_j - 2w_i Y_{ii} w_i, \text{ and}$$

$$M_{ij} = -2w_i Y_{ij} w_j \quad \text{for } j \neq i.$$

Since  $Y_{ij}^+ + Y_{ij}^- = |Y_{ij}|$ , we have

$$\begin{aligned}
 M_{ii} - \sum_{j \neq i} |M_{ij}| &= 2w_i \sum_j |Y_{ij}| w_j - 2w_i Y_{ii} w_i \\
 &\quad - 2w_i \sum_{j \neq i} |Y_{ij}| w_j \\
 &= 2w_i |Y_{ii}| w_i - 2w_i Y_{ii} w_i \geq 0,
 \end{aligned}$$

implying that  $M$  is diagonally dominant. □

---

<sup>1</sup>this was suggested to us by one of the reviewers of [90], it is more elegant and simpler than our original proof.



**Remark 1** (Choice of the auxiliary function). *A simpler choice for the auxiliary function would be to replace  $\Phi(\tilde{w})$  with  $2\lambda_{\max}(Y)I$  where  $\lambda_{\max}(Y)$  is the largest eigenvalue of  $Y$  (the constant 2 appears because  $l(w) = w^T Y w$  while there is a factor 1/2 in front of  $\Phi(\tilde{w})$ ). However, it would lead to a worse approximation. In particular if  $Y$  is a diagonal matrix (since  $Y > 0$ , these diagonal elements are positive), our choice gives  $\Phi(\tilde{w}) = 2Y$  for any  $\tilde{w} > 0$ , meaning that the auxiliary function matches perfectly the function  $l(w)$ . This would not be the case for the choice  $2\lambda_{\max}(Y)I$  (unless  $Y$  is a scaling of the identity matrix).*

### 3.3.3 Auxiliary function for min-vol $\beta$ -NMF

Based on the auxiliary functions presented in Sections 3.3.1 and 3.3.2, we can directly derive a separable auxiliary function  $\bar{F}(W|\tilde{W})$  for min-vol  $\beta$ -NMF (3.1).

**Corollary 3.3.0.1.** *For  $W, H \geq 0$ ,  $\lambda > 0$ ,  $Y = (\tilde{W}^T \tilde{W} + \delta I)^{-1}$  with  $\delta > 0$  and the constant  $c = \log \det(Y^{-1}) + K$ , the function*

$$\bar{F}(W|\tilde{W}) = \sum_f G(w_f|\tilde{w}_f) + \lambda \left( \sum_f \bar{l}(w_f|\tilde{w}_f) + c \right),$$

where  $G$  is given by (3.7) and  $\bar{l}$  by (3.9), is a convex and separable auxiliary function for  $F(W) = D_\beta(V|WH) + \lambda \log \det(W^T W + \delta I)$  at  $\tilde{W}$ .

*Proof.* This follows directly from Lemma 3.3.2, Equation (3.8) and Lemma 3.3.3.  $\square$

In the following section, we provide explicitly MU for the KL divergence ( $\beta = 1$ ) by finding a closed-form solution for the minimization of  $\bar{F}$ . In Section 3.3.5, we provide the MU for the IS divergence ( $\beta = 0$ ). Note that the other cases are not treated explicitly but can be in a similar way. For the same reason, we will only compare KL NMF problems in the numerical experiments (Section 3.4).

### 3.3.4 Algorithm for min-vol KL-NMF

As before, let us focus on a single row of  $W$ , denoted  $w$ , as the objective function  $F(W)$  is separable by row. For  $\beta = 1$ , the derivative of the auxiliary function  $\bar{F}(w|\tilde{w})$  with respect to a specific coefficient  $w_k$  is given by  $\nabla_{w_k} \bar{F}(w|\tilde{w}) = \sum_n h_{kn} - \sum_n h_{kn} \frac{\tilde{w}_k v_n}{w_k \tilde{w}_n} + 2\lambda [Y \tilde{w}]_k + 2\lambda \left[ \text{diag} \left( \frac{Y^+ \tilde{w} + Y^- \tilde{w}}{\tilde{w}} \right) \right]_k w_k - 2\lambda \left[ \text{diag} \left( \frac{Y^+ \tilde{w} + Y^- \tilde{w}}{\tilde{w}} \right) \right]_k \tilde{w}_k$ .

Due to the separability, we set the derivative to zero to obtain the closed-form solution, which is given in Table 3.2 in matrix form.

Note that although the closed-form solution has a negative term in the numerator of the multiplicative factor (see Table 3.2), they always remain nonnegative given that  $V, H$  and  $\tilde{W}$  are nonnegative. In fact, the term before the minus sign is always larger than the term after the minus sign:  $e_{F,N} H^T - 4\lambda(\tilde{W} Y^-)$  is squared (component wise) and added a positive term, hence the component-wise square root of that result is larger than  $e_{F,N} H^T - 4\lambda(\tilde{W} Y^-)$ .

Table 3.2: Multiplicative update for min-vol KL-NMF.

$$W = \tilde{W} \odot \frac{\left[ \left[ e_{F,N} H^T - 4\lambda(\tilde{W}Y^-) \right]^2 + 8\lambda\tilde{W}(Y^+ + Y^-) \odot \left( \frac{[V]}{[\tilde{W}H]} H^T \right) \right]^{\frac{1}{2}} - (e_{F,N} H^T - 4\lambda(\tilde{W}Y^-))}{[4\lambda\tilde{W}(Y^+ + Y^-)]},$$

where  $A \odot B$  (resp.  $\frac{[A]}{[B]}$ ) is the Hadamard product (resp. division) between  $A$  and  $B$ ,  $A^{(\alpha)}$  is the element-wise  $\alpha$  exponent of  $A$ ,  $e_{F,N}$  is the  $F$ -by- $N$  all-one matrix, and  $Y = Y^+ - Y^- = (\tilde{W}^T \tilde{W} + \delta I)^{-1}$  with  $\delta > 0$ ,  $Y^+ \geq 0$ ,  $Y^- \geq 0$ ,  $\lambda > 0$ .

Algorithm 3 summarizes our algorithm to tackle (3.1) for  $\beta = 1$  which we refer to as min-vol KL-NMF LS (Line Search). Note that the update of  $H$  (step 4) is the one from [86]. More importantly, note that we have incorporated a line-search for the update of  $W$ . In fact, although the MU for  $W$  are guaranteed to decrease the objective function, they do not guarantee that  $W$  remains normalized, that is, that  $\|W(:, k)\|_1 = 1$  for all  $k$ . Hence, we normalize  $W$  after it is updated (step 10), and we normalize  $H$  accordingly so that  $WH$  remains unchanged. When this normalization is performed, the  $\beta$ -divergence part of  $F$  is unchanged but the minimum-volume penalty will change so that the objective function  $F$  might increase. In order to guarantee non-increasingness, we integrate a simple backtracking line-search procedure; see steps 11-16 of Algorithm 3. In summary, our MU provide a descent direction that preserved nonnegativity of the iterates, and we use a projection and a simple backtracking line-search to guarantee the monotonicity of the objective function, as in standard projected gradient descent methods.

---

**Algorithm 3** min-vol KL-NMF LS
 

---

**Require:** A matrix  $V \in \mathbb{R}^{M \times T}$ , an initialization  $H \in \mathbb{R}_+^{K \times T}$ , an initialization  $W \in \mathbb{R}^{M \times K}$ , a factorization rank  $K$ , a maximum number of iterations `maxiter`, min-vol weight  $\lambda > 0$  and  $\delta > 0$

**Ensure:** A rank- $K$  NMF  $(W, H)$  of  $V \approx WH$  with  $W \geq 0$  and  $H \geq 0$ .

```

1:  $\gamma = 1, Y = (W^T W + \delta I)^{-1}$ 
2: for  $i = 1$  : maxiter do
3:     % Update of matrix  $H$ 
4:      $H \leftarrow H \odot \frac{W^T \left( \begin{bmatrix} V \\ WH \end{bmatrix} \right)}{W^T e_{F,N}}$ 
5:     % Update of matrix  $W$ 
6:      $Y \leftarrow (W^T W + \delta I)^{-1}$ 
7:      $Y^+ \leftarrow \max(Y, 0)$ 
8:      $Y^- \leftarrow \max(-Y, 0)$ 
9:      $W^+$  is updated according to Table 3.2
10:     $(W_\gamma^+, H_\gamma) = \text{normalize}(W^+, H)$ 
11:    % Line-search procedure
12:    while  $F(W_\gamma^+, H_\gamma) > F(W, H)$  do
13:         $\gamma \leftarrow \gamma \times 0.8$ 
14:         $W_\gamma^+ \leftarrow (1 - \gamma)W + \gamma W^+$ 
15:         $(W_\gamma^+, H_\gamma) \leftarrow \text{normalize}(W_\gamma^+, H)$ 
16:    end while
17:     $(W, H) \leftarrow (W_\gamma^+, H_\gamma)$ 
18:    % Update of  $\gamma$  to avoid a vanishing stepsize
19:     $\gamma \leftarrow \min(1, \gamma \times 1.2)$ 
20: end for

```

---

It can be verified that the computational complexity of the min-vol KL-NMF LS is asymptotically equivalent to the standard MU for  $\beta$ -NMF, that is, it requires  $\mathcal{O}(FNK)$  operations per iteration. Indeed, all the main operations include matrix products with a complexity of  $\mathcal{O}(FNK)$  and element-wise operations on matrices of size  $F \times K$  or  $K \times N$ . Note that the inversion of the  $K$ -by- $K$  matrix  $(W^T W + \delta I)$  requires  $\mathcal{O}(K^3)$  operations which is dominated by  $\mathcal{O}(FNK)$  since  $K \leq \min(F, N)$  (in fact, typically  $K \ll \min(F, N)$  hence this term is negligible). Therefore, although Algorithm 3 will be slower than the baseline KL-NMF (that is, the standard MU) because of the additional terms to be computed and the line-search, the asymptotical computational cost is the same; see Table 3.4 for runtime comparison.

### 3.3.5 Algorithm for min-vol IS-NMF

For  $\beta = 0$  (IS divergence), the derivative of the auxiliary function  $\bar{F}(w|\tilde{w})$  with respect to a specific coefficient  $w_k$  is given by:

$$\begin{aligned} \nabla_{w_k} \bar{F}(w|\tilde{w}) &= \sum_n \frac{h_{kn}}{\tilde{v}_n} - \sum_n h_{kn} \frac{\tilde{w}_k^2 v_n}{w_k^2 \tilde{v}_n^2} + 2\lambda [Y\tilde{w}]_k \\ &\quad + 2\lambda \left[ \text{diag} \left( \frac{Y^+ \tilde{w} + Y^- \tilde{w}}{\tilde{w}} \right) \right]_k w_k \\ &\quad - 2\lambda \left[ \text{diag} \left( \frac{Y^+ \tilde{w} + Y^- \tilde{w}}{\tilde{w}} \right) \right]_k \tilde{w}_k. \end{aligned}$$

Let

$$\begin{aligned} \tilde{a} &= 2\lambda \left[ \text{diag} \left( \frac{Y^+ \tilde{w} + Y^- \tilde{w}}{\tilde{w}} \right) \right]_k, \\ \tilde{b} &= \sum_n \frac{h_{kn}}{\tilde{v}_n} - 4\lambda [Y^- \tilde{w}]_k, \\ \tilde{d} &= - \sum_n h_{kn} \frac{\tilde{w}_k^2 v_n}{\tilde{v}_n^2}. \end{aligned} \tag{3.10}$$

Setting the derivative to zero requires to compute the roots of the following degree-three polynomial  $\tilde{a}w_k^3 + \tilde{b}w_k^2 + \tilde{d}$ . We used the procedure developed in [118] which is based on the explicit calculation of the intermediary root of a canonical form of cubic. This procedure is able to provide highly accurate numerical results even for badly conditioned polynomials. The algorithm for min-vol IS-NMF follows the same steps as for min-vol KL-NMF LS: only the two steps corresponding to the updates of  $W$  and  $H$  have to be modified. For the update of  $H$  (step 4), use the standard MU. For the update of  $W$  (step 9), use

```

for  $f \leftarrow 1$  to  $F$ 
  for  $k \leftarrow 1$  to  $K$ 
    Compute  $\tilde{a}$ ,  $\tilde{b}$  and  $\tilde{d}$  according to equations (3.10)
    Compute the roots of  $\tilde{a}w_k^3 + \tilde{b}w_k^2 + \tilde{d}$ 
    Pick  $y$  among these roots and zero that minimizes
    the objective
     $W_{f,k}^+ \leftarrow \max(10^{-16}, y)$ 
  end for
end for
    
```

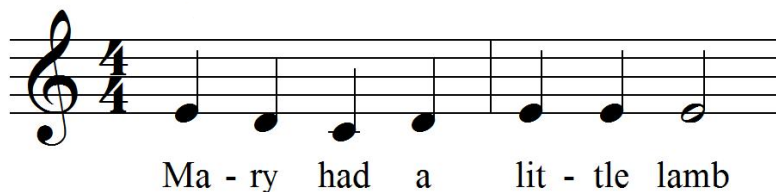
## 3.4 Numerical experiments

In this section we report an experimental comparative study of baseline KL-NMF, min-vol KL-NMF LS (Algorithm 3) and sparse KL-NMF [119] applied to the spectrogram of three monophonic piano sequences and a synthetic mix of a bass and drums. For the two monophonic piano sequences, the audio signals are true life signals with standard quality. Note that the sequences are made of pure piano notes, the number  $K$  should therefore correspond to the number of notes present into the mixed signals. The comparative study

is performed for several values of  $K$  with a focus on the case where the factorization rank  $K$  is overestimated. For all simulations, random initializations are used for  $W$  and  $H$ , the best results among 5 runs are kept for the comparative study. In all cases, we use a Hamming window of size  $F=1024$ , and 50% overlap between two frames. Sparse KL-NMF LS has a similar structure as min-vol KL-NMF LS, with a penalty parameter for the sparsity enforcing regularization. To tune these two parameters, we have used the same strategy for both methods: we manually tried a wide range of values and report the best results.

The code is available from [bit.ly/minvolKLNMF](https://bit.ly/minvolKLNMF) (code written in MATLAB R2017a), and can be used to rerun directly all experiments below. They were run on a laptop computer with Intel Core i7-7500U CPU @ 2.70GHz 4 and 32GB memory. A demonstration video has been recorded and is available online from <https://www.youtube.com/watch?v=1BrpxvpghKQ> which shows the results obtained with our algorithm applied to the Montoise folk song "El Doudou" with an overestimated factorization rank  $K$ .

**Mary had a little lamb** The first audio sample is the first measure of "Mary had a little lamb". As explained in Section 1.5.2 the sequence is composed of three notes;  $E_4$ ,  $D_4$  and  $C_4$ , played all at once. The recorded signal is 4.7 seconds long and downsampled to  $f_s = 16000\text{Hz}$  yielding  $T=75200$  samples. STFT of the input signal  $x$  yields a temporal resolution of 16ms and a frequency resolution of 31.25Hz, so that the amplitude spectrogram  $V$  has  $N=294$  frames and  $F=257$  frequency bins. The musical score is shown on Figure 3.1. The input audio signal  $x(t)$  and the input matrix  $V$  (amplitude spectrogram) are pictured in Figure 1.6.

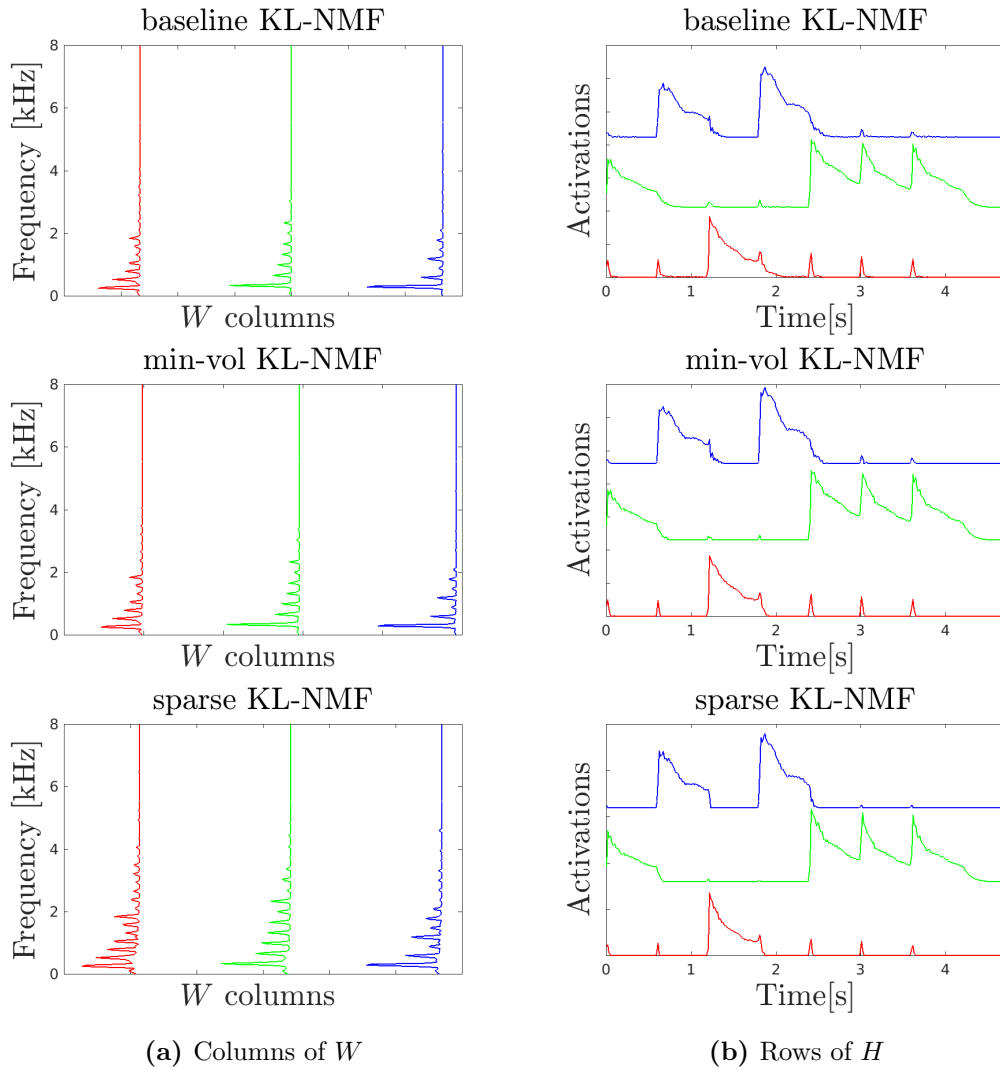


**Fig. 3.1.** Musical score of "Mary had a little lamb".

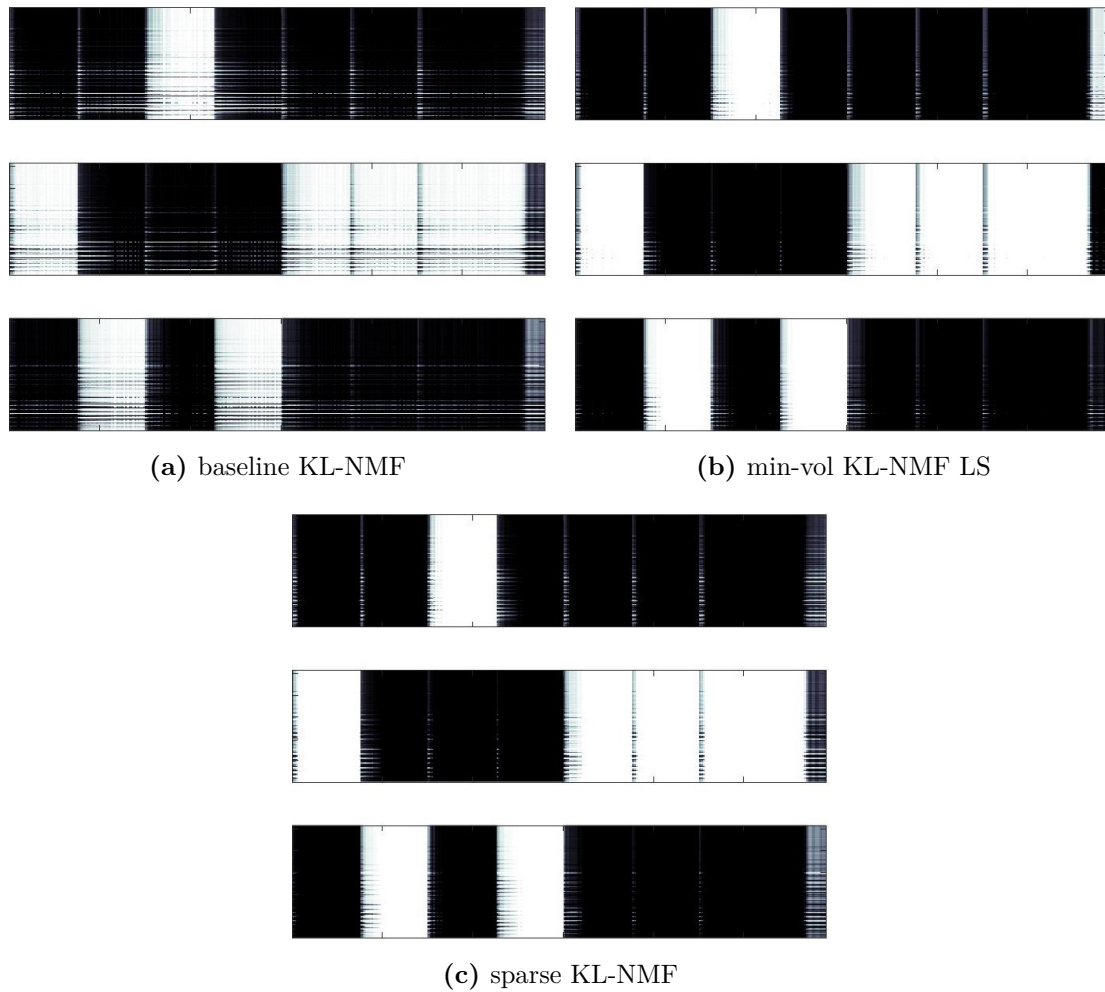
All NMF algorithms were run for 200 iterations which allowed them to converge. Figure 3.2 presents the columns of  $W$  (dictionary matrix) and the rows of  $H$  for baseline KL-NMF and min-vol KL-NMF LS with  $K = 3$ . Figure 3.3 presents the time-frequency masking coefficients. These coefficients are computed as follows

$$\text{mask}_{f,n}^{(k)} = \frac{\hat{X}_{f,n}^{(k)}}{\sum_k \hat{X}_{f,n}^{(k)}} \quad \text{with } k = 1, \dots, K,$$

where  $\hat{X}^{(k)} = W(:,k)H(k,:)$  is the estimated source  $k$ . The masks are nonnegative and sum to one for each pair  $(f,n)$ . This representation allows to identify visually whether



**Fig. 3.2.** Comparative study of baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to “Mary had a little lamb” amplitude spectrogram with  $K=3$ .



**Fig. 3.3.** Masking coefficients obtained with baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to “Mary had a little lamb” amplitude spectrogram with  $K=3$ .

the NMF algorithm was able to separate the sources properly. All the simulations give a nice separation with similar results for  $W$  and  $H$ . The activations are coherent with the sequences of the notes. However, Figure 3.3 shows that min-vol KL-NMF LS and sparse KL-NMF provide a better separation in terms of time-frequency localization compared to the baseline KL-NMF.

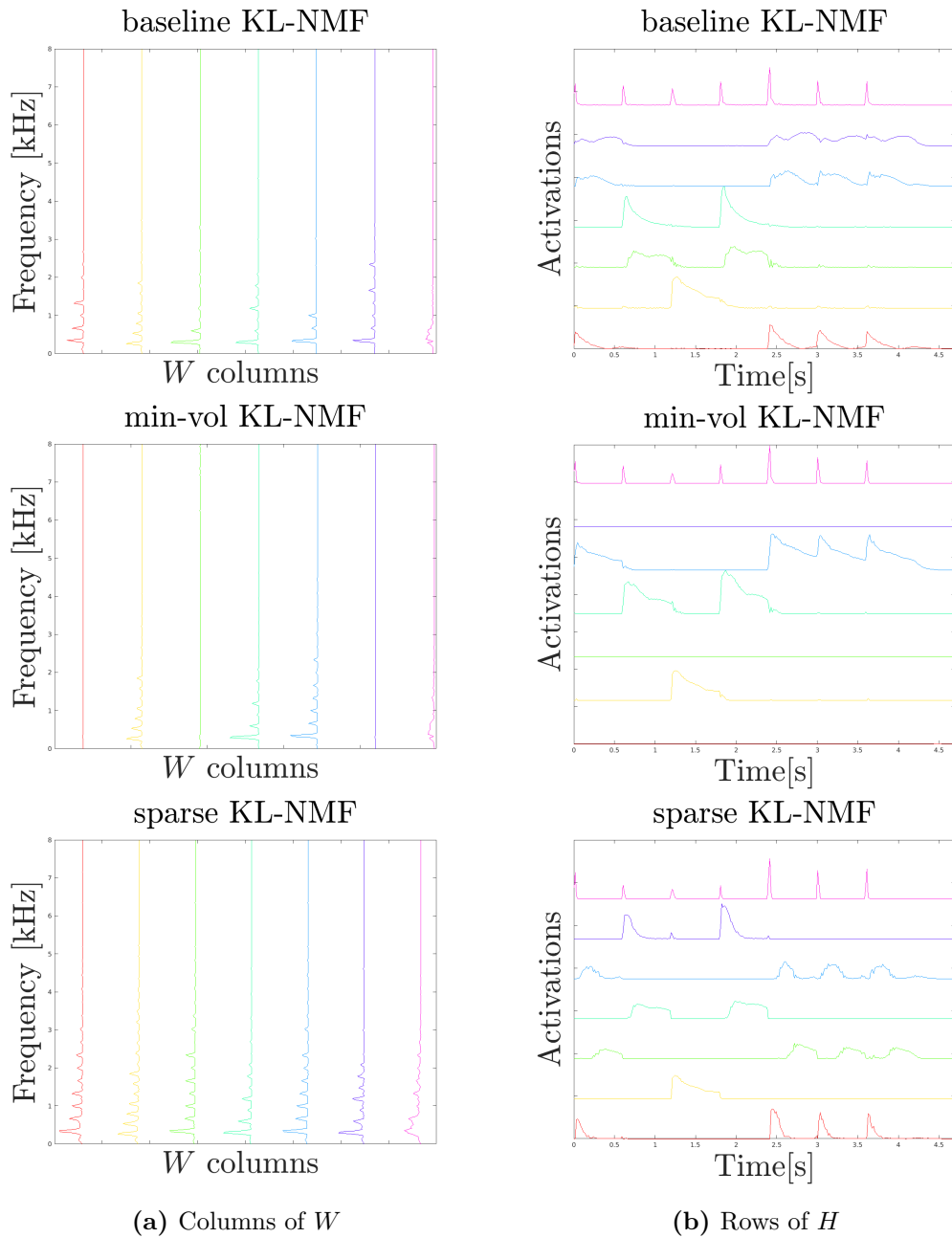
We now perform the same experiment but using  $K=7$ . Figure 3.4 presents the results. This situation corresponds to the situation where the factorization rank is overestimated. Figure 3.5 presents the time-frequency masking coefficients.

We observe that min-vol KL-NMF LS is able to extract the three notes correctly and set automatically to zero three source estimates (more precisely, three rows of  $H$  are set to zero, while the corresponding columns of  $W$  have entries equal to one another as  $\|W(:, k)\|_1 = 1$  for all  $k$ ) while baseline KL-NMF and sparse KL-NMF split the notes in all the sources. One can observe that a fourth note is identified in all simulations (see isolated peaks on Figure 3.5-(b), second row of  $H$  from the top) and corresponds to each very first offset of each note in the musical sequence. This result makes sense and corresponds to some common mechanical vibrations acting in the piano just before triggering a specific note. This observation is confirmed by the fact that the amplitude is proportional to the natural strength of the fingers playing the notes. In this scenario, with  $K$  is overestimated, min-vol KL-NMF LS outperforms baseline KL-NMF and sparse KL-NMF.

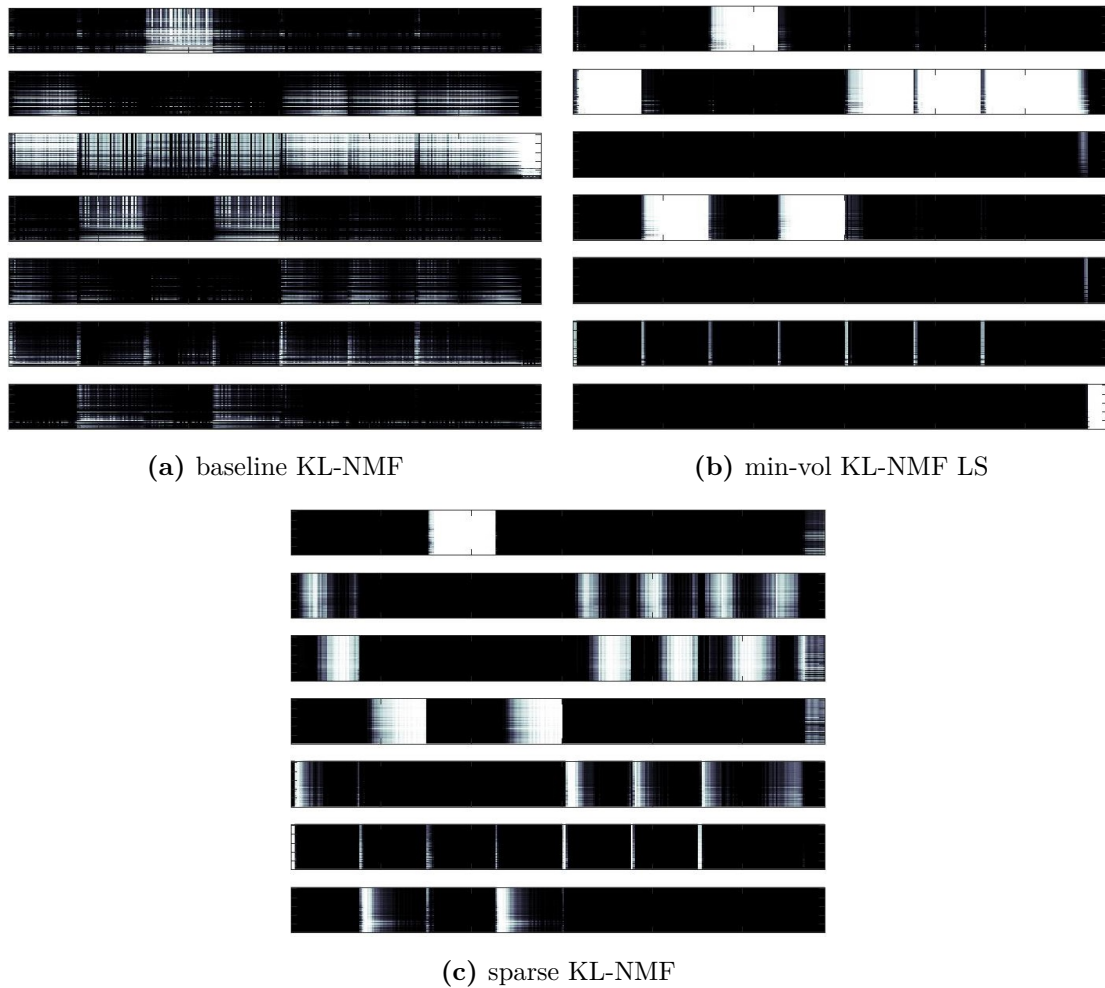
**Prelude of Bach** The second audio sample corresponds to the first 30 seconds of "Prelude and Fugue No.1 in C major" from J. S. Bach played by Glenn Gould<sup>2</sup>. The audio sample is a sequence of 13 notes:  $B_3, C_4, D_4, E_4, F_4^\#, G_4, A_4, C_5, D_5, E_5, F_5, G_5, A_5$ . The recorded signal is downsampled to  $f_s = 11025\text{Hz}$  yielding  $T=330750$  samples. STFT of the input signal  $x$  yields a temporal resolution of 46ms and a frequency resolution of 10.76Hz, so that the amplitude spectrogram  $V$  has  $N=647$  frames and  $F=513$  frequency bins. The musical score is presented on Figure 3.8. All NMF algorithms were run for 300 iterations which allowed them to converge. Figure 3.9 presents the results obtained for  $W$  and  $H$  with a factorization rank  $K = 16$ , hence overestimated. We observe that min-vol KL-NMF LS automatically sets three components to zero (with \* symbol on Figure 3.9) while 13 source estimates are determined. The analysis of the fundamentals (maximum peak frequency) of the 13 source estimates correspond to the theoretical fundamentals of the 13 notes mentioned earlier. Note that using baseline KL-NMF or sparse KL-NMF led to same conclusions as for the first audio sample; these two algorithms generate as many source estimates as imposed by the rank of factorization while min-vol KL-NMF LS algorithm preserves the integrity of the 13 sources. Additionally, the activations are coherent with the sequences of the notes. Figure 3.10 shows (on a limited time interval) that the estimate sequence follows the sequence defined in the score. Note that a threshold and permutations

<sup>2</sup><https://www.youtube.com/watch?v=Z1bK5r5mBH4>

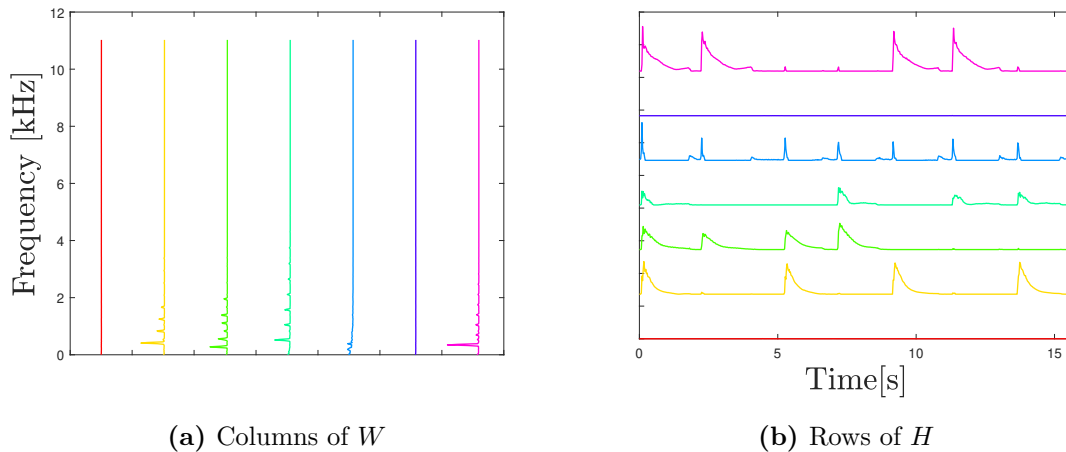




**Fig. 3.4.** Comparative study of baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to “Mary had a little lamb” amplitude spectrogram with  $K=7$ .



**Fig. 3.5.** Masking coefficients obtained with baseline KL-NMF (top), min-vol KL-NMF LS (middle) and sparse KL-NMF (bottom) applied to “Mary had a little lamb” amplitude spectrogram with  $K=7$ .



**Fig. 3.6.** Factors matrices  $W$  and  $H$  obtained with min-vol KL-NMF LS with factorization rank  $K=7$  for the third audio sample.

on rows of  $H$  was used to improve visibility.

**Third piano sequence** The third audio sample is a piano sequence played from the score given in Figure 3.7. The piano sequence is composed of four notes;  $D_4$ ,  $F_4$ ,  $A_4$  and  $C_5$ , played all at once in the first measure and then played by pairs in all possible combinations in the remaining measures. This signal is a real-life sample initially proposed in [43] played on a Yamaha Disklavier MX100A piano and recorded in a small-size room by a Schoeps omnidirectional microphone. The signal is 15.6 seconds long and has a sampling frequency  $f_s = 22050\text{Hz}$  yielding  $T = 345500$  samples. STFT of the input signal  $x$  yields a temporal resolution of 23ms and a frequency resolution of 21.5Hz, so that the amplitude spectrogram  $V$  has  $N=676$  frames and  $F=513$  frequency bins.

All NMF algorithms were run for 500 iterations which allowed them to converge. Figure 3.6 presents the results obtained for  $W$  and  $H$  with a factorization rank  $K = 7$ , hence overestimated.

We observe that min-vol KL-NMF LS automatically sets two components to zero while 5 source estimates are determined. We observe that the notes estimation is equivalent to the ones presented in [43]. We get four notes and one component that corresponds to residual noise and transient events (as such, the hammer hits and pedal releases).

Note that using baseline KL-NMF or sparse KL-NMF led to same conclusions as for the two first audio samples; these two algorithms generate as many source estimates as imposed by the rank of factorization while min-vol KL-NMF LS algorithm preserves the integrity of the four notes.

For illustration purposes, the audio files for the source estimates obtained with Algorithm 3 are available online <sup>3</sup>.

<sup>3</sup><https://www.dropbox.com/sh/i9hgcl8g1fviq3t/AAARYAYfGdN65UsgzIMto5roa?dl=0>



**Fig. 3.7.** Musical score of the third audio sample.

**Bass and drums** The third audio signal is a synthetic mix of a bass and drums<sup>4</sup>. The audio signal is downsampled to  $f_s=16000\text{Hz}$  yielding  $T=104821$  samples. STFT of the input signal  $x$  yields a temporal resolution of 32ms and a frequency resolution of 15.62Hz, so that the amplitude spectrogram  $V$  has  $N=206$  frames and  $F=513$  frequency bins. For this synthetic mix, we have access to the true sources under the form of two audio files. Therefore, we can estimate the quality of the separation with standard metrics, namely the signal to distortion ratios (SDR), the source to interference ratios (SIR) and the sources to artifacts ratios (SAR) [136]. They have been computed with the toolbox BSS Eval<sup>5</sup>. The metrics are expressed in dB and the higher they are the better is the separation. Algorithms min-vol KL-NMF LS, baseline KL-NMF and sparse KL-NMF have been considered for this comparative study. A factorization rank equal to two is used. It is clear that the rank-one approximation is too simplistic for these sources but the goal is to compare the algorithms and show that min-vol KL-NMF LS is able to find a better solution even in this simplified context.

All NMF algorithms were run for 400 iterations which allowed them to converge. Table 3.3 shows the results.

Table 3.3: SDR, SIR and SAR metrics comparison for results obtained with baseline KL-NMF and min-vol KL-NMF LS on a synthetic mix of bass and drums

Algorithms	Source 1: bass			Source 2: drums		
	SDR(dB)	SIR(dB)	SAR(dB)	SDR(dB)	SIR(dB)	SAR(dB)
min-vol KL-NMF LS	<b>-1.14</b>	<b>0.12</b>	<b>7.78</b>	<b>9.60</b>	<b>19.8</b>	10.09
baseline KL-NMF	-4.26	-1.39	2.64	7.97	9.00	<b>15.25</b>
sparse KL-NMF	-4.69	-1.73	2.33	7.89	8.96	14.98

Except for SAR metric for the second source (drums), min-vol KL-NMF LS outperforms baseline KL-NMF and sparse KL-NMF.

**Runtime performance** Let us compare the runtime of baseline KL-NMF, min-vol KL-NMF LS (Algorithm 3) and sparse KL-NMF [119]. The algorithms are compared on the three examples presented in paragraphs 3.4 and 3.4:

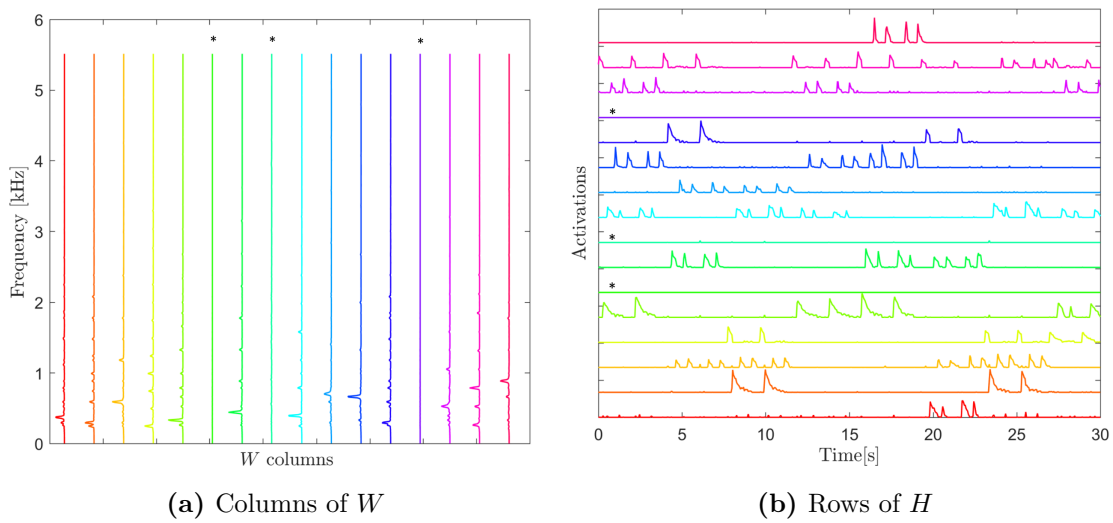
- Setup #1: sample “Mary had a little lamb” with  $K = 3$ , 200 iterations.

<sup>4</sup><http://isse.sourceforge.net/demos.html>

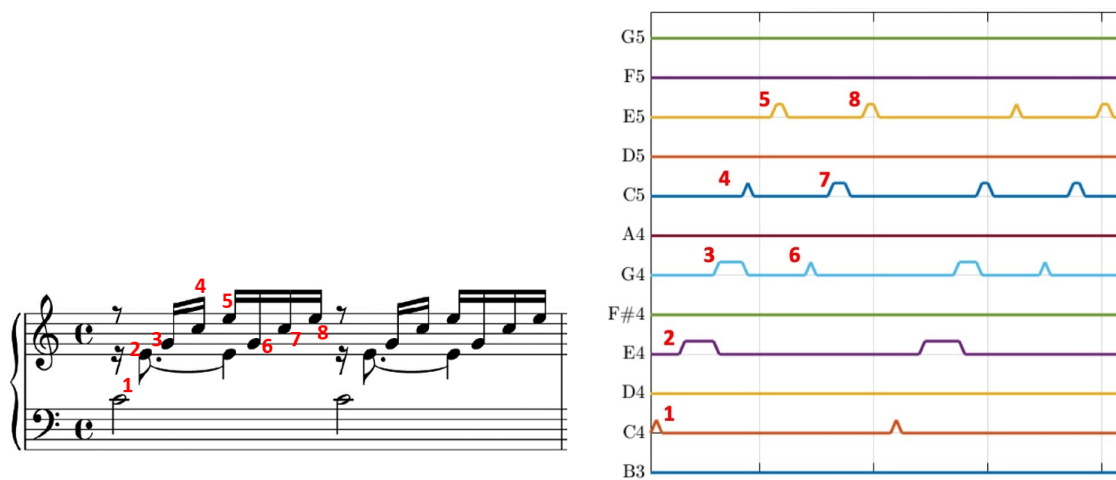
<sup>5</sup>[http://bass-db.gforge.inria.fr/bss\\_eval/](http://bass-db.gforge.inria.fr/bss_eval/)



**Fig. 3.8.** Musical score of the sample “Prelude and Fugue No.1 in C major”.



**Fig. 3.9.** Factors matrices  $W$  and  $H$  obtained with min-vol KL-NMF LS with factorization rank  $K=16$  on the sample “Prelude and Fugue No.1 in C major”.



**Fig. 3.10.** Validation of the estimate sequence obtained with min-vol KL-NMF LS with factorization rank  $K=16$  on the sample “Prelude and Fugue No.1 in C major”.

- Setup #2: sample “Mary had a little lamb” with  $K = 7$ , 200 iterations.
- Setup #3: “Prelude and Fugue No.1 in C major” with  $K = 16$ , 300 iterations.

For each test setup, the algorithms are run for the same 20 random initializations of  $W$  and  $H$ . Table 3.4 reports the average and standard deviation of the runtime (in seconds) over these 20 runs. We observe that the runtime of min-vol KL-NMF LS (Algorithm 3) is slower but not significantly so, as expected. In particular, on the larger setup #3, it is less than three times slower than the standard MU.

Table 3.4: Runtime performance in seconds of baseline KL-NMF, min-vol KL-NMF LS (Algorithm 3) and sparse KL-NMF [119]. The table reports the average and standard deviation over 20 random initializations for three experimental setups described in the text.

Algorithms	runtime in seconds		
	setup #1	setup #2	setup #3
baseline KL-NMF	0.44±0.03	0.43±0.01	3.81±0.19
min-vol KL-NMF LS	3.79±0.13	2.39±0.30	10.19±1.28
sparse KL-NMF	0.20±0.02	0.20±0.01	2.21± 0.26

### 3.5 Conclusion and Perspectives

In this chapter, we have presented a new NMF problem for audio source separation based on the minimization of a cost function that includes a  $\beta$ -divergence (data fitting term) and a penalty term that promotes solutions  $W$  with minimum volume. We have proved the identifiability of the problem in the exact case, under the sufficiently scattered condition for the activation matrix  $H$ . We have provided multiplicative updates to tackle this problem and have illustrated the behaviour of the method on real-world audio signals. We highlighted the capacity of the model to deal with the case where  $K$  is overestimated by setting automatically to zero some components and give good results for the source estimates.

Further work includes tackling the following questions:

- Under which conditions can we prove the identifiability of min-vol  $\beta$ -NMF in the presence of noise, and the rank-deficient case?
- Can we prove that min-vol  $\beta$ -NMF performs model order selection automatically? Under which conditions? We have observed this behaviour on many examples, but the proof remains elusive.
- Can we design more efficient algorithms? For this question, the answer is yes. In Chapter 5, we develop a general framework to tackle  $\beta$ -divergences NMF problems

under linear disjoint constraints. Among others, we derive a new algorithm to tackle problem 3.1.

Further work also includes the use of our new problem and derived algorithms for other applications.

**Acknowledgments** We thank Kejun Huang and Xiao Fu for helpful discussion on Theorem 3.2.1, and giving us the insight to adapt their proof from [49] to our problem (3.1).

## 4 Multi-resolution $\beta$ -NMF for blind spectral unmixing

Blind spectral unmixing is the problem of decomposing the spectrum of a mixed signal or image into a collection of source spectra and their corresponding activations indicating the proportion of each source present in the mixed spectrum. To perform this task, nonnegative matrix factorization (NMF) based on the  $\beta$ -divergence, referred to as  $\beta$ -NMF, is a standard and state-of-the-art technique. Many NMF-based methods factorize a data matrix that is the result of a resolution trade-off between two adversarial dimensions. Two instrumental examples are (1) audio spectral unmixing for which the frequency-by-time data matrix is computed with the short-time Fourier transform and is the result of a trade-off between the frequency resolution and the temporal resolution, and (2) BHU for which the wavelength-by-location data matrix is a trade-off between the number of wavelengths measured and the spatial resolution. In this chapter, we propose a new NMF-based method, dubbed multi-resolution  $\beta$ -NMF (MR- $\beta$ -NMF), to address this issue by fusing the information coming from multiple data with different resolutions in order to produce a factorization with high resolutions for all the dimensions. MR- $\beta$ -NMF relies on a nonnegative joint factorization. To achieve this goal, we propose an optimization problem including the  $\beta$ -divergences as objective functions. In order to solve this problem, we propose multiplicative updates based on a majorization-minimization algorithm. We show on numerical experiments that MR- $\beta$ -NMF is able to obtain high resolutions in both dimensions for two applications: the joint-factorization of two audio spectrograms, and the hyperspectral and multispectral data fusion problem.

The content of this chapter is extracted from: [91]: V. Leplat, N. Gillis, and C. Févotte. *Multi-Resolution Beta-Divergence NMF for Blind Spectral Unmixing*. 2020. arXiv: 2007.03893.

### 4.1 Introduction

As introduced in Section 1.5, spectral unmixing concerns the techniques used to decompose the spectrum of a mixed signal into a set of source spectra and their corresponding activations. The activations give the proportion of each source spectrum present in the mixed spectrum. More specifically, blind spectral unmixing consists in estimating the source spectra with limited prior information; usually, the only known information is the number of sources. Spectral unmixing techniques are applied in many fields such as in audio and



image processing. In this chapter, we introduce a flexible framework to perform spectral unmixing by fusing the information coming from multiple data with different resolutions. We showcase its efficiency on two major applications: audio spectral unmixing and fusion of hyperspectral and multispectral images. For these applications, the input data usually results from a trade-off between two adversarial dimensions. Let us illustrate this assertion in the particular case of audio spectral unmixing that commonly uses the simultaneous time-frequency representation of an input mixed signal. The simultaneous time-frequency representation is here computed with the short-time Fourier transform (STFT).

As explained in Section 1.5.2, the STFT consists in dividing the time signal into short segments of the same length, in multiplying the segments element-wise by a window function of size  $2F$ , and then in computing the Fourier transform of each windowed segment (only half of the frequency coefficients can be retained thanks to the Hermitian symmetry). Therefore, from an input signal  $u \in \mathbb{R}^T$ , we obtain a complex matrix  $U \in \mathbb{C}^{F \times N}$  called spectrogram. The number of rows corresponds to the frequency resolution. Letting  $f_s$  be the sampling rate, consecutive rows correspond to frequency bands that are  $\frac{f_s}{2F}$  Hz apart. Choosing a particular value for the window length  $2F$  is equivalent to fixing the frequency and the time resolutions. A larger window implies a higher frequency resolution but it comes at the cost of lower temporal resolution. Moreover, the trade-off between detailed frequency and temporal information is due to the fundamental physical limit known as the Heisenberg uncertainty principle. A natural solution is to consider multiple audio spectrograms and fuse them into a product with both high frequency and high temporal resolutions. A similar idea has been studied in the hyperspectral imaging community; see for example [114, 54, 25, 26, 110, 116, 4, 141, 77]. Indeed, hyperspectral (HS) images have high spectral resolution (typically between 100 and 200 spectral bands) but low spatial resolution, whereas the opposite is true for multispectral (MS) images. The fusion of HS and MS data, which we refer to as the HS-MS fusion problem, gives the possibility to produce fused data with both high spectral and high spatial resolutions, called the super-resolution (SR) image. The SR image can improve the precision of the unmixing [149].

**Contribution and outline** We propose multi-resolution  $\beta$ -NMF (MR- $\beta$ -NMF) for fusing the information coming from multiple audio amplitude spectrograms with different frequency resolutions. As far as we know, it is the first time such an approach is used in this context. High-frequency-resolution data and high-temporal-resolution data are jointly factorized by MR- $\beta$ -NMF, taking into account the linear mixture model (4.3). Based on these audio spectrograms, we are able to generate a solution  $W$  that exploits the spectral accuracy from the high-frequency-resolution data and a solution for  $H$  exploiting the temporal accuracy from the high-temporal-resolution data. Both frequency and temporal reconstruction qualities are evaluated by numerical simulation using synthetic audio signals. We also show that MR- $\beta$ -NMF is flexible and can be used in other applications. In particular we motivate and show its efficiency to deal with the HS-MS fusion problem. As far as we know, it is the first time that a HS-MS fusion model and algorithm tackles any

$\beta$ -divergence. Most previous works focused on the case  $\beta = 2$ , that is, least squares, which assumes Gaussian noise as a prior. As we will see, considering  $\beta$ -divergences for  $\beta \neq 2$  allows to obtain much better solutions in the presence of non-Gaussian noise. In particular, we show that in the presence of Poisson noise, using  $\beta = 1$  (Kullback-Leibler divergence) outperforms standard approaches.

This chapter is organized as follows. Section 4.2 details the problem formulation, in particular the mixture model and the (optimization) problem for MR- $\beta$ -NMF. Section 4.3 describes the algorithm developed to tackle this problem. Section 4.4 (resp. Section 4.5) presents numerical results on audio datasets (resp. on the HS-MS fusion problem). MR- $\beta$ -NMF is shown to be competitive with state-of-the-art techniques, and allows to obtain solutions with both high spectral resolution and high temporal (resp. spatial) resolution.

## 4.2 Problem formulation

The aim of multi-resolution unmixing, or more generally data fusion, is to estimate non-observable data with high resolutions in adversarial dimensions from observable data that show high resolution in one dimension only. In this chapter, we propose a flexible framework that can be easily adapted to many applications. More particularly, we consider the blind audio spectral unmixing and the HS-MS fusion problem.

In the case of the audio spectral unmixing, the multi-resolution unmixing is based on high-frequency-resolution (HRF) data and low-frequency-resolution (LRF) data. In this chapter we limit the discussion to the use of two input audio amplitude spectrograms  $X \in \mathbb{R}_+^{F_X \times N_X}$  and  $Y \in \mathbb{R}_+^{F_Y \times N_Y}$ .

We assume they are computed with STFTs based on a common input audio signals  $u$ . The windows lengths are respectively  $F_X$  and  $F_Y$  such that  $F_Y > F_X$  with  $\frac{F_Y}{F_X} = d$  where  $d$  is usually referred to as the frequency downsampling ratio. Sizes  $N_X$  and  $N_Y$  denote the number of time frames of LRF and HRF spectrograms, respectively, with  $N_X > N_Y$  as per the trade-off between frequency and temporal resolutions. Given  $X$  and  $Y$ , we are searching for an amplitude audio spectrogram  $V \in \mathbb{R}_+^{F_Y \times N_X}$  that has both high frequency and high temporal resolutions. We suppose in this chapter that the observed LRF spectrogram  $X$  is a frequency downsampled version of  $V$ , that is,

$$X \approx RV, \quad (4.1)$$

where  $R \in \mathbb{R}^{F_X \times F_Y}$  is the frequency downsampling operator. Similarly, the observed HRF spectrogram  $Y$  is a temporally downsampled version of  $V$ , that is,

$$Y \approx VS, \quad (4.2)$$

where  $S \in \mathbb{R}^{N_X \times N_Y}$  is the temporal downsampling operator. For the HS-MS fusion problem, we assume that a high spatial resolution image  $X$  and a high spectral resolution image  $Y$  are available to reconstruct the target image with high-spectral and high-spatial

resolutions  $V$ , the SR image. These images result from the linear spectral and spatial degradations of the SR image  $V$ , given by the same equations (4.1) and (4.2). In this context, the operator  $R$  from (4.1) is the relative spectral bandpass responses from the super-resolution image to the MS image, while the operator  $S$  introduced in (4.2) specifies the spatial blurring and down-sampling responses that result in the HS image. Hence both operators are nonnegative matrices. In the context of HS-MS fusion, the operators  $R$  and  $S$  can be acquired either by cross-calibration [148] or by estimations from the HS and MS images [149, 124]. As far as we know, in the context of audio spectral unmixing, it is unknown how to estimate  $R$  and  $S$ , and we will propose an optimization strategy to do so.

### Linear Spectral Mixture Model

A linear spectral mixture model is commonly used for the audio spectral unmixing or HS unmixing due to its physical meaning and its mathematical simplicity; see sections 1.5.1, 1.5.2 and references [21, 126, 147] for detailed reviews. Under this model and assuming we have noise in the data, the input data matrix  $V$  has the form

$$V \approx WH, \quad (4.3)$$

where  $W \in \mathbb{R}_+^{F_Y \times K}$  is the dictionary matrix and  $H \in \mathbb{R}_+^{K \times N_X}$  is the activation matrix.

Substituting (4.3) into (4.1) and (4.2),  $X$  and  $Y$  are expressed as follows:

$$X \approx RWH, \quad (4.4)$$

$$Y \approx WHS. \quad (4.5)$$

where  $R \in \mathbb{R}_+^{F_X \times F_Y}$  and  $S \in \mathbb{R}_+^{N_X \times N_Y}$ . Equations (4.4) (resp. (4.5)) correspond to the linear spectral mixture model degraded in the frequency (resp. spectral) and temporal (resp. spatial) domains. This leads to our proposed NMF approach described in the next section.

### Multi-Resolution $\beta$ -NMF problem

In this section, we present a new approach for spectral unmixing based on the minimization of  $\beta$ -divergences. To solve the multi-resolution problem and estimate the signal  $V$ , we need to estimate  $W$  and  $H$ . From (4.4) and (4.5), we propose to solve the following optimization problem, which we refer to as MR- $\beta$ -NMF problem:

$$\min_{W \geq 0, H \geq 0, R \geq 0, S \geq 0} D_\beta(X|RWH) + \lambda D_\beta(Y|WHS), \quad (4.6)$$

where  $A \geq 0$  means that  $A$  is component-wise nonnegative,  $\lambda$  is a positive penalty parameter, and  $D_\beta(Z|ABC) = \sum_{fn} d_\beta(Z_{fn}|[ABC]_{fn})$  with  $d_\beta(x|y)$  is the  $\beta$ -divergence between scalars  $x$  and  $y$ , see Section 1.9.1 for more details. In the general case, MR- $\beta$ -NMF is also able to estimate the downsampling operators  $R$  and  $S$ , which is a contribution. Note that when the downsampling operators  $R$  and  $S$  are known, the objective function is

minimized over  $W$  and  $H$  only, see section 4.3 for more details. Note also that in general  $R$  and  $S$  have a particular structure where some entries are fixed to zero; see Section 4.2. As our algorithm will rely on multiplicative updates, entries initialized at zero remain zero in the course of the optimization process. As explained in sections 1.9.2 and 1.9.3, the error measure should be chosen depending on the noise statistic assumed on the data. Let us briefly recall that the Euclidean distance ( $\beta = 2$ ) assumes i.i.d. Gaussian noise, KL divergence ( $\beta = 1$ ) assumes Poisson noise, and the IS divergence ( $\beta = 0$ ) assumes multiplicative noise following exponential distributions. KL and IS divergences are usually considered for amplitude spectrogram and power spectrogram, respectively. Both KL and IS divergences are more adapted to audio spectral unmixing than Euclidean distance; see [88] and [45]. The Euclidean distance is the most widely used to tackle the HS unmixing problem as well as the HS-MS data fusion problem. However, when no obvious choice of a specific divergence is available, finding the right measure of fit, namely the value for  $\beta$ , can be seen as a model selection problem [44]. Therefore an objective function with an adjustable  $\beta$  is fully justified. Moreover, divergences are often log-likelihoods in disguise (see Section 1.9.2) and therefore choosing a divergence boils down to choosing a noise statistic as mentioned earlier. For example, sensors embedded in cameras can be seen as photon counters, and the Poisson distribution makes particular sense for count data. This assumption supports once again our motivation to consider an adjustable  $\beta$ , in this case with  $\beta = 1$ . Based on numerical experiments, we will show that the KL-divergence is also well suited for the HS-MS fusion problem.

### Downsampling operators

As mentioned earlier, for HS-MS data fusion, downsampling operators can usually be estimated and hence are assumed to be known. In the context of audio spectral unmixing, the downsampling operators in (4.6) are unknown. Different structures for downsampling operators  $R$  and  $S$  have been tested, and we report here the form for  $R$  that shows the best results in practice, while  $S$  is obtained in the same way. Let us illustrate this on the simple example of the frequency downsampling of a matrix  $W \in \mathbb{R}^{8 \times 3}$  with a downsampling ratio  $d = 2$ . A possible structure for the matrix  $R \in \mathbb{R}_+^{4 \times 8}$  is as follows:

$$R = \begin{pmatrix} \underline{r_{11}} & \underline{r_{12}} & \mathbf{r_{13}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{r_{21}} & \underline{r_{22}} & \underline{r_{23}} & \mathbf{r_{24}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{r_{31}} & \underline{r_{32}} & \underline{r_{33}} & \mathbf{r_{34}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{r_{41}} & \underline{r_{42}} & \underline{r_{43}} \end{pmatrix},$$

This downsampling operator  $R$  performs a weighted arithmetic mean over a set of rows of the matrix it is applied on; here,  $W \in \mathbb{R}_+^{8 \times 3}$  is downsampled as  $RW \in \mathbb{R}_+^{4 \times 3}$ . The structure of the matrix  $R$  relies on two parameters:  $d$  and  $f$ . As mentioned earlier,  $d$  corresponds to the downsampling ratio. Each row of  $R$  has at least  $d$  non-zero values that correspond to the rows in  $W$  that are combined to form the rows of  $RW$ ; see the underlined entries of  $R$

above. The parameter  $f$  controls the overlap between the linear combinations of the rows of  $W$ .

In the example above,  $f = 1$  and one positive value is added to the left and the right end of the  $d$  non-zero entries corresponding to the downsampling parameter; see the bold entries in matrix  $R$  above. These positive values allow an overlap (or coupling) within the downsampling process. If we consider two consecutive frequency bins that result from a downsampling operation, it is reasonable to consider that they share common frequency bins in the original frequency space. We imposed  $f \leq \frac{d}{2}$  to avoid too much non-physical coupling. This limitation is also based on numerical experiments that show a degradation of the results when  $f$  exceeds  $\frac{d}{2}$ . When  $f = 0$ , the downsampling operator  $R$  performs a weighted arithmetic mean over  $d$  rows without overlapping. Note that these downsampling operators are sparse nonnegative matrices.

### 4.3 Algorithm for MR- $\beta$ -NMF problem

As explained in Section 1.10 many state-of-the-art NMF optimization algorithms rely on a BCD scheme by optimizing alternatively over  $W$  for  $H$  fixed and vice versa, and we adopt this approach in this chapter. Recall that the  $\beta$ -divergence is only convex with respect to its second argument when  $\beta \in [1, 2]$ . The goal in this section is to derive an algorithm to tackle MR- $\beta$ -NMF problem (4.6). For  $R, S$  and  $W$  fixed, let us consider the subproblem in  $H$ :

$$\min_{H \geq 0} L(H) = D_\beta(X|RWH) + \lambda D_\beta(Y|WHS). \quad (4.7)$$

As we will see, the subproblems in  $W, R$  and  $S$  for the other variables fixed are similar. To tackle this problem, we follow the standard majorization-minimization framework [130]. We start by constructing an auxiliary function denoted  $\bar{L}$  which is a tight upper-bound for the objective  $L$  at the current iterate, see Definition 3.3.1. The optimization problem with  $L$  is then replaced by a sequence of simpler problems for which the objective is  $\bar{L}$ . The new iterate  $H^{(i+1)}$  is computed by minimizing the auxiliary function at the previous iterate  $H^{(i)}$ , either approximately or exactly. This guarantees  $L$  to decrease at each iteration as per Lemma 3.3.1.

The most difficult part in using the majorization-minimization framework is to design an auxiliary function that is easy to optimize. Usually such auxiliary functions are separable (that is, there is no interaction between the variables so that each entry of  $H$  can be updated independently) and convex. We will construct an auxiliary function for  $L(H)$  from (4.7) by a positive linear combination of two auxiliary functions, one for each term of  $L(H)$ .

#### Separable auxiliary function for the first term of $L(H)$

The function  $D_\beta(X|RWH)$  separates into  $\sum_n D_\beta(x_n|RW h_n)$ , where  $x_n$  and  $h_n$  are the  $n$ th column of  $X$  and  $H$  respectively. Therefore we only consider the optimization over one

specific column  $x$  of  $X$  and  $h$  of  $H$ . To simplify notation, we denote the current iterate as  $\tilde{h}$ .

We now use the separable auxiliary function presented in [45] which consists in majorizing the convex part of the  $\beta$ -divergence using Jensen's inequality and majorizing the concave part by its tangent (first-order Taylor approximation). Note that the divergence can always be expressed as the sum of a convex, concave, and constant part, such that:

$$d_\beta(x|y) = \check{d}_\beta(x|y) + \hat{d}_\beta(x|y) + \bar{d}_\beta(x|y),$$

where  $\check{d}$  is convex function of  $y$ ,  $\hat{d}$  is a concave function of  $y$  and  $\bar{d}$  is a constant of  $y$ , see [45] for the definition of these terms for different values of  $\beta$ .

By denoting  $RW$  by  $P$  and  $RW\tilde{h}$  by  $\tilde{x}$  with entries  $[RW\tilde{h}]_f = \tilde{x}_f$  for  $f \in [1, F_X]$ , the auxiliary function for  $\sum_f d_\beta(x_f | [Ph]_f)$  at  $\tilde{h}$  is given by:

$$\begin{aligned} G_X(h|\tilde{h}) &= \sum_f^{F_X} \left[ \sum_k \frac{p_{fk}\tilde{h}_k}{\tilde{x}_f} \check{d}_\beta(x_f|\tilde{x}_f \frac{h_k}{\tilde{h}_k}) \right] + \bar{d}_\beta(x_f|\tilde{x}_f) \\ &+ \left[ \check{d}_\beta(x_f|\tilde{x}_f) \sum_k p_{fk}(h_k - \tilde{h}_k) + \hat{d}_\beta(x_f|\tilde{x}_f) \right]. \end{aligned} \quad (4.8)$$

Therefore the function

$$G_X(H|\tilde{H}) = \sum_n G_X(h_n|\tilde{h}_n) \quad (4.9)$$

is an auxiliary function (convex and separable) for  $D_\beta(X|RWH)$  at  $\tilde{H}$  where  $G_X(h|\tilde{h})$  is given by (4.8).

### Separable auxiliary function for the second term of $L(H)$

Let  $\tilde{y}_{fn} = [WHS]_{fn}$ :

$$\begin{aligned} D_\beta(Y|WHS) &= \sum_{fn} d_\beta(y_{fn} | [WHS]_{fn}) \\ &= \sum_{fn} d_\beta(y_{fn} | \sum_{kj} w_{fk} h_{kj} s_{jn}) \\ &= \sum_{fn} d_\beta(y_{fn} | \sum_{kj} (w_{fk} s_{jn}) h_{kj}) \\ &= \sum_{fn} \check{d}_\beta(y_{fn} | \sum_{kj} (w_{fk} s_{jn}) h_{kj}) \\ &\quad + \sum_{fn} \hat{d}_\beta(y_{fn} | [WHS]_{fn}) \\ &\quad + \sum_{fn} \bar{d}_\beta(y_{fn} | [WHS]_{fn}) \end{aligned} \quad (4.10)$$

where the indices  $f \in [1, F_Y]$ ,  $n \in [1, N_Y]$  and  $j \in [1, N_X]$ . If we introduce  $\tilde{\lambda}_{fkn} = \frac{(w_{fk}s_{jn})\tilde{h}_{kj}}{\sum_{kj} w_{fk}s_{jn}\tilde{h}_{kj}}$ , then (4.10) can be written as follows:

$$\begin{aligned} D_\beta(Y|WHS) &= \sum_{fn} \check{d}_\beta(y_{fn} | \sum_{kj} \tilde{\lambda}_{fkn} \frac{(w_{fk}s_{jn})\tilde{h}_{kj}}{\tilde{\lambda}_{fkn}}) \\ &\quad + \sum_{fn} \hat{d}_\beta(y_{fn} | [WHS]_{fn}) \\ &\quad + \sum_{fn} \bar{d}_\beta(y_{fn} | [WHS]_{fn}). \end{aligned} \quad (4.11)$$

Let  $\tilde{y}_{fn} = [WHS]_{fn}$  and let us remark that  $\sum_{kj} \tilde{\lambda}_{fkn} = 1$ . Therefore we can majorize the convex part of (4.11) using Jensen's inequality and majorize the concave part of (4.11) by its first-order Taylor approximation and we get the following function:

$$\begin{aligned} G_Y(H|\tilde{H}) &= \sum_{f,n} \left[ \sum_{k,j} \frac{(w_{fk}s_{jn})\tilde{h}_{kj}}{\tilde{y}_{fn}} \check{d}_\beta(y_{fn} | \tilde{y}_{fn} \frac{h_{kj}}{\tilde{h}_{kj}}) \right] \\ &\quad + \bar{d}_\beta(y_{fn} | \tilde{y}_{fn}) + \hat{d}_\beta(y_{fn} | \tilde{y}_{fn}) \\ &\quad + \check{d}'_\beta(y_{fn} | \tilde{y}_{fn}) \sum_{k,j} w_{fk}(h_{kj} - \tilde{h}_{kj})s_{jn}. \end{aligned} \quad (4.12)$$

In [45], the authors show that (4.12) is an auxiliary function (separable and convex) to  $D_\beta(Y|WHS)$  at  $\tilde{H}$ . Indeed  $G_Y(H|\tilde{H})$  is an upper-bound to  $D_\beta(Y|WHS)$  at  $\tilde{H}$  by construction and is tight when  $H = \tilde{H}$ .

### Auxiliary function for multi-resolution $\beta$ -NMF

Based on the auxiliary functions presented in previous sections, we can directly derive a separable auxiliary function  $\bar{F}(H|\tilde{H})$  for multi-resolution  $\beta$ -NMF (4.7).

**Corollary 4.3.0.1.** *For  $H \geq 0$ ,  $\lambda > 0$ , the function*

$$\bar{L}(H|\tilde{H}) = G_X(H|\tilde{H}) + \lambda G_Y(H|\tilde{H}),$$

where  $G_X$  is given by (4.9) and  $G_Y$  by (4.12), is a convex and separable auxiliary function for  $L(H) = D_\beta(X|RWH) + \lambda D_\beta(Y|WHS)$ .

*Proof.* This follows directly from (4.9) and (4.12). □

### Algorithm for MR- $\beta$ -NMF

Given the convexity and the separability of the auxiliary function, the optimum is obtained by canceling the gradient. The derivative of the auxiliary function  $\bar{L}(H|\tilde{H})$  with respect to a specific coefficient  $h_{kz}$ , with index  $z$  identifying the same column specified by  $n$  in (4.8)

and specified by  $j$  in (4.12), is given by:

$$\begin{aligned}
 \nabla_{h_{kz}} \bar{L} &= \nabla_{h_{kz}} G_X(H|\tilde{H}) + \lambda \nabla_{h_{kz}} G_Y(H|\tilde{H}) \\
 &= \sum_f^{F_X} p_{fk} \left[ \tilde{d}'_{\beta} \left( x_{fz} | \tilde{x}_{fz} \frac{h_{kz}}{\tilde{h}_{kz}} \right) + \tilde{d}'_{\beta}(x_{fz} | \tilde{x}_{fz}) \right] \\
 &\quad + \lambda \sum_f^{F_Y} \sum_n^{N_Y} w_{fk} s_{zn} \left[ \tilde{d}'_{\beta} \left( y_{fn} | \tilde{y}_{fn} \frac{h_{kz}}{\tilde{h}_{kz}} \right) \right. \\
 &\quad \left. + \tilde{d}'_{\beta}(y_{fn} | \tilde{y}_{fn}) \right].
 \end{aligned} \tag{4.13}$$

For  $\beta = 1$ , (4.13) becomes:

$$\begin{aligned}
 \nabla_{h_{kz}} \bar{L} &= \sum_f^{F_X} p_{fk} \left[ 1 - \frac{x_{fz} \tilde{h}_{kz} \tilde{x}_{fz}^{-1}}{h_{kz}} \right] \\
 &\quad + \lambda \sum_f^{F_Y} \sum_n^{N_Y} w_{fk} s_{zn} \left[ 1 - \frac{y_{fn} \tilde{h}_{kz} \tilde{y}_{fn}^{-1}}{h_{kz}} \right].
 \end{aligned} \tag{4.14}$$

We set (4.14) to zero and get the following closed-form solution for the  $h_{kz}$  coefficient of  $H$ :

$$h_{kz} = \tilde{h}_{kz} \frac{\sum_f^{F_X} p_{fk} x_{fz} \tilde{x}_{fz}^{-1} + \lambda \sum_f^{F_Y} \sum_n^{N_Y} w_{fk} s_{zn} y_{fn} \tilde{y}_{fn}^{-1}}{\sum_f^{F_X} p_{fk} + \lambda \sum_f^{F_Y} \sum_n^{N_Y} w_{fk} s_{zn}} \tag{4.15}$$

The generalization of the closed-form solution (4.15) for any  $\beta$  for  $H$  is given in Table 4.1 in matrix forms. Table 4.1 gives also the closed-form solution for  $W$  which is derived with the same rationale. As mentioned in Section 4.2, in the general case, operators  $R$  and  $S$  are unknown. We propose here to derive updates for  $R$  and  $S$  so that these operators can be learned from the data and sensible estimates for  $W$  and  $H$  during the optimization scheme. The updates for  $R$  and  $S$  have been derived in a similar fashion as for matrices  $W$  and  $H$ . For the update of  $R$  for instance, one has simply to note it corresponds to the update of  $W$  where we only keep the terms multiplied by  $\lambda = 1$  and where the roles of  $Y$ ,  $W$ ,  $H$  and  $S$  are exchanged with  $X$ ,  $R$ ,  $W$  and  $H$ , respectively.

Algorithm 4 summarizes our method to tackle (4.6) which is referred as MR- $\beta$ -NMF. It consists in two optimization loops:

- Loop 1: matrices  $W$  and  $H$  are alternatively updated with downsampling operators  $R$  and  $S$  kept fixed so that we obtain good estimates for  $W$  and  $H$ . The updates are performed for a maximum number of iterations imposed by the parameter MAXITERL1.
- Loop 2: matrices  $W$ ,  $H$ ,  $S$  and  $R$  are alternatively updated so that the algorithm learns the downsampling operators during the optimization process. The maximum number of iterations for loop 2 is controlled by parameter MAXITERL2.



Table 4.1: Multiplicative updates for MR- $\beta$ -NMF.

$$\begin{aligned}
H &= \tilde{H} \odot \left( \frac{\left[ W^T \left( R^T \left( (RW\tilde{H})^{(\beta-2)} \odot X \right) + \lambda \left( (W\tilde{H}S)^{(\beta-2)} \odot Y \right) S^T \right) \right]}{\left[ W^T \left( R^T (RW\tilde{H})^{(\beta-1)} + \lambda (W\tilde{H}S)^{(\beta-1)} S^T \right) \right]} \right)^{\cdot\gamma(\beta)}, \\
W &= \tilde{W} \odot \left( \frac{\left[ \left( R^T \left( (R\tilde{W}H)^{(\beta-2)} \odot X \right) + \lambda \left( (\tilde{W}HS)^{(\beta-2)} \odot Y \right) S^T \right) H^T \right]}{\left[ \left( R^T (R\tilde{W}H)^{(\beta-1)} + \lambda (\tilde{W}HS)^{(\beta-1)} S^T \right) H^T \right]} \right)^{\cdot\gamma(\beta)}, \\
S &= \tilde{S} \odot \left( \frac{\left[ H^T \left( W^T \left( (WH\tilde{S})^{(\beta-2)} \odot Y \right) \right) \right]}{\left[ H^T \left( W^T (WH\tilde{S})^{(\beta-1)} \right) \right]} \right)^{\cdot\gamma(\beta)}, \\
R &= \tilde{R} \odot \left( \frac{\left[ \left( \left( (\tilde{R}WH)^{(\beta-2)} \odot X \right) H^T \right) W^T \right]}{\left[ \left( (\tilde{R}WH)^{(\beta-1)} H^T \right) W^T \right]} \right)^{\cdot\gamma(\beta)},
\end{aligned}$$

where  $\gamma(\beta) = \frac{1}{2-\beta}$  for  $\beta < 1$ ,  $\gamma(\beta) = 1$  for  $\beta \in [1, 2]$  and  $\gamma(\beta) = \frac{1}{\beta-1}$  for  $\beta > 2$  [45].

For the HS-MS fusion problem, the operators  $R$  and  $S$  are usually known and therefore the parameter MAXITERL2 is set to zero. In this chapter, the second optimization loop is considered only for the audio spectral unmixing application since the operators  $R$  and  $S$  are unknown.

After  $W$  and  $H$  are updated, we normalize  $W$  such that  $\|W(:,k)\|_1 = 1$  for all  $k$ , and we normalize  $H$  accordingly so that  $WH$  remains unchanged. This normalization is commonly used for NMF-based methods and is mainly performed to remove the scaling degree of freedom. As a convergence condition, we consider the relative change ratio of the cost function  $L$  from (4.6), namely  $|L^i - L^{i+1}| \leq \kappa L^i$  where  $\kappa$  is a given threshold, and  $i$  is the iteration counter. We also stop the optimization process if the number of iterations exceeds the predefined maximum number of iterations.

It can be verified that the computational complexity of the MR- $\beta$ -NMF is asymptotically equivalent to the standard MU for  $\beta$ -NMF, that is, it requires  $\mathcal{O}(FNK)$  operations per iteration.

### Parallel computing

We remark that some of the most computationally intensive steps of the proposed algorithm can be easily ran onto a parallel computation platform. Indeed, the complexity of our multiplicative updates detailed in Table 4.1 is mainly driven by the matrix products in which matrix  $S$  is involved. On Matlab for example, we can easily take of advantage of a GPU compatible with CUDA libraries by simply transforming usual arrays into GPU arrays. In our case, on a desktop equipped with a Intel Core<sup>TM</sup> i7-8700 CPU and a GeForce RTX 2070 Super GPU, the runtime can be up to 5 times shorter.

## 4.4 Numerical experiments on audio data sets

In this section, we perform numerical experiments to validate the effectiveness of MR- $\beta$ -NMF on two synthetic audio data sets.

**Algorithm 4** Multiplicative updates for MR- $\beta$ -NMF

**Require:** A matrix  $X \in \mathbb{R}_+^{F_X \times N_X}$ , a matrix  $Y \in \mathbb{R}_+^{F_Y \times N_Y}$ , an initialization  $H \in \mathbb{R}_+^{K \times N_X}$ , an initialization  $W \in \mathbb{R}_+^{F_Y \times K}$ , a matrix  $R \in \mathbb{R}_+^{F_X \times F_Y}$ , a matrix  $S \in \mathbb{R}_+^{N_X \times N_Y}$ , a factorization rank  $K$ , a maximum number of iterations MAXITERL1, a maximum number of iterations MAXITERL2, a threshold  $\kappa$  and a weight  $\lambda > 0$

**Ensure:** A rank- $K$  NMF  $(W, H)$  of  $V \approx WH$  with  $W \geq 0$  and  $H \geq 0$ , and operators  $R$  and  $S$  such that  $X \approx RWH$  and  $Y \approx WHS$ .

```

1: % Loop 1
2:  $i \leftarrow 0$ 
3: while  $i < \text{MAXITERL1}$  and  $\left| \frac{L^i - L^{i+1}}{L^i} \right| > \kappa$  do
4:   % Update of matrices  $H$  and  $W$ 
5:   Update  $H$  and  $W$  sequentially; see Table 4.1
6:    $(W, H) \leftarrow \text{normalize}(W, H)$ ,  $i \leftarrow i + 1$ 
7: end while
8: % Loop 2
9:  $i \leftarrow 0$ 
10: while  $i < \text{MAXITERL2}$  and  $\left| \frac{L^i - L^{i+1}}{L^i} \right| > \kappa$  do
11:   % Update of matrices  $H, W, S$  and  $R$ 
12:   Update  $H, W, S, R$  sequentially; see Table 4.1
13:    $(W, H) \leftarrow \text{normalize}(W, H)$ ,  $i \leftarrow i + 1$ 
14: end while

```

#### 4.4.1 Experimental setup and evaluation

##### Data

The proposed technique for joint-factorization of amplitude audio spectrograms is applied to two synthetic audio samples. Both music samples, respectively referred as dataset 1 and dataset 2, have been generated with a professional audio software called Sibelius based on the musical score shown in Figures 3.1 and 3.7.

The two following subsections respectively introduce a dedicated test procedure and the quantitative criteria to evaluate the performance of MR- $\beta$ -NMF algorithm.

##### Experimental comparison

This section describes the test procedure elaborated to evaluate the quality of the results obtained with MR- $\beta$ -NMF (4.6) that jointly factorizes two audio spectrograms  $X$  and  $Y$ . In the following, matrices  $W$  and  $H$  stand for the solutions computed with Algorithm 4 that solves MR- $\beta$ -NMF problem (4.6). We aim at showing that the factor  $W$  has a high frequency resolution whereas the matrix  $H$  has a high temporal resolution. To achieve this goal, we compare  $W$  to a dictionary matrix denoted  $W_Y$  computed with a baseline  $\beta$ -NMF approach that factorizes the high frequency spectrogram  $Y$  only. The baseline  $\beta$ -NMF

applied on  $Y$  solves the following optimization problem:

$$\min_{W_Y \geq 0, H_Y \geq 0} D_\beta(Y|W_Y H_Y). \quad (4.16)$$

Due to the trade-off between detailed frequency and temporal information, the activation matrix  $H_Y$  shows a low temporal resolution. To compare the accuracy of the solutions  $W$  and  $W_Y$ , we need to have access to an oracle matrix  $W_\#$  that is the reference for the comparison. For instance, for the dataset 1, each column of  $W_\#$  is supposedly the "true" spectral signature of each of the three notes;  $E_4$ ,  $D_4$  and  $C_4$ . We estimated  $W_\#$  as follows:

- We synthetically generate three audio signals and each one contains the sequence of one note in particular.
- Based on the three audio signals, we generate three amplitude spectrograms that have high frequency resolution with the same window size as the one used to generate  $Y$ .
- For each amplitude spectrogram, we perform a rank-1 NMF. The resulting  $F_Y$ -dimensional vectors are concatenated to form the oracle matrix  $W_\#$ .

We show the accuracy of  $H$  with a similar procedure;  $H$  is compared to an activation matrix  $H_X$  obtained by solving

$$\min_{W_X \geq 0, H_X \geq 0} D_\beta(X|W_X H_X), \quad (4.17)$$

using multiplicative updates. The oracle matrix  $H_\#$ , that is, the reference for the comparison, is computed by performing three independent rank-1 NMF on three amplitude spectrograms that have high temporal resolution, all generated with the same window size as the one used to generate  $X$ .

## Performance Evaluation

This section presents the qualitative criteria for evaluating the performance of the solutions obtained with Algorithm 4. We compute the following measures of reconstruction.

- *Activation matrices*: in order to avoid the scaling and permutation ambiguities inherent to the considered NMF models, we first normalize in L-1 norm the rows of the activations matrices  $H$  and solve an assignment problem w.r.t. the oracle matrix  $H_\#$ . The quality of the activation matrix  $H$  is compared to  $H_X$  w.r.t.  $H_\#$  by computing the following signal-to-noise ratios (SNR): for all  $k$ ,

$$SNR_{H_k} = 20 \log_{10} \left( \frac{\|\bar{H}(k, \cdot)\|_F}{\|\bar{H}(k, \cdot) - \bar{H}_\#(k, \cdot)\|_F} \right), \quad (4.18)$$

where  $\bar{H}(k, \cdot) = \frac{H(k, \cdot)}{\|H(k, \cdot)\|_1}$  and  $\|H(k, \cdot)\|_1 = \sum_j |H(k, j)|$ , and

$$SNR_{H_{X,k}} = 20 \log_{10} \left( \frac{\|\bar{H}_X(k, \cdot)\|_F}{\|\bar{H}_X(k, \cdot) - \bar{H}_\#(k, \cdot)\|_F} \right). \quad (4.19)$$

The higher the SNRs (4.18) and (4.19), the better is the estimation for the activation matrix.

- *Dictionary matrices:* The quality of the dictionary matrix  $W$  is evaluated in the same fashion, except that the normalization is performed by columns.

#### 4.4.2 Results

In this section, we use the following setting:

- 100 random initializations for  $W$  and  $H$  for each NMF.
- the window lengths are set to 1024 (23ms) and 4096 (93ms), then the downsampling ratio  $d$  is equal to 4. For the generation of  $R$  and  $S$ , parameter  $f$  is set to 2.
- $\beta = 1$ , and we consider the amplitude spectrograms as the input data.
- we use  $\lambda = 1$  in all our experiments.

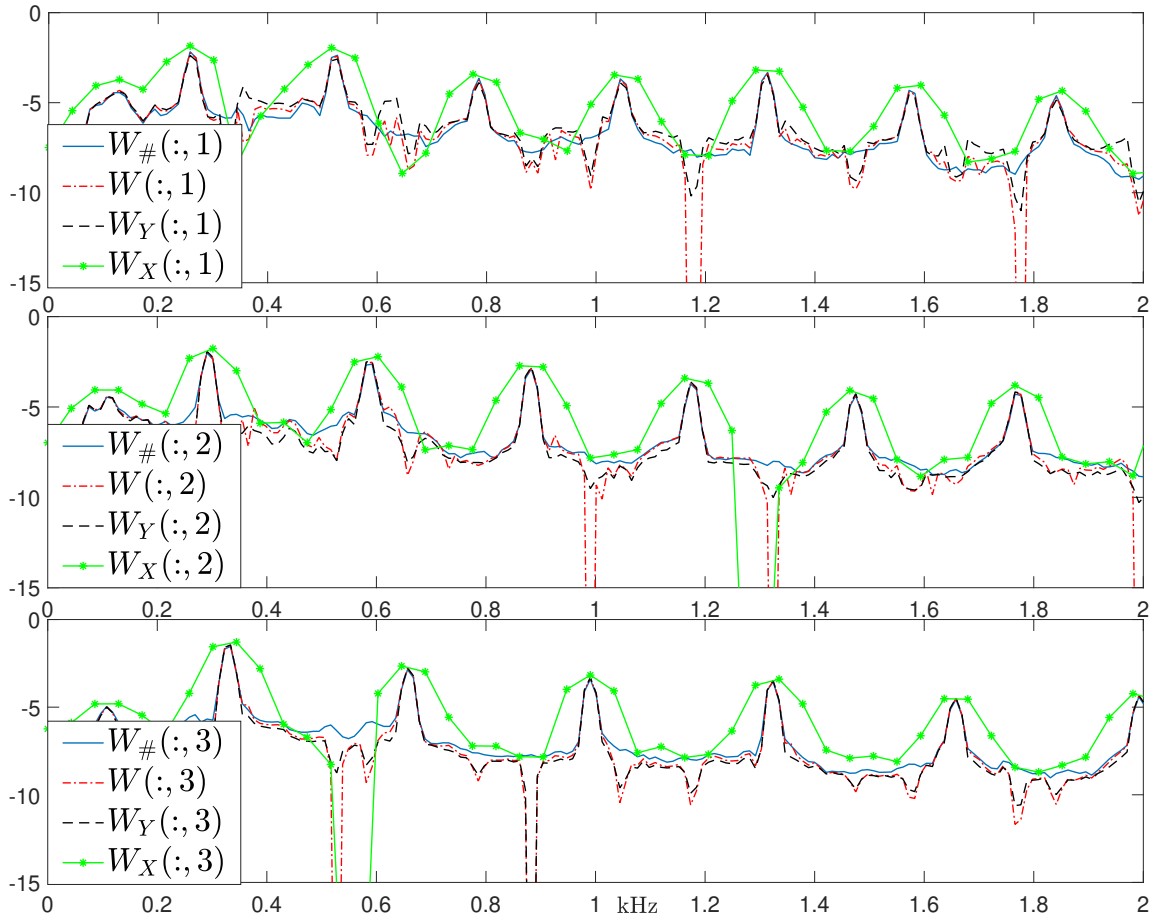
Note that we finally report qualitative numerical results obtained for a real-life recording of data set 2 used in [43]. In particular we showcase the results obtained for  $W$  and  $H$  with Algorithm 4 for two values of  $\beta$ , namely  $\beta = 0$  and  $\beta = 1$  with increasing values of penalty weight  $\lambda$ . Algorithm 4 has been implemented using Matlab R2018a, the code is available from <https://bit.ly/2J6dKtc>.

#### Dataset 1: "Mary had a little lamb"

In this section we report the numerical results obtained after the completion of the test set up presented in Section 4.4.1, and using MAXITERL1=100 and MAXITERL2=400. for Algorithm 4.

Table 4.2 reports the average SNR, the standard deviation and the best SNR computed for the activations and dictionary vectors obtained with the models described in Section 4.4.1 over the 100 initializations. As it can be observed, activations  $H$  are slightly better than activations  $H_X$ , and with a significant lower standard deviation for each note. The results for the dictionary are even more conclusive; MR- $\beta$ -NMF outperforms baseline NMF (4.16) for which the SNR (best case) can be up to two times larger. Moreover, the standard deviations of MR- $\beta$ -NMF are significantly lower than those obtained with baseline NMF (4.16). It appears that the second term in the objective function in (4.6) acts as a regularizer so that MR- $\beta$ -NMF is more robust to different initializations.

Figure 4.1 shows the dictionary matrices  $W_{\#}$ ,  $W$ ,  $W_Y$  and  $W_X$ . For more clarity, the frequency range is limited to 2 kHz. This limited range includes all the most significant peaks in terms of magnitude. We observe that all the frequency peaks are accurately estimated by MR- $\beta$ -NMF for each note. Figure 4.1 also integrates the dictionary matrix  $W_X$  to highlight the impact of using baseline NMF (4.17) that uses a higher temporal resolution.



**Fig. 4.1.** Columns of  $W_{\#}$ ,  $W$ ,  $W_Y$  and  $W_X$  in semi-log scale. Top, middle and bottom sub-figures show the spectral content respectively for  $C_4$ ,  $D_4$  and  $E_4$ .

We conclude that MR- $\beta$ -NMF is able to obtain more robust and more accurate results than baseline  $\beta$ -NMFs that factorize a single spectrogram.

Table 4.2: Comparison of MR- $\beta$ -NMF with baseline  $\beta$ -NMF in terms of SNR on the activations and the dictionary vectors with respect to true factors on the dataset 1. The table reports the average, standard deviation and the best SNR over 100 random initializations for  $W$  and  $H$ . Bold numbers indicate the highest SNR.

Note	Activation SNR's (dB)				Basis SNR's (dB)			
	$SNR_{H_k}$		$SNR_{H_{X,k}}$		$SNR_{W_k}$		$SNR_{W_{Y,k}}$	
	average $\pm$ std	best	average $\pm$ std	best	average $\pm$ std	best	average $\pm$ std	best
$C_4$	12.33 $\pm$ 0.17	<b>12.74</b>	3.89 $\pm$ 8.99	12.19	21.35 $\pm$ 1.77	<b>22.66</b>	7.95 $\pm$ 7.84	12.38
$D_4$	14.50 $\pm$ 0.08	<b>14.62</b>	8.57 $\pm$ 6.44	14.38	21.25 $\pm$ 0.35	<b>21.61</b>	14.71 $\pm$ 6.06	18.23
$E_4$	19.68 $\pm$ 0.04	<b>19.82</b>	15.28 $\pm$ 5.06	19.74	22.71 $\pm$ 0.36	<b>23.02</b>	19.36 $\pm$ 2.02	20.66

## Dataset 2

In this section we report the numerical results obtained for dataset 2, using MAXITERL1=500 and MAXITERL2=1500 for Algorithm 4.

Table 4.3 reports the average SNR, the standard deviation and the best SNR computed for activations and dictionary vectors obtained with the methods described in 4.4.1 over 100 initializations. We observe that:

- MR- $\beta$ -NMF algorithm provides results that show high resolutions in both frequency and temporal domains,
- the regularization effect of MR- $\beta$ -NMF w.r.t. baseline NMFs is less stunning than observed for dataset 1. However the standard deviations obtained with MR- $\beta$ -NMF for the dictionary are significantly lower than those obtained with the baseline NMFs.
- by looking more accurately at the results for the dictionary, MR- $\beta$ -NMF globally performs better than baseline NMFs. For the activations, baseline NMFs perform slightly better than MR- $\beta$ -NMF for three scores, with an improvement of at most 1.9% (for the  $F_4$  score).

Table 4.3: Comparison of MR- $\beta$ -NMF with baseline  $\beta$ -NMF in terms of SNR on the activations and the dictionary vectors with respect to true factors on the dataset 2. The table reports the average, standard deviation and the best SNR over 100 random intializations for  $W$  and  $H$ . Bold numbers indicate the highest SNR.

Note	Activation SNR's (dB)				dictionary SNR's (dB)			
	$SNR_{H_k}$		$SNR_{H_{X,k}}$		$SNR_{W_k}$		$SNR_{W_{Y,k}}$	
	average $\pm$ std	best	average $\pm$ std	best	average $\pm$ std	best	average $\pm$ std	best
$A_4$	11.98 $\pm$ 0.01	12.03	12.17 $\pm$ 0.01	<b>12.17</b>	16.24 $\pm$ 0.02	<b>16.43</b>	16.29 $\pm$ 0.26	16.42
$C_5$	9.54 $\pm$ 0.02	<b>9.57</b>	9.43 $\pm$ 0.01	9.43	9.41 $\pm$ 0.02	<b>9.42</b>	8.61 $\pm$ 0.72	8.73
$D_4$	14.81 $\pm$ 0.01	14.82	14.92 $\pm$ 0.01	<b>14.92</b>	16.20 $\pm$ 0.06	<b>16.33</b>	15.24 $\pm$ 2.37	15.64
$F_4$	11.23 $\pm$ 0.01	11.32	11.52 $\pm$ 0.01	<b>11.54</b>	16.47 $\pm$ 0.05	16.50	16.76 $\pm$ 0.99	<b>16.93</b>

**Qualitative numerical results on real-life recording for data set 2** In this section we report qualitative numerical results obtained for a real-life recording of data set 2 used in [43]. The objective is to highlight the effect of increasing the penalty weight  $\lambda$  in MR- $\beta$ -NMF problem on the solutions  $W$  and  $H$  obtained with Algorithm 4. Two values for parameter  $\beta$  are considered;  $\beta = 1$  (KL) and  $\beta = 0$  (IS). For the analysis, the following parameters have been considered:

- values for penalty weight  $\lambda$  in the range  $[0.001; 2]$  are considered. For clarity purpose only results with  $\lambda = 0.001$  and  $\lambda = 1$  are presented in this section.

- initializations for  $W$  and  $H$  are performed with a continuation method; with  $\lambda$  set to zero, 50 analysis with MR- $\beta$ -NMF with random initialization for  $W$  and  $H$  have been conducted and best solutions (in terms of final value for the objective function) among them have been considered for initialization of Algorithm 4 for the different values of  $\lambda$ .
- we use MAXITERL1=100 and MAXITERL2=900.
- For  $\beta = 1$ , we have chosen the amplitude spectrograms as input data for Algorithm 4 with the factorization rank set to 5 as suggested in [43].
- For  $\beta = 0$ , we have chosen the power spectrograms as input data for Algorithm 4 with the factorization rank set to 6 as suggested in [43].

Figures 4.2 and 4.3 respectively show the activations and the basis obtained with MR-KL-NMF for both values of parameter  $\lambda$ . For more clarity, frequency range is limited to 5 kHz.

First of all, one can observe that the pitch estimation is equivalent to the ones presented in [43]. We get four pitches and one extra component that correspond to a mix of residual noise and transient events (as such, the hammer hits and pedal releases). Further, one observe that:

- as the value for  $\lambda$  increases, the frequency peaks estimation is more accurate and the frequency spread around the peaks decreases.
- increase the value for  $\lambda$  has no impact on the activations accuracy. For  $\lambda = 0.001$ , the factorization is mainly driven by input data  $V_X$  which ensures high-temporal resolution for the solution. When  $\lambda = 1$ , we get solutions with high frequency and temporal resolutions.

Let us finally report that the conclusions of the analysis conducted for  $\beta = 0$  are identical.

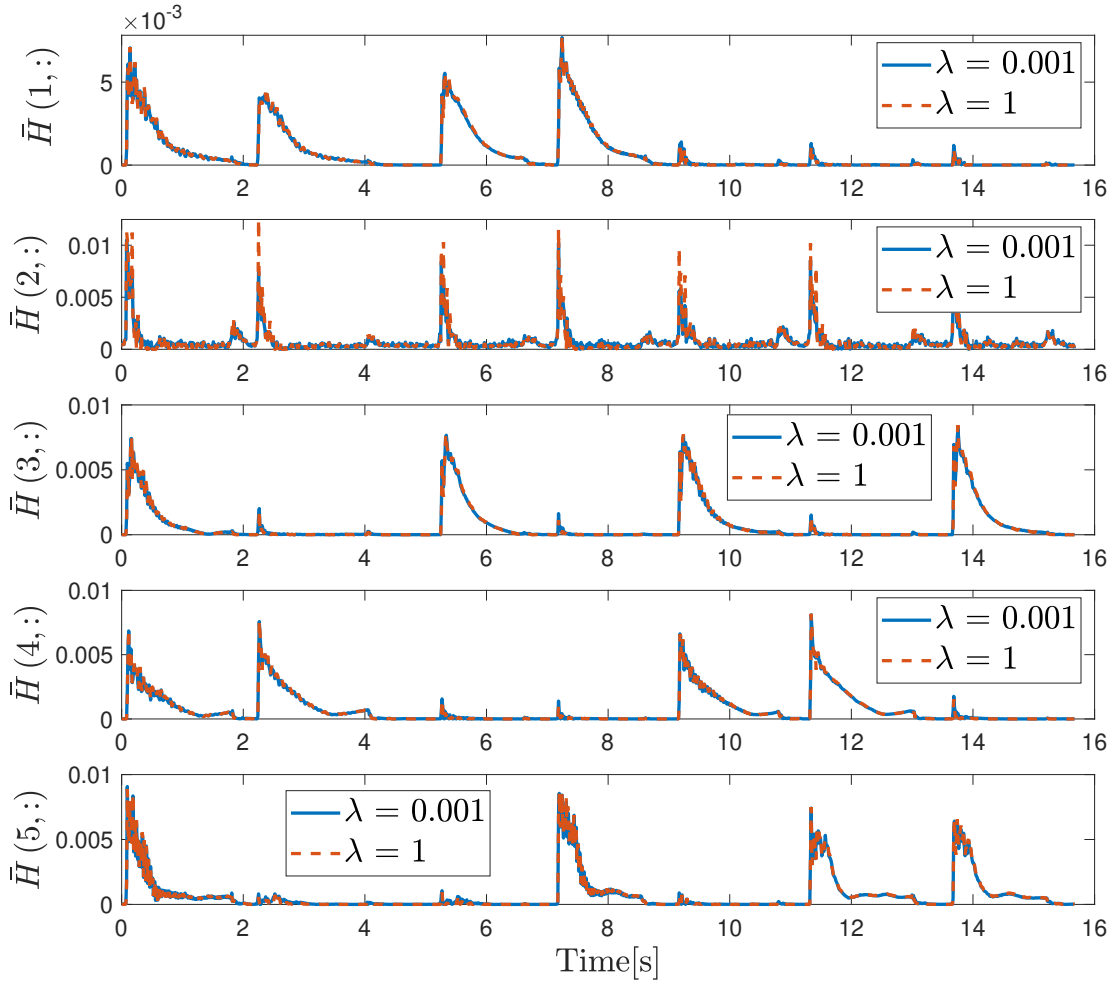
## 4.5 Numerical experiments on HS-MS fusion

In this section, we perform numerical experiments to validate the effectiveness of MR- $\beta$ -NMF on the HS-MS fusion problem.

### 4.5.1 Test setup and criteria

#### Test data

The proposed MR- $\beta$ -NMF algorithm is tested on semi-real datasets against several methods and algorithms widely used to tackle the HS-MS data fusion problem, namely GSA [6], CNMF [149], HySure [123], FUMI [142], GLP [5], MAPSMM [40], SFIM [96] and Lanaras's method [83]. In a nutshell: GSA, SFIM and GLP are pansharpening-based methods, the



**Fig. 4.2.** Rows of  $H$  obtained with MR-KL-NMF with  $K = 5$  for  $\lambda = 0.001$  and  $\lambda = 1$ .

remaining methods belong to subspace-based methods that can be splitted into unmixing methods (CNMF, Lanaras’s method and HySure) and Bayesian-based approaches (FUMI, MAPSMM) [147].

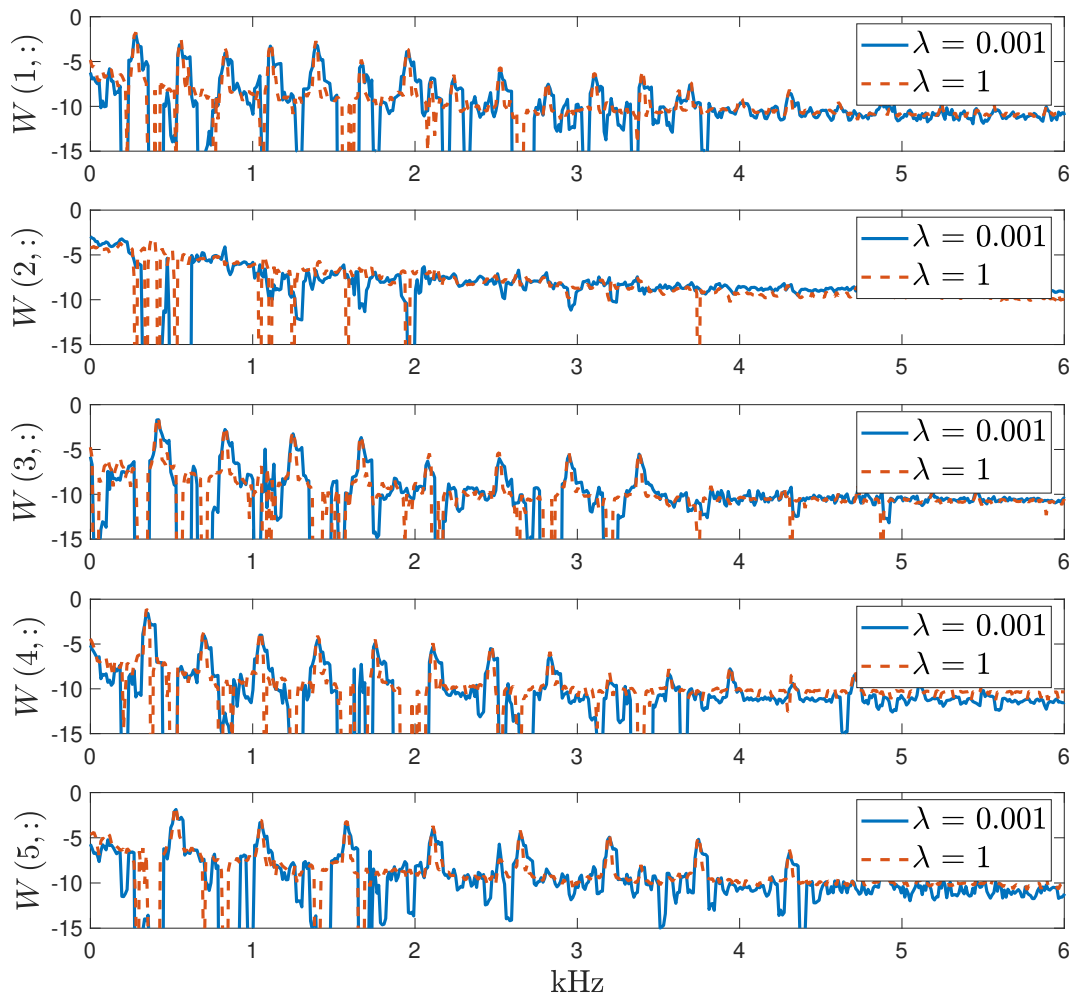
All the algorithms are implemented and tested on a desktop computer with Intel Core i7-8700@3.2GHz CPU, Geforce RTX 2070 Super GPU and 32GB memory. The codes<sup>1</sup> are written in MATLAB R2018a. The implementation for benchmarked algorithms comes from the comparative review of the recent literature for HS and MS data fusion detailed in [147]. We consider the following real HS datasets:

- HYDICE Urban: this data set has been acquired with HS Digital Imagery Collection Experiment (HYDICE) HS sensor [13] over an urban area at Copperas Cove, TX, U.S. in October 1995. The Urban dataset<sup>2</sup> consists of  $307 \times 307$  pixels and 162 spectral reflectance bands in the wavelength range 400 nm to 2500 nm. We extract a  $120 \times 120$

<sup>1</sup><https://naotoyokoya.com/Download.html>

<sup>2</sup><http://lesun.weebly.com/hyperspectral-data-set.html>





**Fig. 4.3.** Columns of  $W$  ( $\log_{10}$  scale) obtained with MR-KL-NMF with  $K = 5$  for  $\lambda = 0.001$  and  $\lambda = 1$ .

subimage from this dataset.

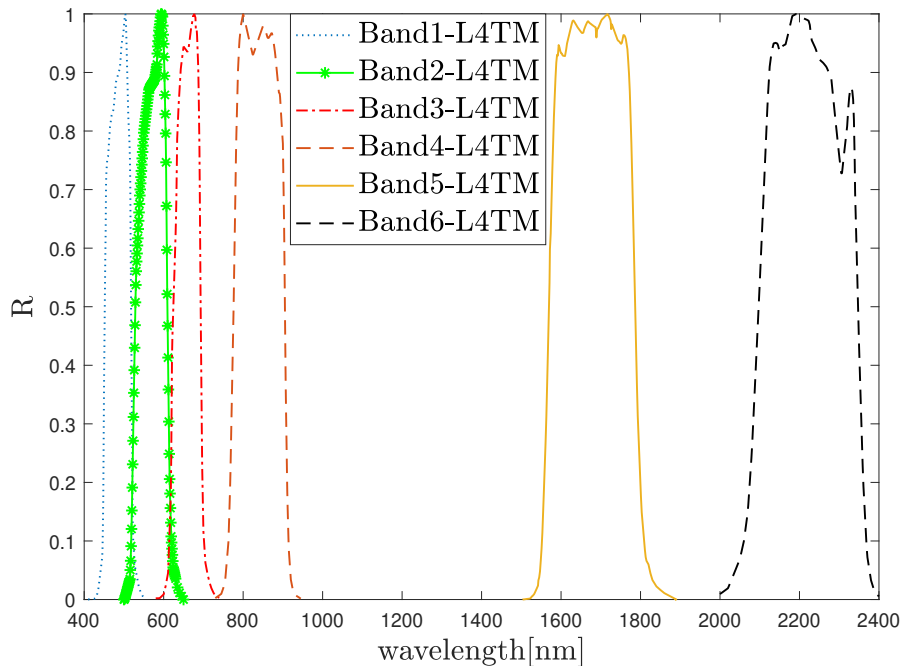
- HYDICE Washington DC Mall: this dataset<sup>3</sup> has been acquired with HYDICE HS sensor over the Washington DC Mall and consists of  $1208 \times 307$  pixels and 191 spectral reflectance bands in the wavelength range 400 nm to 2500 nm. We extract a  $240 \times 240$  subimage from this dataset.
- AVIRIS Indian Pines: this dataset has been acquired with NASA Airborne Visible/Infrared Imaging (AVIRIS) Spectrometer [134] over the Indian Pines test site in North-western Indiana and consists of  $145 \times 145$  pixels and 200 spectral reflectance bands in the wavelength range 400 nm to 2500 nm. We extract a  $120 \times 120$  subimage from this dataset.

<sup>3</sup><https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html>

Note that entries of the datasets are uncalibrated relative values, also referred as Digital Numbers (DN). As the goal is to fuse data and not to perform HS unmixing and classification, we do not convert these values into reflectances.

### Test procedure

In this chapter we consider semi-real data by conducting the numerical experiments based on the widely used Wald’s protocol [138]. This protocol consists in simulating input MS and HS images from a reference high-resolution HS image. In this chapter, MS image  $X$  and HS image  $Y$  have been derived from high-resolution HS image  $V$  through the models (4.4) and (4.5) respectively. Let us recall that the operator  $R$  from (4.1) designates the relative spectral responses from the super-resolution image to the MS image. In other words, it defines how the satellite instruments measure the intensity of the wavelengths (colors) of light. We generate a six-band MS image  $X$  by filtering the reference image  $V$  with the Landsat 4 TM-like reflectance spectral responses<sup>4</sup> depicted in Figure 4.4. The Landsat 4 TM sensor [105] has a spectral coverage from 400 nm to 2500 nm so that it is consistent with the spectral coverage of the datasets.



**Fig. 4.4.** Landsat 4 TM relative spectral responses.

The operator  $S$  (4.5) corresponds to the process of spatial blurring and downsampling. The high spectral low spatial resolution HS image  $Y$  is generated by applying a  $11 \times 11$  Gaussian spatial filter with a standard deviation of 1.7 on each band of the reference image  $V$  and downsampling every 4 pixels, both horizontally and vertically. The HS and MS images are finally both contaminated with noise. The level of noise is usually characterized

<sup>4</sup><https://landsat.usgs.gov/spectral-characteristics-viewer>

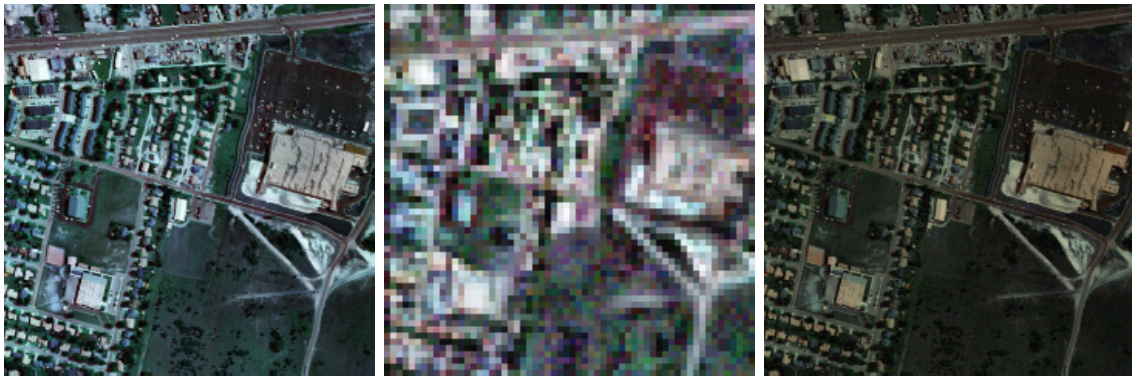
by the SNR expressed in dB. Here,  $\text{SNR}_X$  and  $\text{SNR}_Y$  refer to the noise level for the MS and HS images respectively. In this chapter, we apply the same level of noise for each spectral band. Let us give more insights on the last step of the MS image generation:  $X = \max(0, RV + \epsilon_X)$  where the noise matrix  $\epsilon_X$  is constructed as follows: we introduce  $x_i$  for  $i = 1, 2$ , some binary coefficients, and

$$\tilde{N} = x_1 \frac{N_P}{\|N_P\|_F} + x_2 \frac{N_F}{\|N_F\|_F},$$

where

- Each entry of  $N_P$  is generated using the Poisson distribution of parameter  $(R\tilde{V})_{i,j}$  for all  $(i, j)$ , where  $\tilde{V}$  is a noiseless low-rank approximation of  $V$  that is computed separately. More precisely, by setting  $\epsilon_X = 0_{F_X \times N_X}$  where  $0_{F_X \times N_X}$  is all-zero matrix, a solution  $(W, H)$  for MR- $\beta$ -NMF (4.6) is first computed with Algorithm 4, and the parameter for the Poisson distribution is defined as  $\tilde{V} = WH$ .
- Each entry of  $N_F$  is generated using the normal distribution of mean 0 and variance 1.

We set  $\epsilon_X = \eta \frac{\|RV\|_F}{\|\tilde{N}\|_F} \tilde{N}$  with  $\eta = \frac{1}{10^{\frac{\text{SNR}_X}{20}}}$ . For example, if we fix  $\text{SNR}_X = 25\text{dB}$ ,  $V_1 = \max(0, RV + \epsilon_X)$  is a MS image contaminated with 5.62% of noise (that is,  $\|\epsilon_X\|_F = 0.0562\|RV\|_F$ ) and projected onto the nonnegative orthant. The noise matrix  $\epsilon_Y$  is obtained in the same way. As an illustration, the reference image of Urban data set as the noise-contaminated HS and MS images are displayed in Figure 4.5.



**Fig. 4.5.** Urban data set: (Left) reference input image, (Middle) contaminated HS input image with  $\text{SNR}_X = 25\text{dB}$  and (Right) contaminated MS input image with  $\text{SNR}_Y = 25\text{dB}$ . All the images are in composite RGB color based on spectral bands 13 (460 nm), 33 (560 nm) and 39 (600 nm) for HS images and bands 1 to 3 for MS image.

The benchmarked algorithms listed in 4.5.1 are configured as recommended in the comparative review [147] with the following variations:

- The number of endmembers is a key parameter for unmixing-based methods. For MR- $\beta$ -NMF, CNMF, Lanaras's method and HySure,  $K$  is set to the 5 and 6 for HYDICE Urban and HYDICE Washington DC Mall datasets respectively as done in [154]. For the Indian Pine dataset,  $K = 16$  as in [128].
- The benchmarked algorithms are stopped when the relative change of the objective function is below  $10^{-4}$  or when the number of iterations exceeds 500. For algorithms such as CNMF that include outer and inner loops, we contacted the authors to set up the best balance for the maximum number of inner ( $I_1$ ) and outer ( $I_2$ ) loop iterations to fairly compare the methods, the following couples of values are considered:  $I_1 = 100$  and  $I_2 = 5$  and  $I_1 = 250$  and  $I_2 = 2$ . The couple of values that gives the best results for each dataset is considered in Section 4.5.2, that is  $I_1 = 100$  and  $I_2 = 5$ .
- The matrix  $R$  is known for all algorithms that make use of it. For MR- $\beta$ -NMF, it means we use MAXITERL1=500 and MAXITERL2=0.

Finally, let us summarize the initialization strategy:

- MR- $\beta$ -NMF uses random nonnegative initializations for  $W$  and  $H$ .
- CNMF starts by unmixing the HS image using VCA [106] to initialize the endmember signatures,
- SISAL [20] is used to initialize the endmembers for Lanaras's method.

Four variants of the MR- $\beta$ -NMF are considered, namely  $\beta = 2$ ,  $\beta = \frac{3}{2}$ ,  $\beta = 1$  and  $\beta = \frac{1}{2}$ . We test the algorithms under a scenario where no noise is added (that is,  $\tilde{N} = 0$ ), and a scenario where noise is added so that the SNRs for the noise terms in  $\epsilon_X$  and  $\epsilon_Y$  are  $SNR_X = 25dB$  and  $SNR_Y = 25dB$ .

### Performance evaluation

In order to assess the fusion quantitatively, we use the following five complementary and widely used quality measurements:

- Peak SNR (PSNR): the PSNR is used to assess the spatial reconstruction quality of each band. It corresponds to the ratio between the maximum power of a signal and the power of residual errors. The PSNR of the  $f$ -th band is defined as:

$$\text{PSNR}(v_f, \tilde{v}_f) = 10 \log_{10} \left( \frac{\max(v_f)^2}{\|v_f - \tilde{v}_f\|_2^2 / N_X} \right)$$

where  $N_X$  is the number of pixels,  $\max(v_f)$  is the maximum pixel value in the  $f$ -th reference band image. A larger PSNR value indicates a higher quality of spatial reconstruction. We use the average PSNR with respect to bands for the quality index of the entire fused image.

- The root-mean-square error (RMSE): RMSE is a similarity measure between the super-resolution image  $V$  and the fused image  $\tilde{V} = WH$  defined as:

$$\text{RMSE}(V, \tilde{V}) = \frac{1}{F_Y N_X} \|V - \tilde{V}\|_F^2$$

where  $F_Y$  is the number of spectral bands in the super-resolution image. The smaller the RMSE is, the better the fusion quality is.

- Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS): ERGAS provides a macroscopic statistical measure of the quality of the fused data. More precisely, ERGAS calculates the amount of spectral distortion in the image [139] which is defined as:

$$\text{ERGAS}(V, \tilde{V}) = 100d \sqrt{\frac{1}{F_Y} \sum_{f=1}^{F_Y} \frac{\|v_f - \tilde{v}_f\|_2^2}{(1/N_X e^T v_f)^2}}$$

where  $d$  is the spatial downsampling ratio between the higher and lower spatial resolution input images and  $e$  is a all-ones column vector of appropriate size. ERGAS calculates the band-wise normalized RMSE and multiplies it with  $d$  ratio to account for the difficulty in the fusion problem. The best value is at 0.

- Spectral Angle Mapper (SAM): SAM is used to quantify the spectral information preservation at each pixel. More precisely, SAM determines the spectral distance by computing the angle between two vectors of the estimated and reference spectra. The SAM index at the  $n$ -th pixel is defined as follows:

$$\text{SAM}(v_n, \tilde{v}_n) = \arccos\left(\frac{\langle v_n | \tilde{v}_n \rangle}{\|v_n\|_2 \|\tilde{v}_n\|_2}\right)$$

where  $\langle \cdot | \cdot \rangle$  stands for the inner product. The overall SAM is obtained by averaging the SAMs computed for all image pixels. The smaller the absolute value of SAM is, the better the fusion quality is.

- The universal image quality index (UIQI) introduced in [140]: UIQI evaluates the similarity between two single-band images. It is related to the correlation, luminance distortion, and contrast distortion of the estimated image w.r.t. reference image. The UIQI between two single-band images  $v_f$  and  $\tilde{v}_f$  is defined as follows:

$$\text{UIQI}(v_f, \tilde{v}_f) = \frac{4\sigma_{v_f, \tilde{v}_f}^2 \mu_{v_f} \mu_{\tilde{v}_f}}{(\sigma_{v_f}^2 + \sigma_{\tilde{v}_f}^2) (\mu_{v_f}^2 + \mu_{\tilde{v}_f}^2)}$$

where  $(\mu_{v_f}, \mu_{\tilde{v}_f}, \sigma_{v_f}^2, \sigma_{\tilde{v}_f}^2)$  are the sample means and variances of  $v_f$  and  $\tilde{v}_f$  and  $\sigma_{v_f, \tilde{v}_f}^2$  is the sample covariance of  $(v_f, \tilde{v}_f)$ . UIQI indicator is in the range  $[-1, 1]$ . For multiband images, the overall UIQI is computed by averaging the UIQI computed band by band. The best value for UIQI is at 1.

For more details about these quality measurements, we refer the reader to [97] and [142].

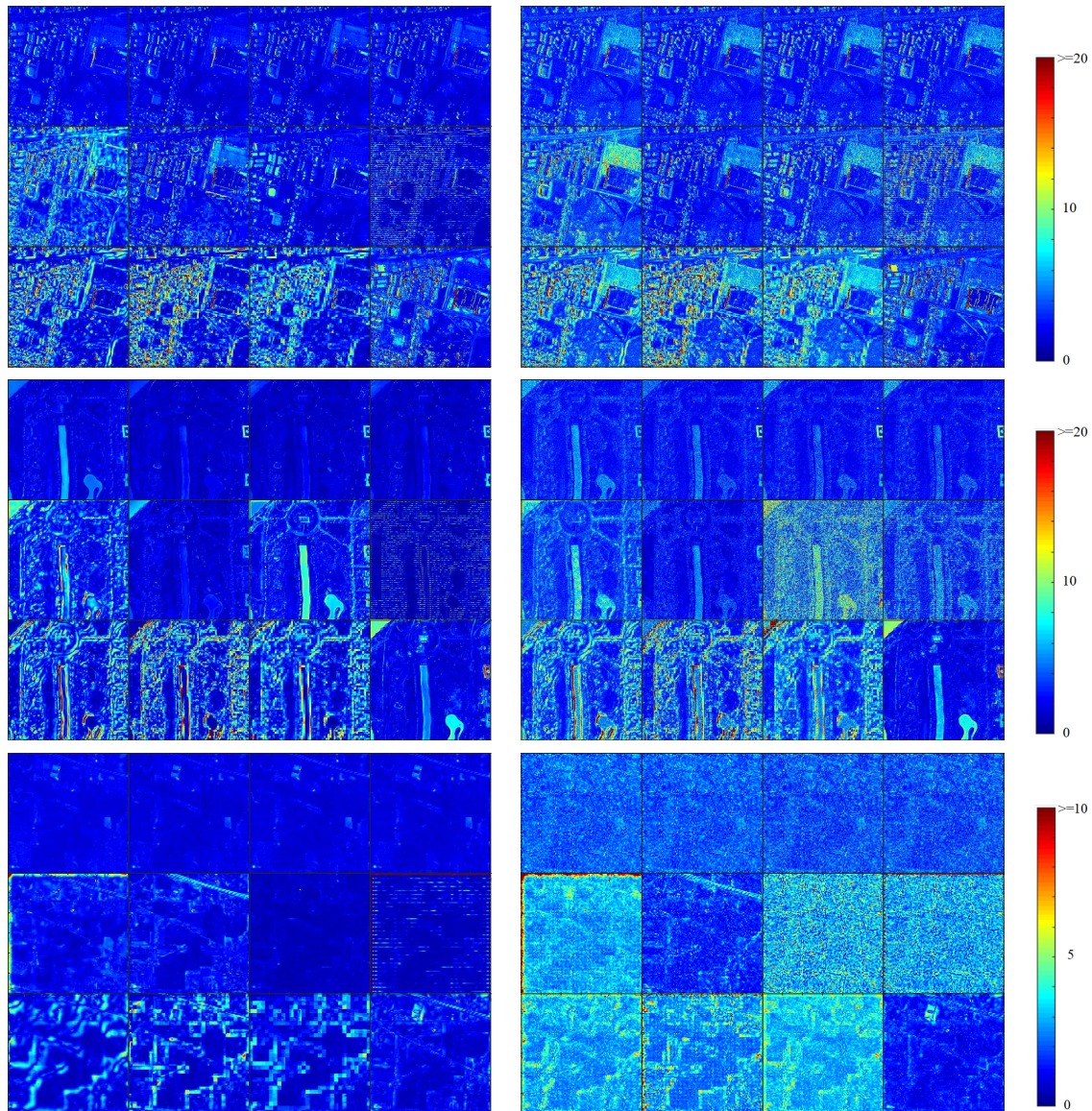
### 4.5.2 Experimental results

We ran 20 independent trials for each dataset detailed in 4.5.1. The average performance of each algorithm is shown in Tables 4.4 to 4.6.

Except for runtimes, MR- $\beta$ -NMF generally rank in the fifth first for all the quality measurements. For Urban dataset with noise added, MR- $\beta$ -NMF with  $\beta = 1$ ,  $\beta = 1/2$  and  $\beta = 3/2$  respectively rank first, second and third for all the metrics except for SAM for which CNMF ranks first. For the condition with no noise added, MR- $\beta$ -NMF with  $\beta = 1$ ,  $\beta = 1/2$  ranks first and second for all metrics. MR- $\beta$ -NMF with  $\beta = 3/2$ , FUMI and HySure give similar results. For Washington DC Mall without noise added, MR- $\beta$ -NMF with  $\beta = 1$ ,  $\beta = 1/2$  ranks first and second for all metrics. For Indian Pines dataset without noise added, MR- $\beta$ -NMF with  $\beta = 1$  ranks second while HySure ranks first. When noise is added, Lanaras’s method ranks first while MR- $\beta$ -NMF with  $\beta = 1/2$ ,  $\beta = 1$  rank second and third for most criteria. In order to give more insights on the performance comparison between algorithms, Figure 4.6 displays the SAM maps obtained for one trial for the Urban, Washington DC Mall and Indian Pines datasets. Visually, the proposed method performs competitively with other state-of-the-art methods. Indeed, as already observed with the SAM comparison in Tables 4.4 to 4.6, the variants of MR- $\beta$ -NMF show in general lower values for SAM errors across the images. For the Urban dataset, the highest SAM errors obtained with the variants of MR- $\beta$ -NMF are less widespread and localized at some specific spots which correspond to the edges of the roofs and trees. This observation makes sense as those regions show more atypic reflectance angles and therefore more non-linear effects in terms of spectral mixture. The same observations apply for the Washington DC Mall dataset with and without noise added. For the Indian Pines dataset without noise added, HySure and FUMI algorithms show lower SAM errors accross images, we visually confirm that MR- $\beta$ -NMF with  $\beta = 1, 1/2, 3/2$  rank third to fifth. When the noise is added, Lanaras’s method gives the lowest SAM errors and is less widespread, while MR- $\beta$ -NMF with  $\beta = 1, 1/2, 3/2$  appear to provide less accurate estimates than CNMF that visually looks better.

## 4.6 Conclusions and outlooks

In this chapter, we have presented a new NMF approach for blind spectral unmixing, called multi-resolution  $\beta$ -NMF (MR- $\beta$ -NMF). The estimation relies on the minimization of the  $\beta$ -divergence, a flexible family of measures of fit. MR- $\beta$ -NMF addresses the resolution trade-off between two adversarial dimensions by fusing the information coming from multiple data with different resolutions in order to produce a factorization with high resolutions for all the dimensions. We have provided multiplicative updates to tackle the minimization problem and we showed that MR- $\beta$ -NMF is flexible and can be successfully applied to



**Fig. 4.6.** SAM maps for the different hyperspectral images. From top to bottom: Urban dataset with  $K = 5$ , Washington DC Mall dataset with  $K = 6$ , and Indian Pines dataset with  $K = 16$ . On the left column: SAM maps without added noise. On the right column: SAM maps with added noise ( $SNR_X = SNR_Y = 25dB$ ). For each image, the 12 SAM maps correspond to the different benchmark algorithms; from left to right, top to bottom: MR-2-NMF, MR-3/2-NMF, MR-1-NMF, MR-1/2-NMF, GSA, CNMF, HySure, FUMI, GLP, MAPSMM, SFIM, and Lanaras's method.

Table 4.4: Comparison of MR- $\beta$ -NMF with state-of-the-arts methods for HS-MS fusion problem on dataset HYDICE Urban. The table reports the average, standard deviation for the quantitative quality assessments over 20 trials. Bold, underlined and italic to highlight the three best algorithms.

Method	Runtime (seconds)	PSNR (dB)	RMSE	ERGAS	SAM	UIQI
Best value	0	$\infty$	0	0	0	1
Dataset - HYDICE Urban - SNR = 25dB						
MR- $\beta = 2$ -NMF	52.25 $\pm$ 2.45	33.88 $\pm$ 0.10	16.26 $\pm$ 0.19	2.48 $\pm$ 0.03	4.13 $\pm$ 0.06	0.97 $\pm$ 0.00
MR- $\beta = 3/2$ -NMF	54.46 $\pm$ 2.31	<i>34.54 <math>\pm</math> 0.06</i>	<i>14.92 <math>\pm</math> 0.09</i>	<i>2.28 <math>\pm</math> 0.01</i>	3.65 $\pm$ 0.04	<i>0.98 <math>\pm</math> 0.00</i>
MR- $\beta = 1$ -NMF	52.20 $\pm$ 2.03	<b>34.85 <math>\pm</math> 0.10</b>	<b>14.51 <math>\pm</math> 0.14</b>	<b>2.22 <math>\pm</math> 0.03</b>	<b>3.49 <math>\pm</math> 0.06</b>	<b>0.98 <math>\pm</math> 0.00</b>
MR- $\beta = 1/2$ -NMF	54.47 $\pm$ 1.96	<u>34.81 <math>\pm</math> 0.10</u>	<u>14.65 <math>\pm</math> 0.15</u>	<u>2.24 <math>\pm</math> 0.02</u>	<u>3.52 <math>\pm</math> 0.06</u>	<u>0.98 <math>\pm</math> 0.00</u>
GSA	<i>0.72 <math>\pm</math> 0.05</i>	32.52 $\pm$ 0.00	19.41 $\pm$ 0.00	2.87 $\pm$ 0.00	5.63 $\pm$ 0.00	0.96 $\pm$ 0.00
CNMF	9.73 $\pm$ 1.84	34.33 $\pm$ 0.50	15.45 $\pm$ 0.85	2.37 $\pm$ 0.17	<i>3.64 <math>\pm</math> 0.27</i>	0.98 $\pm$ 0.00
HySure	31.57 $\pm$ 2.93	33.90 $\pm$ 0.00	16.44 $\pm$ 0.00	2.57 $\pm$ 0.00	4.17 $\pm$ 0.00	0.97 $\pm$ 0.00
FUMI	<u>0.39 <math>\pm</math> 0.03</u>	32.92 $\pm$ 0.00	20.30 $\pm$ 0.00	2.85 $\pm$ 0.00	4.92 $\pm$ 0.00	0.96 $\pm$ 0.00
GLP	6.05 $\pm$ 0.42	27.24 $\pm$ 0.00	34.37 $\pm$ 0.00	5.10 $\pm$ 0.00	6.27 $\pm$ 0.00	0.91 $\pm$ 0.00
MAPSMM	44.12 $\pm$ 2.60	25.57 $\pm$ 0.00	41.95 $\pm$ 0.00	6.15 $\pm$ 0.00	6.82 $\pm$ 0.00	0.87 $\pm$ 0.00
SFIM	<b>0.24 <math>\pm</math> 0.03</b>	26.32 $\pm$ 0.00	37.89 $\pm$ 0.00	5.71 $\pm$ 0.00	5.90 $\pm$ 0.00	0.90 $\pm$ 0.00
Lanaras's method	8.12 $\pm$ 8.71	29.33 $\pm$ 0.29	26.84 $\pm$ 0.85	4.39 $\pm$ 0.23	4.88 $\pm$ 0.26	0.94 $\pm$ 0.00
Dataset - HYDICE Urban - No added noise						
MR- $\beta = 2$ -NMF	49.55 $\pm$ 0.31	38.10 $\pm$ 0.40	10.94 $\pm$ 0.31	1.67 $\pm$ 0.07	3.28 $\pm$ 0.10	0.99 $\pm$ 0.00
MR- $\beta = 3/2$ -NMF	51.54 $\pm$ 0.52	40.01 $\pm$ 0.50	<i>8.82 <math>\pm</math> 0.32</i>	<i>1.35 <math>\pm</math> 0.09</i>	2.60 $\pm$ 0.10	<i>0.99 <math>\pm</math> 0.00</i>
MR- $\beta = 1$ -NMF	49.71 $\pm$ 0.12	<u>41.53 <math>\pm</math> 0.56</u>	<u>7.86 <math>\pm</math> 0.28</u>	<b>1.19 <math>\pm</math> 0.07</b>	<u>2.27 <math>\pm</math> 0.10</u>	<b>0.99 <math>\pm</math> 0.00</b>
MR- $\beta = 1/2$ -NMF	52.09 $\pm$ 0.35	<b>41.69 <math>\pm</math> 0.64</b>	<b>7.81 <math>\pm</math> 0.35</b>	<u>1.19 <math>\pm</math> 0.08</u>	<b>2.23 <math>\pm</math> 0.12</b>	<u>0.99 <math>\pm</math> 0.00</u>
GSA	<i>0.67 <math>\pm</math> 0.04</i>	32.93 $\pm$ 0.00	22.17 $\pm$ 0.00	2.87 $\pm$ 0.00	5.25 $\pm$ 0.00	0.97 $\pm$ 0.00
CNMF	10.56 $\pm$ 2.02	35.35 $\pm$ 0.64	13.91 $\pm$ 1.81	2.18 $\pm$ 0.32	3.26 $\pm$ 0.53	0.98 $\pm$ 0.00
HySure	28.51 $\pm$ 1.09	40.27 $\pm$ 0.00	9.67 $\pm$ 0.00	1.46 $\pm$ 0.00	<i>2.50 <math>\pm</math> 0.00</i>	0.99 $\pm$ 0.00
FUMI	<u>0.36 <math>\pm</math> 0.02</u>	<i>41.01 <math>\pm</math> 0.00</i>	14.14 $\pm$ 0.00	1.67 $\pm$ 0.00	2.71 $\pm$ 0.00	0.99 $\pm$ 0.00
GLP	5.61 $\pm$ 0.09	27.97 $\pm$ 0.00	31.97 $\pm$ 0.00	4.65 $\pm$ 0.00	4.78 $\pm$ 0.00	0.94 $\pm$ 0.00
MAPSMM	42.19 $\pm$ 0.84	25.92 $\pm$ 0.00	40.56 $\pm$ 0.00	5.89 $\pm$ 0.00	5.66 $\pm$ 0.00	0.89 $\pm$ 0.00
SFIM	<b>0.21 <math>\pm</math> 0.03</b>	27.05 $\pm$ 0.00	35.19 $\pm$ 0.00	5.21 $\pm$ 0.00	4.21 $\pm$ 0.00	0.93 $\pm$ 0.00
Lanaras's method	4.72 $\pm$ 4.72	29.50 $\pm$ 0.35	26.54 $\pm$ 0.69	4.26 $\pm$ 0.23	4.57 $\pm$ 0.21	0.95 $\pm$ 0.00



Table 4.5: Comparison of MR- $\beta$ -NMF with state-of-the-arts methods for HS-MS fusion problem on dataset HYDICE Washington DC Mall. The table reports the average, standard deviation for the quantitative quality assessments over 20 trials. Bold, underlined and italic to highlight the three best algorithms.

Method	Runtime (seconds)	PSNR (dB)	RMSE	ERGAS	SAM	UIQI
Best value	0	$\infty$	0	0	0	1
Dataset - HYDICE Washington DC Mall - $SNR = 25dB$						
MR- $\beta = 2$ -NMF	57.59 $\pm$ 0.32	<u>26.77 <math>\pm</math> 0.25</u>	202.02 $\pm$ 3.59	18.21 $\pm$ 0.13	3.38 $\pm$ 0.11	<b>0.90 <math>\pm</math> 0.01</b>
MR- $\beta = 3/2$ -NMF	60.04 $\pm$ 0.39	<i>26.37 <math>\pm</math> 0.32</i>	<i>194.40 <math>\pm</math> 6.38</i>	<i>18.07 <math>\pm</math> 0.23</i>	3.05 $\pm$ 0.18	0.87 $\pm$ 0.01
MR- $\beta = 1$ -NMF	57.95 $\pm$ 0.24	26.29 $\pm$ 0.20	<b>188.42 <math>\pm</math> 11.18</b>	18.50 $\pm$ 0.25	<i>2.83 <math>\pm</math> 0.28</i>	0.86 $\pm$ 0.01
MR- $\beta = 1/2$ -NMF	60.38 $\pm$ 0.20	25.68 $\pm$ 0.28	201.62 $\pm$ 14.05	19.46 $\pm$ 0.41	3.06 $\pm$ 0.30	0.83 $\pm$ 0.01
GSA	<i>0.79 <math>\pm</math> 0.04</i>	23.00 $\pm$ 0.00	235.64 $\pm$ 0.00	32.25 $\pm$ 0.00	4.20 $\pm$ 0.00	0.74 $\pm$ 0.00
CNMF	7.25 $\pm$ 1.26	<b>27.60 <math>\pm</math> 0.09</b>	<u>192.67 <math>\pm</math> 6.50</u>	<u>17.37 <math>\pm</math> 0.10</u>	<b>2.55 <math>\pm</math> 0.14</b>	<i>0.89 <math>\pm</math> 0.00</i>
HySure	34.14 $\pm$ 0.94	24.01 $\pm$ 0.00	351.13 $\pm$ 0.00	33.51 $\pm$ 0.00	6.15 $\pm$ 0.00	0.75 $\pm$ 0.00
FUMI	<u>0.42 <math>\pm</math> 0.02</u>	24.67 $\pm$ 0.00	243.06 $\pm$ 0.00	19.73 $\pm$ 0.00	4.04 $\pm$ 0.00	0.80 $\pm$ 0.00
GLP	6.42 $\pm$ 0.24	19.85 $\pm$ 0.00	423.89 $\pm$ 0.00	33.64 $\pm$ 0.00	5.28 $\pm$ 0.00	0.67 $\pm$ 0.00
MAPSMM	40.91 $\pm$ 0.46	19.34 $\pm$ 0.00	494.39 $\pm$ 0.00	32.18 $\pm$ 0.00	5.91 $\pm$ 0.00	0.65 $\pm$ 0.00
SFIM	<b>0.24 <math>\pm</math> 0.01</b>	18.08 $\pm$ 0.00	892.35 $\pm$ 0.00	42.23 $\pm$ 0.00	5.45 $\pm$ 0.00	0.64 $\pm$ 0.00
Lanaras's method	3.11 $\pm$ 1.94	25.95 $\pm$ 0.06	235.62 $\pm$ 2.67	<b>17.36 <math>\pm</math> 0.02</b>	<u>2.78 <math>\pm</math> 0.03</u>	<u>0.90 <math>\pm</math> 0.00</u>
Dataset - HYDICE Washington DC Mall - No added noise						
MR- $\beta = 2$ -NMF	58.55 $\pm$ 1.50	32.61 $\pm$ 0.28	128.50 $\pm$ 5.87	5.54 $\pm$ 0.13	2.59 $\pm$ 0.12	0.97 $\pm$ 0.00
MR- $\beta = 3/2$ -NMF	60.95 $\pm$ 1.58	35.36 $\pm$ 0.38	<i>104.11 <math>\pm</math> 5.89</i>	2.41 $\pm$ 0.22	1.89 $\pm$ 0.12	<i>0.98 <math>\pm</math> 0.00</i>
MR- $\beta = 1$ -NMF	59.01 $\pm$ 2.02	<u>37.80 <math>\pm</math> 0.75</u>	<b>89.20 <math>\pm</math> 5.43</b>	<u>1.76 <math>\pm</math> 0.27</u>	<b>1.47 <math>\pm</math> 0.07</b>	<b>0.99 <math>\pm</math> 0.00</b>
MR- $\beta = 1/2$ -NMF	61.21 $\pm$ 1.05	<b>38.27 <math>\pm</math> 0.83</b>	<u>90.88 <math>\pm</math> 6.26</u>	<b>1.55 <math>\pm</math> 0.20</b>	<u>1.48 <math>\pm</math> 0.10</u>	<u>0.99 <math>\pm</math> 0.00</u>
GSA	<i>0.81 <math>\pm</math> 0.08</i>	29.93 $\pm$ 0.00	262.27 $\pm$ 0.00	3.11 $\pm$ 0.00	3.84 $\pm$ 0.00	0.97 $\pm$ 0.00
CNMF	7.90 $\pm$ 2.67	31.46 $\pm$ 1.07	152.95 $\pm$ 14.25	5.93 $\pm$ 8.92	2.01 $\pm$ 0.49	0.96 $\pm$ 0.03
HySure	35.85 $\pm$ 2.19	31.23 $\pm$ 0.00	190.57 $\pm$ 0.10	3.21 $\pm$ 0.00	3.21 $\pm$ 0.00	0.96 $\pm$ 0.00
FUMI	<u>0.43 <math>\pm</math> 0.03</u>	<i>36.52 <math>\pm</math> 0.00</i>	142.92 $\pm$ 0.00	<i>2.32 <math>\pm</math> 0.00</i>	<i>1.76 <math>\pm</math> 0.00</i>	0.98 $\pm$ 0.00
GLP	6.95 $\pm$ 0.52	26.19 $\pm$ 0.00	373.07 $\pm$ 0.00	4.53 $\pm$ 0.00	4.16 $\pm$ 0.00	0.93 $\pm$ 0.00
MAPSMM	42.88 $\pm$ 0.85	24.42 $\pm$ 0.00	459.09 $\pm$ 0.00	5.61 $\pm$ 0.00	4.98 $\pm$ 0.00	0.88 $\pm$ 0.00
SFIM	<b>0.27 <math>\pm</math> 0.05</b>	25.12 $\pm$ 0.00	408.40 $\pm$ 0.00	6.53 $\pm$ 0.00	3.95 $\pm$ 0.00	0.92 $\pm$ 0.00
Lanaras's method	4.70 $\pm$ 3.55	28.46 $\pm$ 0.36	230.31 $\pm$ 7.44	3.94 $\pm$ 0.21	2.55 $\pm$ 0.03	0.96 $\pm$ 0.00

Table 4.6: Comparison of MR- $\beta$ -NMF with state-of-the-arts methods for HS-MS fusion problem on dataset AVIRIS Indian Pines. The table reports the average, standard deviation for the quantitative quality assessments over 20 trials. Bold, underlined and italic to highlight the three best algorithms.

Method	Runtime (seconds)	PSNR (dB)	RMSE	ERGAS	SAM	UIQI
Best value	0	$\infty$	0	0	0	1
Dataset - AVIRIS Indian Pines - $SNR = 25dB$						
MR- $\beta = 2$ -NMF	15.48 $\pm$ 0.53	27.11 $\pm$ 0.03	187.37 $\pm$ 0.80	1.64 $\pm$ 0.01	2.26 $\pm$ 0.02	0.78 $\pm$ 0.00
MR- $\beta = 3/2$ -NMF	16.76 $\pm$ 0.75	27.29 $\pm$ 0.02	183.47 $\pm$ 0.56	1.57 $\pm$ 0.00	2.14 $\pm$ 0.01	<i>0.78 <math>\pm</math> 0.00</i>
MR- $\beta = 1$ -NMF	15.57 $\pm$ 0.53	<i>27.38 <math>\pm</math> 0.02</i>	<i>181.77 <math>\pm</math> 0.51</i>	<i>1.55 <math>\pm</math> 0.00</i>	2.09 $\pm$ 0.01	<u>0.78 <math>\pm</math> 0.00</u>
MR- $\beta = 1/2$ -NMF	16.90 $\pm$ 0.55	<u>27.55 <math>\pm</math> 0.03</u>	<u>179.10 <math>\pm</math> 0.41</u>	<u>1.52 <math>\pm</math> 0.01</u>	<i>2.03 <math>\pm</math> 0.01</i>	<b>0.79 <math>\pm</math> 0.00</b>
GSA	<i>0.31 <math>\pm</math> 0.04</i>	21.79 $\pm$ 0.00	326.23 $\pm$ 0.00	2.94 $\pm$ 0.00	3.28 $\pm$ 0.00	0.64 $\pm$ 0.00
CNMF	2.13 $\pm$ 0.10	24.05 $\pm$ 0.21	241.72 $\pm$ 5.39	2.33 $\pm$ 0.07	<u>1.68 <math>\pm</math> 0.04</u>	0.60 $\pm$ 0.01
HySure	22.70 $\pm$ 0.43	24.82 $\pm$ 0.28	241.17 $\pm$ 3.31	2.33 $\pm$ 0.13	3.25 $\pm$ 0.05	0.64 $\pm$ 0.01
FUMI	<b>0.12 <math>\pm</math> 0.02</b>	24.71 $\pm$ 0.00	242.25 $\pm$ 0.00	2.27 $\pm$ 0.00	3.19 $\pm$ 0.00	0.66 $\pm$ 0.00
GLP	2.36 $\pm$ 0.07	20.24 $\pm$ 0.00	403.70 $\pm$ 0.00	3.47 $\pm$ 0.00	3.14 $\pm$ 0.00	0.49 $\pm$ 0.00
MAPSMM	10.63 $\pm$ 0.21	18.35 $\pm$ 0.00	519.28 $\pm$ 0.00	4.30 $\pm$ 0.00	3.36 $\pm$ 0.00	0.42 $\pm$ 0.00
SFIM	<u>0.20 <math>\pm</math> 0.02</u>	19.74 $\pm$ 0.00	423.46 $\pm$ 0.00	3.68 $\pm$ 0.00	3.31 $\pm$ 0.00	0.48 $\pm$ 0.00
Lanaras's method	2.82 $\pm$ 1.69	<b>29.59 <math>\pm</math> 0.71</b>	<b>149.59 <math>\pm</math> 13.20</b>	<b>1.19 <math>\pm</math> 0.09</b>	<b>1.43 <math>\pm</math> 0.06</b>	0.76 $\pm$ 0.05
Dataset - AVIRIS Indian Pines - No added noise						
MR- $\beta = 2$ -NMF	14.55 $\pm$ 0.07	36.43 $\pm$ 0.15	69.71 $\pm$ 1.65	0.65 $\pm$ 0.02	1.23 $\pm$ 0.03	0.92 $\pm$ 0.00
MR- $\beta = 3/2$ -NMF	15.69 $\pm$ 0.09	38.09 $\pm$ 0.09	57.69 $\pm$ 0.89	0.48 $\pm$ 0.00	1.00 $\pm$ 0.02	0.93 $\pm$ 0.00
MR- $\beta = 1$ -NMF	14.56 $\pm$ 0.03	<u>39.30 <math>\pm</math> 0.13</u>	<u>51.66 <math>\pm</math> 0.79</u>	<u>0.41 <math>\pm</math> 0.01</u>	<u>0.90 <math>\pm</math> 0.01</u>	<i>0.94 <math>\pm</math> 0.00</i>
MR- $\beta = 1/2$ -NMF	16.00 $\pm$ 0.05	<i>39.15 <math>\pm</math> 0.20</i>	<i>52.98 <math>\pm</math> 1.18</i>	<i>0.42 <math>\pm</math> 0.01</i>	0.91 $\pm$ 0.02	0.94 $\pm$ 0.00
GSA	<i>0.29 <math>\pm</math> 0.03</i>	23.33 $\pm$ 0.00	300.32 $\pm$ 0.00	2.42 $\pm$ 0.00	1.38 $\pm$ 0.00	0.90 $\pm$ 0.00
CNMF	1.94 $\pm$ 0.09	26.72 $\pm$ 0.16	184.42 $\pm$ 2.95	1.71 $\pm$ 0.04	1.17 $\pm$ 0.03	0.74 $\pm$ 0.01
HySure	20.83 $\pm$ 0.17	<b>40.96 <math>\pm</math> 0.03</b>	<b>44.29 <math>\pm</math> 0.18</b>	<b>0.34 <math>\pm</math> 0.00</b>	<b>0.56 <math>\pm</math> 0.00</b>	<b>0.96 <math>\pm</math> 0.00</b>
FUMI	<b>0.11 <math>\pm</math> 0.02</b>	39.13 $\pm$ 0.00	115.58 $\pm$ 0.00	0.83 $\pm$ 0.00	<i>0.90 <math>\pm</math> 0.00</i>	<u>0.95 <math>\pm</math> 0.00</u>
GLP	2.24 $\pm$ 0.05	23.12 $\pm$ 0.00	312.46 $\pm$ 0.00	2.48 $\pm$ 0.00	1.42 $\pm$ 0.00	0.85 $\pm$ 0.00
MAPSMM	10.09 $\pm$ 0.14	22.27 $\pm$ 0.00	346.40 $\pm$ 0.00	2.74 $\pm$ 0.00	1.54 $\pm$ 0.00	0.78 $\pm$ 0.00
SFIM	<u>0.18 <math>\pm</math> 0.01</u>	22.66 $\pm$ 0.00	328.92 $\pm$ 0.00	2.62 $\pm$ 0.00	1.39 $\pm$ 0.00	0.85 $\pm$ 0.00
Lanaras's method	2.05 $\pm$ 1.90	29.89 $\pm$ 0.54	155.03 $\pm$ 7.39	1.15 $\pm$ 0.06	1.18 $\pm$ 0.02	0.81 $\pm$ 0.00

various problems. In particular, we have showcased its efficiency on two instrumental examples. The first is the audio spectral unmixing for which the frequency-by-time data matrix is computed with the short-time Fourier transform and is the result of a trade-off between the frequency resolution and the temporal resolution. We highlighted the capacity of this approach to provide solutions that show high frequency and high temporal accuracy taking advantage from the input data. Based on these results, MR- $\beta$ -NMF seems to be well suited for audio applications such as transcription problems and performs in general better than baseline NMF methods. The second is BHU for which the wavelength-by-location data matrix is a trade-off between the number of wavelengths measured and the spatial resolution. We demonstrated the efficiency of MR- $\beta$ -NMF to tackle the HS-MS data fusion problem. Based on various quantitative quality assessments, the proposed method performs competitively with the state of the art.

Further work includes:

- The theoretical guarantees for recoverability or identifiability of the latent factors for the model and the associated problem.
- The design of more efficient algorithms to reduce the runtime.
- The design of more accurate downsampling operators in the frame of audio spectral unmixing.

## 5 Multiplicative Updates for NMF with $\beta$ -divergences Under Disjoint Equality Constraints

In this chapter, we introduce a general framework to design multiplicative updates (MU) for NMF based on  $\beta$ -divergences ( $\beta$ -NMF) with disjoint equality constraints, and with penalty terms in the objective function. By disjoint, we mean that each variable appears in at most one equality constraint. Our MU satisfy the set of constraints after each update of the variables during the optimization process, while guaranteeing that the objective function decreases monotonically. We showcase this framework on three NMF optimization problems, and show that it outperforms the state of the art: (1)  $\beta$ -NMF with sum-to-one constraints on the columns of  $H$ , (2) minimum-volume  $\beta$ -NMF with sum-to-one constraints on the columns of  $W$ , and (3) sparse  $\beta$ -NMF with  $\ell_2$ -norm constraints on the columns of  $W$ .

The content of this chapter is extracted from [92]: V. Leplat, N. Gillis, and J. Idier. *Multiplicative Updates for NMF with  $\beta$ -Divergences under Disjoint Equality Constraints*. 2020. arXiv: 2010.16223

### 5.1 Introduction

As introduced in sections 1.4 and 1.9, the most standard approach to compute  $W$  and  $H$  for the standard approximate NMF model  $V \approx WH$  (1.2) is to solve the following optimization problem

$$\min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} D(V|WH) \quad \text{such that} \quad H \geq 0 \text{ and } W \geq 0, \quad (5.1)$$

where  $D(V|WH) = \sum_{f,n} d(V_{fn}|[WH]_{fn})$ ,  $d(x|y)$  is a measure of distance between two scalars. In this chapter, we consider the  $\beta$ -divergences  $d_\beta(x|y)$ , see Section 1.9.1 for more details.

In Section 1.10 we explained that most NMF algorithms developed to tackle (5.1) are based on iterative schemes that alternatively updates the factors  $W$  and  $H$ . At each iteration, the minimization over one factor,  $W$  or  $H$ , is performed with various optimization methods. For  $\beta$ -divergences, the most popular approach is to use multiplicative updates (MU) which were introduced in the seminal papers of Lee and Seung [87, 85]. In all application we are aware of,  $\beta$  is always chosen smaller than two. The reason is that,

for  $\beta > 2$ ,  $\beta$ -divergences become more and more sensitive to outliers. Already for  $\beta = 2$ , it is well-known that the squared Frobenius norm is sensitive to outliers. However, the case  $\beta = 2$  is particular because the subproblem in  $W$  and  $H$  are nonnegative least squares problems, that is, convex quadratic problems with Lipschitz continuous gradient. Therefore, highly efficient schemes exist when  $\beta = 2$  that outperform the MU; for example exact block coordinate descent methods [29, 60, 79], or fast gradient methods [69]. In this chapter, we focus on the case  $\beta < 2$ .

In many applications, on top of the nonnegative constraints on the variables, additional constraints are needed to provide a meaningful solution. An instrumental example is the constraint that the entries in each column of  $H$  sum to one; this is the so-called sum-to-one constraint that is crucial in BHU; see Section 5.3 for more details. Another example from Chapter 3 is the sum-to-one constraint on the columns of  $W$  along with a volume regularizer on  $W$ . This model and associated optimization problem (3.2) leads to identifiability of the factors  $W$  and  $H$  under mild conditions; see Section 5.4 for more details. Most algorithms that deal with such equality constraints do it a posteriori with a projection onto the feasible set, or with a renormalization of the columns of  $W$  and the rows of  $H$  (that is, replace  $W(:, k)$  and  $H(k, :)$  by  $\alpha_k W(:, k)$  and  $H(k, :)/\alpha_k$  for some  $\alpha_k > 0$ ), so that their product  $WH$  remains unchanged, and hence  $D(V|WH)$  remains unchanged. Such approaches are not ideal:

- Projection requires to perform a line-search to ensure the monotonicity of the algorithm, that is, to ensure that the objective does not increase after each iteration, which may be computationally heavy.
- Renormalization of the columns of  $W$  and the rows of  $H$  is only useful when each constraint applies to the columns of  $W$  or the rows of  $H$ . It is not applicable for example for the sum-to-one constraint on the columns of  $H$  mentioned above. Moreover, in the presence of regularization terms in the objective function, it may destroy the monotonicity of the algorithm.

Another approach is to use parametrization. However, as far as we know, it does not guarantee the monotonicity of the algorithm; see Section 5.3 for more details.

**Outline and contribution** In this chapter, we introduce a general framework to design MU for  $\beta$ -NMF with disjoint linear equality constraints, and with penalty terms in the objective function. By disjoint, we mean that each variable appears in at most one equality constraint. This framework, presented in Section 5.2, does not resort to projection, renormalization, or parametrization. Our MU satisfy the set of constraints after each update of the variables during the optimization process, while guaranteeing that the objective function decreases monotonically. This framework works as follows:

- First, as for the standard MU for  $\beta$ -NMF, we majorize the objective function using a separable majorizer, that is, the majorizer is the sum of functions involving a single

variable.

- Second, we construct the augmented Lagrangian for the majorizer. Because the majorizer is separable, the problem can be decomposed into independent subproblems involving only variables that occur in the same equality constraint since they are disjoint. For a fixed value of the Lagrange multipliers, we prove that the solution of these subproblems are unique, under mild conditions (Proposition 1). Moreover, they can be written in closed form via MU for specific values of  $\beta$  and depending on the regularizer used (this is summarized in Table 5.1).
- Finally, we prove that, under mild conditions, there is a unique solution for the Lagrange multipliers so that the equality constraints are satisfied (Proposition 2). This allows us to apply the Newton-Raphson method to compute the Lagrange multipliers while guaranteeing quadratic convergence (Proposition 3).

We then showcase this framework on two NMF optimization problems, and show that it outperforms the state of the art:

1. A  $\beta$ -NMF problem with sum-to-one constraints on the columns of  $H$ , which we refer to as simplex-structured  $\beta$ -NMF (Section 5.3), and
2. A minimum-volume  $\beta$ -NMF problem with sum-to-one constraints on the columns of  $W$  (Section 5.4).

Finally, Section 5.5 shows that the framework can be extended to the case of quadratic disjoint constraints, which we showcase on sparse  $\beta$ -NMF with  $\ell_2$ -norm constraints on the columns of  $W$ .

## 5.2 General framework to design MU for $\beta$ -NMF under disjoint linear equality constraints and penalization

In this chapter, we introduce a general framework to tackle  $\beta$ -NMF with disjoint linear equality constraints, and with penalty terms in the objective function. Let us first introduce specific notations: given a matrix  $A \in \mathbb{R}^{F \times N}$  and a list of indices  $\mathcal{K} \subseteq \{(f, k) \mid 1 \leq f \leq F, 1 \leq k \leq N\}$ , we denote  $A(\mathcal{K})$  the vector of dimension  $|\mathcal{K}|$  whose entries are the entries of  $A$  corresponding to the indices within  $\mathcal{K}$ . Let us introduce  $\mathcal{K}_i$  ( $1 \leq i \leq I$ ) and  $\mathcal{B}_j$  ( $1 \leq j \leq J$ ) to be disjoint sets of indices for the entries of  $W$  and  $H$ , respectively, that is,

- $\mathcal{K}_i \subseteq \{(f, k) \mid 1 \leq f \leq F, 1 \leq k \leq K\}$  for  $i = 1, 2, \dots, I$ ,
- $\mathcal{B}_j \subseteq \{(k, n) \mid 1 \leq k \leq K, 1 \leq n \leq N\}$  for  $j = 1, 2, \dots, J$ ,
- $\mathcal{K}_u \cap \mathcal{K}_v = \emptyset$  for all  $1 \leq u, v \leq I$  and  $u \neq v$ ,
- $\mathcal{B}_q \cap \mathcal{B}_p = \emptyset$  for all  $1 \leq q, p \leq J$  and  $q \neq p$ .

We now define penalized  $\beta$ -NMF with disjoint linear equality constraints as follows

$$\begin{aligned} & \min_{W \in \mathbb{R}_+^{F \times K}, H \in \mathbb{R}_+^{K \times N}} D_\beta(V|WH) + \lambda_1 \Phi_1(W) + \lambda_2 \Phi_2(H) \\ & \text{such that} \quad \alpha_i^T W(\mathcal{K}_i) = b_i \text{ for } 1 \leq i \leq I, \\ & \quad \quad \quad \gamma_j^T H(\mathcal{B}_j) = c_j \text{ for } 1 \leq j \leq J, \end{aligned} \tag{5.2}$$

where

- the penalty functions  $\Phi_1(W)$  and  $\Phi_2(H)$  are lower bounded and admit a particular upper approximation; see Assumption 5.2.1 below.
- $\lambda_1$  and  $\lambda_2$  are the penalty weights (nonnegative scalars).
- $\alpha_i \in \mathbb{R}_{++}^{|\mathcal{K}_i|}$  ( $1 \leq i \leq I$ ) and  $\gamma_j \in \mathbb{R}_{++}^{|\mathcal{B}_j|}$  ( $1 \leq j \leq J$ ) are vectors with positive entries. Note that if  $\alpha_i$  or  $\gamma_j$  contains zero entries, the corresponding indices can be removed from  $\mathcal{K}_i$  and  $\mathcal{B}_j$ .
- $b_i$  ( $1 \leq i \leq I$ ) and  $c_j$  ( $1 \leq j \leq J$ ) are positive scalars.

As for most NMF algorithms, we propose to resort to a block coordinate descent (BCD) framework (see Algorithm 1) to solve problem (5.2): at each iteration we tackle two subproblems separately; one in  $W$  and the other in  $H$ .

The subproblems in  $W$  and  $H$  are essentially the same, by symmetry of (5.2). Hence, we may focus on solving the subproblem in  $H$  only, namely

$$\min_{H \in \mathbb{R}_+^{K \times N}} D_\beta(V|WH) + \lambda_2 \Phi_2(H) \quad \text{such that} \quad \gamma_j^T H(\mathcal{B}_j) = c_j \text{ for } 1 \leq j \leq J. \tag{5.3}$$

In order to tackle (5.3), we will design MU based on the majorization-minimization (MM) framework [129], which is the standard in the NMF literature; see [45] and the references therein. As introduced in Chapter 3, it consists in two steps: find a function that is an upper approximation of the objective and is tight at the current iterate, which is referred to as a majorizer, then minimize the majorizer to obtain the next iterate. This guarantees the objective function to decrease at each step of this iterative process.

To do so, we first provide a majorizer for the objective of (5.3) in Section 5.2.1. This majorizer has the property to be separable in each entry of  $H$ . In order to handle the equality constraints, we introduce Lagrange dual variables in Section 5.2.2, and explain how they can be computed efficiently. This allows us to derive general MU in Section 5.2.3 in the case of non-penalized  $\beta$ -NMF under disjoint linear equality constraints. This is showcased on simplex-structured  $\beta$ -NMF in Section 5.3. In section 5.4, we will illustrate on minimum-volume KL-NMF how to derive MU in the presence of penalty terms.

### 5.2.1 Separable majorizer for the objective function

Let us derive a majorizer for

$$\Psi(H) := D_\beta(V|WH) + \lambda \Phi(H), \tag{5.4}$$

that is, a function  $G(H|\tilde{H})$  satisfying (i)  $G(H|\tilde{H}) \geq \Psi(H)$  for all  $H$ , and (ii)  $G(\tilde{H}|\tilde{H}) = \Psi(\tilde{H})$ . Note that, to simplify the presentation, we denote  $\Phi_2(H) = \Phi(H)$  and  $\lambda = \lambda_2$ . To do so, let us analyze each term of  $\Psi(H)$  independently.

**Majorizing  $D_\beta(V|WH)$**  The first term  $D_\beta(V|WH)$  can be decoupled into  $n$  independent terms, one for each column  $h_n$  of  $H$ , that is,  $D_\beta(V|WH) = \sum_{n=1}^N D_\beta(v_n|Wh_n)$ . Let us focus on a specific column of  $H$ , denoted  $h \in \mathbb{R}_+^K$ . We majorize  $D_\beta(v|Wh) = \sum_{f=1}^F d_\beta(v_f|(Wh)_f)$  following the methodology introduced in [45], which consists in applying a convex-concave procedure [151] to  $d_\beta$ , as presented in Appendix 5. The resulting upper bound is given by

$$d_\beta(v_f|(Wh)_f) \leq \sum_{k=1}^K \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} \check{d}\left(v_f|\tilde{v}_f\frac{h_k}{\tilde{h}_k}\right) + \hat{d}(v_f|\tilde{v}_f) \sum_{k=1}^K w_{fk}(h_k - \tilde{h}_k) + \hat{d}(v_f|\tilde{v}_f), \quad (5.5)$$

where  $w_{fk}$  denotes the entry of matrix  $W$  at position  $(f, k)$ ,  $\tilde{v}_f := (W\tilde{h})_f$  denotes the  $f$ th entry of  $\tilde{v}$ , and  $\hat{d}$  and  $\check{d}$  are the concave and convex parts of  $d$ , respectively.

**Majorizing  $\Phi(H)$**  For the second term  $\Phi(H)$ , we rely on the following assumption for  $\Phi$ .

**Assumption 5.2.1.** *The function  $\Phi : \mathbb{R}_+^{K \times N} \mapsto \mathbb{R}$  is lower bounded, and for any  $\tilde{H} \in \mathbb{R}_+^{K \times N}$  there exists constants  $L_{kn}$  ( $1 \leq k \leq K, 1 \leq n \leq N$ ) such that the inequality*

$$\Phi(H) \leq \Phi(\tilde{H}) + \left\langle \nabla\Phi(\tilde{H}), H - \tilde{H} \right\rangle + \sum_{k,n} \frac{L_{k,n}}{2} (H - \tilde{H})_{kn}^2 \quad (5.6)$$

is satisfied for all  $H \in \mathbb{R}_+^{K \times N}$ . (Note that the constants  $L_{kn}$ 's may depend on  $\tilde{H}$ .)

Let us mention two important classes of functions satisfying Assumption 5.2.1.

1. Smooth concave functions that are lower bounded on the nonnegative orthant. For such functions, we can take  $L_{kn} = 0$  for all  $k, n$  since they are upper approximated by their first-order Taylor approximation. Note that, in this case,

$$\nabla\Phi(\tilde{H}) \geq 0, \quad (5.7)$$

otherwise we would have  $\lim_{y \rightarrow \infty} \Phi(H + ye_i e_j^T) = -\infty$ , where  $e_i$  is the  $i$ th unit vector, and this would contradict the fact that  $\Phi$  is bounded from below. This observation will be useful in the proof of Proposition 1 and is only valid for the special case  $L_{kn} = 0$  for all  $k, n$ .

Examples of such penalty functions include the sparsity-promoting regularizers  $\Phi(H) = \|H\|_p^p = \sum_{k,n} H(k, n)^p$  for  $0 < p \leq 1$  since  $H \geq 0$ .

2. Lower-bounded functions with Lipschitz continuous gradient for which (5.6) follows from the descent lemma [18].

Examples of such penalty functions include any smooth convex functions; for example any quadratic penalty, such as  $\|AH - B\|_2^2$  for some matrices  $A$  and  $B$  in which case



$L_{kn} = \sigma_1(A)^2$  for all  $k, n$ . We will encounter another example later in this chapter, namely  $\log\det(HH^\top + \delta I)$  for  $\delta > 0$  which allows to minimize the volume of the rows of  $H$ ; see Section 5.4 for the details. (Note that we will use this regularizer for  $W$ .)

**Majorizing  $\Psi(H)$**  Combining (5.5) and (5.6), we can construct a majorizer for  $\Psi(H)$ . Since both (5.5) and (5.6) are separable in each entry of  $H$ , their combination is also separable into a sum of  $K \times N$  component-wise majorizers, up to an additive constant:

$$G(H|\tilde{H}) = \sum_{n=1}^N \sum_{k=1}^K g(h_{kn}|\tilde{H}) + C(\tilde{H}), \quad (5.8)$$

where

$$\begin{aligned} g(h_{kn}|\tilde{H}) &= \sum_{f=1}^F \frac{w_{fk}\tilde{h}_{kn}}{\tilde{v}_{fn}} \check{d}\left(v_{fn}|\tilde{v}_{fn} \frac{h_{kn}}{\tilde{h}_{kn}}\right) + ah_{kn}^2 + p_{kn}h_{kn}, \\ C(\tilde{H}) &= \sum_{n=1}^N \sum_{f=1}^F \left( \hat{d}(v_{fn}|\tilde{v}_{fn}) - \sum_{k=1}^K \hat{d}'(v_{fn}|\tilde{v}_{fn})w_{fk}\tilde{h}_{kn} \right) + a_{kn}\tilde{h}_{kn}^2, \end{aligned} \quad (5.9)$$

with  $a_{kn} = \lambda \frac{L_{kn}}{2}$ , and

$$p_{kn} = \sum_{f=1}^F w_{fk} \hat{d}'(v_{fn}|\tilde{v}_{fn}) + \lambda \left( \frac{\partial \Phi}{\partial h_{kn}}(\tilde{H}) - L_{kn}\tilde{h}_{kn} \right).$$

This will allow us to minimize the majorizer  $G(H|\tilde{H})$  under the equality constraints efficiently, as presented in the next section.

### 5.2.2 Dealing with equality constraints via Lagrange dual variables

In the previous section, we derived a majorizer for  $\Psi(H)$ ,  $G(H|\tilde{H})$ , which is separable in each entry of  $H$ . Without the equality constraints, we could then compute closed-form solutions to univariate problems to minimize  $G(H|\tilde{H})$  to obtain the standard MU for NMF as in [45].

However, in problem (5.3), the entries of  $H$  in the subsets  $\mathcal{B}_j$  are not independent as they are linked with the equality constraints  $\gamma_j^T H(\mathcal{B}_j) = c_j$  for  $j = 1, 2, \dots, J$ . In fact, to minimize the majorizer under the equality constraints, we need to solve

$$\min_{H \in \mathbb{R}_+^{K \times N}} G(H|\tilde{H}) \quad \text{such that} \quad \gamma_j^T H(\mathcal{B}_j) = c_j \text{ for } 1 \leq j \leq J. \quad (5.10)$$

The variables in different sets  $\mathcal{B}_j$  can be optimized independently, as they do not interact in the majorizer nor in the constraints. Note that, for the entries of  $H$  that do not appear in any constraints, the standard MU [45] can be used. For simplicity, let us fix  $j$  and denote  $\mathcal{B} = \mathcal{B}_j$ ,  $Q = |\mathcal{B}|$ ,  $y = H(\mathcal{B}) \in \mathbb{R}_+^Q$ ,  $\gamma = \gamma_j \in \mathbb{R}_+^Q$ , and  $c = c_j > 0$ . The problems we need to solve have the form

$$\min_{y \in \mathcal{Y}} G(y|\tilde{H}), \quad (5.11)$$

where  $\mathcal{Y} = \{y \in \mathbb{R}_+^Q \mid \gamma^T y = c\}$  and

$$G(y|\tilde{H}) = \sum_{(k,n) \in \mathcal{B}} g(h_{kn}|\tilde{h}_n), \quad (5.12)$$

where the component-wise majorizers  $g(h_{kn}|\tilde{h}_n)$  are defined by (5.9). Let us introduce a convenient notation: for  $q = 1, 2, \dots, Q$ , we denote by  $(k(q), n(q))$  the  $q$ th pair belonging to  $\mathcal{B}$ . Hence the Lagrangian function of (5.12) can be written as

$$G^\mu(y|\tilde{H}) = G(y|\tilde{H}) - \mu(\gamma^T y - c) = \mu c + C(\tilde{H}) + \sum_{q=1}^Q g^\mu(y_q|\tilde{H}), \quad (5.13)$$

where

$$\begin{aligned} g^\mu(y_q|\tilde{H}) &= g(y_q|\tilde{H}) - \mu\gamma_q y_q \\ &= \sum_{f=1}^F \frac{w_{fk(q)}\tilde{y}_q}{\tilde{v}_{fn(q)}} \check{d} \left( v_{fn(q)}|\tilde{v}_{fn(q)} \frac{y_q}{\tilde{y}_q} \right) + a_q y_q^2 + (p_q - \mu\gamma_q) y_q, \end{aligned} \quad (5.14)$$

$$p_q = \sum_{f=1}^F w_{fk(q)} \hat{d} \left( v_{fn(q)}|\tilde{v}_{fn(q)} \right) + \lambda \left( \frac{\partial \Phi}{\partial y_q}(\tilde{H}) - L_{k(q)n(q)} \tilde{y}_q \right), \quad (5.15)$$

and  $\mu \in \mathbb{R}$ . Note that  $G^\mu$  is separable, as is  $G$ , because the term  $\gamma^T y$  is linear.

Assume for now that the Lagrangian multiplier  $\mu$  is known, and let us minimize  $G^\mu(y|\tilde{H})$  on  $(0, \infty)^Q$ . Such a problem is separable under the form of  $L$  subproblems, consisting in minimizing univariate functions  $g^\mu(\cdot|\tilde{H})$  separately over  $(0, \infty)$ . We now show in Proposition 1 that, under mild conditions, each subproblem admits a unique solution over  $(0, \infty)$ .

**Proposition 1.** *Let  $q \in \{1, 2, \dots, Q\}$ . Assume that  $\beta < 2$  and  $\tilde{y}_q, v_{fn(q)}, w_{fk(q)} > 0$  for all  $f$ . Moreover, if  $\beta \leq 1$  and  $a = 0$ , assume that  $\mu < \frac{p_q}{\gamma_q}$ . Then there exists a unique minimizer  $y_q^*(\mu)$  of  $g^\mu(y_q|\tilde{H})$  in  $(0, \infty)$ .*

*Proof.* According to Proposition 4 (see Appendix 5), each  $g^\mu$  is  $C^\infty$  and strictly convex on  $(0, \infty)$ , so its infimum is uniquely attained in the closure of  $(0, \infty)$ . We have to prove that it is neither reached at 0 nor at  $\infty$ .

On the one hand, from (5.14), we have

$$(g^\mu)'(y_q|\tilde{H}) = \sum_{f=1}^F w_{fk(q)} \check{d}' \left( v_{fn(q)}|\tilde{v}_{fn(q)} \frac{y_q}{\tilde{y}_q} \right) + 2a_q y_q + p_q - \gamma_q \mu \quad (5.16)$$

and, for any  $\beta < 2$  and any  $x > 0$ ,

$$\lim_{y \rightarrow 0^+} \check{d}'(x|y) = -\infty,$$

so  $\lim_{y_q \rightarrow 0^+} (g^\mu)'(y_q|\tilde{H}) = -\infty$ , which ensures that the infimum is not reached at 0.

On the other hand,

$$\lim_{y \rightarrow \infty} \check{d}'(x|y) = \begin{cases} 0 & \text{if } \beta \leq 1, \\ \infty & \text{otherwise.} \end{cases} \quad (5.17)$$

According to (5.16) and (5.17), the distinction must be made between two cases:

- If  $a_q > 0$  or  $\beta \in (1, 2)$ :  $\lim_{y_q \rightarrow \infty} (g^\mu)'(y_q|\tilde{H}) = \infty$ , so the infimum is reached for a finite  $y_q$ .
- If  $a_q = 0$  and  $\beta \leq 1$ :  $\lim_{y_q \rightarrow \infty} (g^\mu)'(y_q|\tilde{H}) = p_q - \gamma_q \mu$ , so the same conclusion holds if  $\mu < \frac{p_q}{\gamma_q}$ .

□

We just proved that, under mild conditions, each  $g^\mu$  has a unique minimizer over  $(0, \infty)$ .

However we assumed that the value of  $\mu$  is fixed. Now given  $y^*(\mu) = [y_1^*(\mu), \dots, y_Q^*(\mu)]^T$ , let us show that the solution to  $\gamma^T y^*(\mu) = c$  is unique. The corresponding value of  $\mu$ , which we denote  $\mu^*$ , provides the minimizer  $y^*(\mu^*)$  of  $G^\mu(y|\tilde{H})$  that satisfies the linear constraint  $\gamma^T y^*(\mu^*) = c$ . Moreover,  $\mu^*$  naturally fulfills  $\mu^* < \frac{p_q}{\gamma_q}$  for all  $q$  when  $\beta \leq 1$  and  $a_q = 0$ , as required in Proposition 1.

**Proposition 2.** *Assume that  $\beta < 2$  and  $\tilde{y}_q, v_{fn(q)}, w_{fk(q)} > 0$  for all  $q, f$ . Then the scalar equation  $\gamma^T y^*(\mu) = c$  in the variable  $\mu$  admits a unique solution  $\mu^*$  in  $(-\infty, t)$ , where*

$$t = \begin{cases} \min_q \frac{p_q}{\gamma_q} \geq 0 & \text{if } \beta \leq 1 \text{ and } a_q = 0, \\ \infty & \text{otherwise,} \end{cases}$$

such that  $y^*(\mu^*)$  is the unique solution to problem (5.11).

*Proof.* Under the conditions of Proposition 1,  $g^\mu(y_q|\tilde{H})$  has a unique minimizer  $y_q^*(\mu)$  for each  $q$ . By the first-order optimality condition,  $y_q^*(\mu)$  is a solution of  $(g^\mu)'(y_q|\tilde{H}) = 0$  or equivalently, by (5.16), a solution of  $\gamma_q^{-1} g'(y_q|\tilde{H}) = \mu$  over  $(0, \infty)$  where

$$\gamma_q^{-1} g'(y_q|\tilde{H}) = \gamma_q^{-1} \sum_{f=1}^F w_{fk(q)} \check{d}' \left( v_{fn(q)} |\tilde{v}_{fn(q)} \frac{y_q}{\tilde{y}_q} \right) + 2 \frac{a_q}{\gamma_q} y_q + \frac{p_q}{\gamma_q} \quad (5.18)$$

is strictly increasing on  $(0, \infty)$  (since  $g$  is strictly convex) and one-to-one from  $(0, \infty)$  to an open interval  $T_q = (t_q^-, t_q^+)$  where

$$t_q^- = \lim_{y_q \rightarrow 0} g'(y_q|\tilde{H}) = -\infty, \quad (5.19)$$

$$t_q^+ = \lim_{y_q \rightarrow \infty} g'(y_q|\tilde{H}) = \begin{cases} \frac{p_q}{\gamma_q} & \text{if } \beta \leq 1 \text{ and } y_q = 0, \\ \infty & \text{otherwise.} \end{cases} \quad (5.20)$$

Moreover,  $p_q \geq 0$  if  $y_q = 0$  (then  $L_{k(q)n(q)} = 0$ ) and  $\beta \leq 1$  according to (5.7) and (5.15). As a consequence,  $\gamma_q^{-1} g'(y_q^*|\tilde{y}_q) = \mu$  is equivalent to

$$y_q^*(\mu) = (g')^{-1}(\gamma_q \mu), \quad (5.21)$$

where  $\mu \in T_q$  and  $(g')^{-1}$  denotes the inverse function of  $g'$ .

Coming back to the multivariate problem (5.11), we must find a value  $\mu^*$  of the Lagrangian multiplier such that the constraint  $\gamma^T y^*(\mu) = c$  is satisfied. Given (5.21),  $\mu^*$  is a solution of

$$\sum_{q=1}^Q \gamma_q (g')^{-1}(\gamma_q \mu) = c. \quad (5.22)$$

Each  $g'(y_q|\tilde{H})$  being strictly increasing on  $(0, \infty)$ ,  $(g')^{-1}(\gamma_q \mu)$  is also strictly increasing (from  $T_q$  to  $(0, \infty)$ ), this is a direct consequence of  $(f^{-1})' = \frac{1}{f' \circ f^{-1}}$  where  $f$  is any strictly increasing function on some interval. Finally  $\sum_{q=1}^Q \gamma_q (g')^{-1}(\gamma_q \mu)$  is strictly increasing from  $\cap_{j=1}^J T_q = (-\infty, t)$  to  $(0, \infty)$ , with  $t \geq 0$ . Therefore, the solution  $\mu^*$  is unique.  $\square$

Proposition 2 shows that the optimal Lagrangian multiplier is the unique solution of (5.22). It is clear that finding the solution of (5.22) is equivalent to finding the root of a function  $r(\mu)$ . We propose here-under to use a Newton-Raphson method to compute  $\mu^*$ , and show that this method generates a sequence of iterates  $\mu_n$  that converges towards  $\mu^*$  at a quadratic speed.

**Proposition 3.** *Assume that  $\beta < 2$  and  $\tilde{y}_q, v_{fn(q)}, w_{fk(q)} > 0$  for all  $q, f$ . Let*

$$r(\mu) = \sum_{q=1}^Q \gamma_q (g')^{-1}(\gamma_q \mu) - c$$

for  $\mu \in (-\infty, t)$ , and denote  $\mu^*$  the unique solution of  $r(\mu) = 0$ . From any initial point  $\mu_0 \in (\mu^{**}, t)$ , Newton-Raphson's iterates

$$\mu_{n+1} = \mu_n - \frac{r(\mu_n)}{r'(\mu_n)}$$

decrease towards  $\mu^*$  at a quadratic speed.

*Proof.* We already know that  $r$  is strictly increasing from  $(-\infty, t)$  to  $(0, \infty)$ . Let us show that  $r$  is also strictly convex. According to the third item of Proposition 4 in Appendix 5,  $\tilde{d}''(x|y)$  is completely monotonic, so it is strictly decreasing in  $y$ . Equivalently,  $\tilde{d}'(x|y)$  is strictly concave in  $y$ , and each  $g'$  is also strictly concave according to (5.18). Since the inverse of a strictly increasing, strictly concave function  $f$  is strictly increasing and strictly convex, which is a direct consequence of  $(f^{-1})'' = -\frac{f'' \circ f^{-1}}{(f' \circ f^{-1})^3}$ , then each  $(g')^{-1}$  is strictly convex, and finally,  $r$  is strictly convex.

For any  $\mu_0 \in (\mu^*, t)$ , we have  $r(\mu_0) > 0$ , so  $\mu_1 = \mu_0 - \frac{r(\mu_0)}{r'(\mu_0)} < \mu_0$ . We have also  $\mu_1 > \mu^*$  as a consequence of the strict convexity of  $r$ . By immediate recurrence, we obtain that  $\mu_n$  is a decreasing series that converges towards  $\mu^*$ . According to [111], it converges at a quadratic speed since  $|r'|$  and  $|r''|$  are bounded away from 0 in  $[\mu^*, \mu_0]$ .  $\square$

**Discussion** At this point, we have derived an optimization framework to tackle problem (5.11). The optimal Lagrangian multiplier value is determined before each majorization-minimization update using a Newton-Raphson algorithm. However, such a formal solution

is implementable if and only if each  $y_q^*(\mu)$  can be actually computed as the minimizer of  $g^\mu(y_q|\tilde{H})$  in  $(0, \infty)$ . In some cases, computing  $y_q^*(\mu)$  is equivalent to extracting the roots of a polynomial of a degree smaller or equal to four, which is possible in closed form. In other cases, we have to solve a polynomial equation of degree larger than four, or even an equation that is not polynomial. Table 5.1 indicates the cases where a closed-form solution is available, and hence when our framework can be efficiently implemented. We observe

	$\beta \in (-\infty, 1) \setminus \{0\}$	$\beta = 0$	$\beta = 1$	$\beta \in (1, 2)$			
				$\frac{5}{4}$	$\frac{4}{3}$	$\frac{3}{2}$	other
No penalization	1	1	1	1	1	1	1
$L_{kn} = 0$ for all $k, n$	1	1	1	3	4	2	‡
$L_{kn} > 0$ for some $k, n$	‡	3	2	‡	‡	3	‡

Table 5.1: Cases where (5.21) can be computed in closed form. They are indicated by the degree of the corresponding polynomial equation, otherwise the symbol ‡ is used. The constants  $L_{kn}$ 's is the one needed in Assumption 5.2.1 for the penalization functions  $\Phi_1(H)$  and  $\Phi_2(W)$ ; see (5.6).

that, without penalization, the polynomial equation is of degree one, and hence always admit a closed form. This particular case is discussed in the next Section, which we will exemplify in Section 5.3 with  $\beta$ -NMF with sum-to-one constraints on the columns of  $H$ . In Section 5.4, we will present an important example with  $L_{kn} > 0$  for all  $k, n$  and  $\beta = 1$ , namely minimum-volume KL-NMF.

### 5.2.3 MU for $\beta$ -NMF with disjoint linear equality constraints without penalization

In this section, we derive an algorithm based on the general framework presented in the previous section to tackle the  $\beta$ -NMF problem under disjoint linear equality constraints without penalization, that is, problem (5.2) with  $\lambda_1 = \lambda_2 = 0$ . We consider this simplified case here as it allows to provide explicit MU for any value of  $\beta < 2$ ; see the first row 'No penalization' of Table 5.1. These updates satisfy the constraints after each update of  $W$  or  $H$ , and monotonically decrease the objective function  $D_\beta(V|WH)$ .

Let us then consider the subproblem of (5.2) over  $H$  when  $W$  is fixed and with  $\lambda_2 = 0$ , that is,

$$\min_{H \in \mathbb{R}_+^{K \times N}} D_\beta(V|WH) \quad \text{such that} \quad \gamma_j^T H(\mathcal{B}_j) = c_j \text{ for } 1 \leq j \leq J. \quad (5.23)$$

Let us follow the framework presented above. First, an auxiliary function, which we denote  $G(H|\tilde{H})$ , is constructed at the current iterate  $\tilde{H}$  so that it majorizes the objective

for all  $H$  and is defined as follows:

$$G(H|\tilde{H}) = \sum_{f,n} \left[ \sum_k \frac{w_{fk} \tilde{h}_{kn}}{\tilde{v}_{fn}} \check{d} \left( v_{fn} | \tilde{v}_{fn} \frac{h_{kn}}{\tilde{h}_{kn}} \right) \right] + \left[ \hat{d}'(v_{fn}|\tilde{v}_{fn}) \sum_k w_{fk} (h_{kn} - \tilde{h}_{kn}) + \hat{d}(v_{fn}|\tilde{v}_{fn}) \right], \quad (5.24)$$

where  $\check{d}(\cdot|\cdot)$  and  $\hat{d}(\cdot|\cdot)$  are the ones defined in Appendix 5. Second, we need to minimize  $G(H|\tilde{H})$  while imposing the set of linear constraints  $\gamma_j^T H(\mathcal{B}_j) = c_j$ . The Lagrangian function of  $G$  is given by

$$G^\mu(H|\tilde{H}) = G(H|\tilde{H}) - \sum_j [\mu_j (\gamma_j^T H(\mathcal{B}_j) - c_j)], \quad (5.25)$$

where  $\mu_j$  are the Lagrange multipliers associated to each linear constraint  $\gamma_j^T H(\mathcal{B}_j) = c_j$ . We observe that  $G^\mu$  in (5.25) is a separable majorizer in the variables  $H$  of the Lagrangian function  $D_\beta(V|WH) - \sum_j [\mu_j (\gamma_j^T H(\mathcal{B}_j) - c_j)]$ . Due to the disjointness of each subset of variables  $\mathcal{B}_j$  (5.25), we only consider the optimization over one specific subset  $\mathcal{B}_j$ . The minimizer (5.21) of  $G^\mu(H(\mathcal{B}_j)|\tilde{H}(\mathcal{B}_j))$  has the following component-wise expression:

$$H^*(\mathcal{B}_j) = \tilde{H}(\mathcal{B}_j) \odot \left( \frac{[C(\mathcal{B}_j)]}{[D(\mathcal{B}_j) - \mu_j \gamma_j]} \right)^{(\gamma(\beta))}, \quad (5.26)$$

where  $C = W^T \left( (WH)^{(\beta-2)} \odot V \right)$ ,  $D = W^T (WH)^{(\beta-1)}$ .

Finally, we need now to find the optimal value for  $\mu_j$ , denoted  $\mu_j^*$ , which is the solution of  $\gamma_j^T H^*(\mathcal{B}_j) = c_j$ .

It requires to finding the root of the function

$$r_j(\mu_j) = \sum_{q=1}^Q \gamma_{j,q} \left[ \tilde{H}(\mathcal{B}_j) \odot \left( \frac{[C(\mathcal{B}_j)]}{[D(\mathcal{B}_j) - \mu_j \gamma_j]} \right)^{(\gamma(\beta))} \right]_q - c_j, \quad (5.27)$$

where  $[A]_q$  denotes the  $q$ -th entry of expression  $A$ . Proposition 2 shows that  $\mu_j^*$  is unique on some interval  $(-\infty, t)$ . Indeed,  $r_j(\mu_j)$  is a finite sum of elementary rationale functions of  $\mu_j$  and each of them is an increasing, convex function in  $\mu_j$  over  $(-\infty, t_q)$  with  $t_q = \frac{D_q(\mathcal{B}_j)}{\gamma_{j,q}}$  for all  $\beta$ . It is even completely monotone for all  $\beta$  in  $(-\infty, t_q)$  because  $\gamma(\beta) > 0$  [103]. As a consequence  $r_j(\mu_j)$  is also a completely monotone, convex increasing function of  $\mu_j$  in  $(-\infty, t)$ , where  $t = \min(t_q)$ . Finally, we can easily show that the function  $r_j(\mu_j)$  changes of sign on the interval  $(-\infty, t)$  by computing two limits at the closure of the interval. As  $\mu^* \in (-\infty, t)$ , the update (5.26) is nonnegative. To evaluate  $\mu^*$ , we use a Newton-Raphson method, with any initial point  $\mu_0 \in (\mu^*, t)$ , with a quadratic rate of convergence as demonstrated in Proposition 3. Algorithm 5 summarizes our method to tackle (5.2) for all the  $\beta$ -divergences which we refer to as disjoint-constrained  $\beta$ -NMF algorithm. The update for matrix  $W$  can be derived in the same way, by symmetry of the problem.

**Computational cost** The computational cost of Algorithm 5 is asymptotically equivalent to the standard MU for  $\beta$ -NMF, that is, it requires  $\mathcal{O}(FNK)$  operations per iteration. Indeed, the complexity is mainly driven by matrix products required to compute  $C$  and  $D$ ; see (5.26). To compute the roots of (5.27) corresponding to  $H$  using Newton-Raphson, each iteration requires to compute  $r_j(\mu_j)/r'_j(\mu_j)$  for all  $j$  which requires at most  $\mathcal{O}(KN)$  operations. Finding the roots therefore requires  $\mathcal{O}(KN)$  operations times the number of Newton-Raphson iterations. By symmetry, it requires  $\mathcal{O}(KF)$  operations to compute the roots corresponding to  $W$ . Because of the quadratic convergence, the number of iterations required for the convergence of the Newton-Raphson method is typically small, namely between 10 to 100 in our experiments using the stopping criterion  $|r(\mu_j)| \leq 10^{-6}$  for all  $j$ . Therefore, in practice, the overall complexity of Algorithm 5 is dominated by the matrix products that require  $\mathcal{O}(FNK)$  operations. The same conclusions apply to the algorithms presented in Sections 5.3, 5.4 and 5.5, and this will be confirmed by our numerical experiments.

---

**Algorithm 5**  $\beta$ -NMF with disjoint linear constraints

---

**Require:** A matrix  $V \in \mathbb{R}^{F \times N}$ , an initialization  $H \in \mathbb{R}_+^{K \times N}$  and  $W \in \mathbb{R}^{F \times K}$ , a factorization rank  $K$ , a maximum number of iterations, maxiter, a value for  $\beta$ , and the linear constraints defined by  $\mathcal{K}_i$ ,  $\alpha_j$  and  $b_i$  for  $i = 1, 2, \dots, I$ , and  $\mathcal{B}_j$ ,  $\gamma_j$  and  $c_j$  for  $j = 1, 2, \dots, J$ .

**Ensure:** A rank- $K$  NMF  $(W, H)$  of  $V$  satisfying constraints in (5.2).

```

1: for it = 1 : maxiter do
2:     % Update of matrix H
3:      $C \leftarrow W^T \left( (WH)^{\cdot(\beta-2)} \odot V \right)$ 
4:      $D \leftarrow W^T (WH)^{\cdot(\beta-1)}$ 
5:     for j = 1 : J do
6:          $\mu_j \leftarrow \text{root} (r_j(\mu_j))$  % see Equation (5.27)
7:          $H(\mathcal{B}_j) \leftarrow H(\mathcal{B}_j) \odot \left( \frac{[C(\mathcal{B}_j)]}{[D(\mathcal{B}_j) - \mu_j \gamma_j]} \right)^{\cdot(\gamma(\beta))}$ 
8:     end for
9:      $\mathcal{B}_c = \{(k, n) \mid 1 \leq k \leq K, 1 \leq n \leq N\} \setminus (\cup_j^J \mathcal{B}_j)$ . %  $\mathcal{B}_c$  is the complement of  $\cup_j^J \mathcal{B}_j$ 
10:     $H(\mathcal{B}_c) \leftarrow H(\mathcal{B}_c) \odot \left( \frac{[C(\mathcal{B}_c)]}{[D(\mathcal{B}_c)]} \right)^{\cdot(\gamma(\beta))}$ 
11:    % Update of matrix W
12:    W is updated in the same way as H, by symmetry of the problem.
13: end for
```

---

### 5.3 Showcase 1: Simplex-structured $\beta$ -NMF

In this section, we showcase a particularly important example of  $\beta$ -NMF with linear disjoint constraints and no penalization, namely, the simplex-structured matrix factorization (SSMF) problem. It is defined as follows: given a data matrix  $V \in \mathbb{R}^{F \times N}$  and a factorization rank  $K$ , SSMF refers to the problem of computing  $W$  and  $H$  such that  $V \approx WH$  and the columns of  $H$  lie on the unit simplex, that is, the entries of each column of  $H$  are nonnegative and sum to one. SSMF is a powerful tool in many applications such as hyperspectral unmixing in geoscience and remote sensing [21, 98, 1], document analysis and time-resolved Raman spectroscopy. We refer the reader to the recent survey [50] for more applications and details about SSMF.

To understand the underlying significance of SSMF, it is necessary to give more insights on a research topic for which important SSMF techniques were initially developed which is the BHU, a main research topic in remote sensing. As explained in Section 1.5.1, the task of BHU is to decompose a remotely sensed hyperspectral image into endmember spectral signatures and the corresponding abundance maps with limited prior information, usually the only known information being the number of endmembers. In this context, the columns of  $W$  correspond to the endmembers spectral signatures and the columns of  $H$  contain the proportion of the endmembers in each column of  $V$ , so the column-stochastic assumption for  $H$  naturally holds. For many SSMF-based methods, the nonnegativity constraint  $W \geq 0$  is assumed as many hyperspectral images are such that  $V \geq 0$ . The resulting NMF model is referred to as *SSNMF* and was previously introduced in Section 1.8.4 in the exact case. Since there is potentially noise within the input data, we consider the following approximate SSNMF model

$$V \approx WH \text{ such that } W \in \mathbb{R}_+^{F \times K}, H \in \mathbb{R}_+^{K \times N}, e^T H = e^T, \text{ with } K \ll \min(F, N).$$

We refer to the associated optimization problem as simplex-structured nonnegative matrix factorization with the  $\beta$ -divergence ( $\beta$ -SSNMF), and is formulated as follows:

$$\min_{W \in \mathbb{R}_+^{F \times K}, H \in \mathbb{R}_+^{K \times N}} D_\beta(V|WH) \quad \text{such that} \quad e^T h_j = 1 \text{ for } 1 \leq j \leq N, \quad (5.28)$$

where  $e$  is the vector of all ones of appropriate dimension. This is particular case of (5.2) where

- the subsets  $\mathcal{B}_j$  correspond to the columns of  $H$ , and there is no subset  $\mathcal{K}_i$  (no constraint on  $W$ ),
- $\gamma_j^T = e$  and  $c_j = 1$  for  $j = 1, 2, \dots, N$ .

Hence Algorithm 5 can be directly applied to (5.28).

**Numerical experiments** Let us perform numerical experiments to evaluate the effectiveness of Algorithm 5 on the simplex-structure  $\beta$ -NMF problem against existing methods.



To the best of our knowledge, the so-called group robust NMF (GR-NMF) algorithm<sup>1</sup> from [44] is the most recent algorithm that is able to tackle problem (5.28) for the full range of  $\beta$ -divergences. The approach is not based on Lagrangian multipliers but introduces a change of variables for matrix  $H$ . This approach, initially used for NMF in [39], does not provide an auxiliary function for the subproblem in  $H$  and resort to a heuristic commonly used in NMF, see, *e.g.*, [137, 43]. Therefore there is no guarantee that the objective function is decreasing at each update of the abundance matrix, unlike Algorithm 5.

We apply Algorithm 5 and GR-NMF on three widely used real hyperspectral data sets<sup>2</sup> [154]:

- Samson: 156 spectral bands with  $95 \times 95$  pixels, containing mostly 3 materials ( $K = 3$ ), namely "Soil", "Tree" and "Water".
- Jasper Ridge: 198 spectral bands with  $100 \times 100$  pixels, containing mostly 4 materials ( $K = 4$ ), namely "Road", "Soil", "Water" and "Tree".
- Cuprite: 188 spectral bands with  $250 \times 190$  pixels, containing mostly 12 types of minerals ( $K = 12$ ).

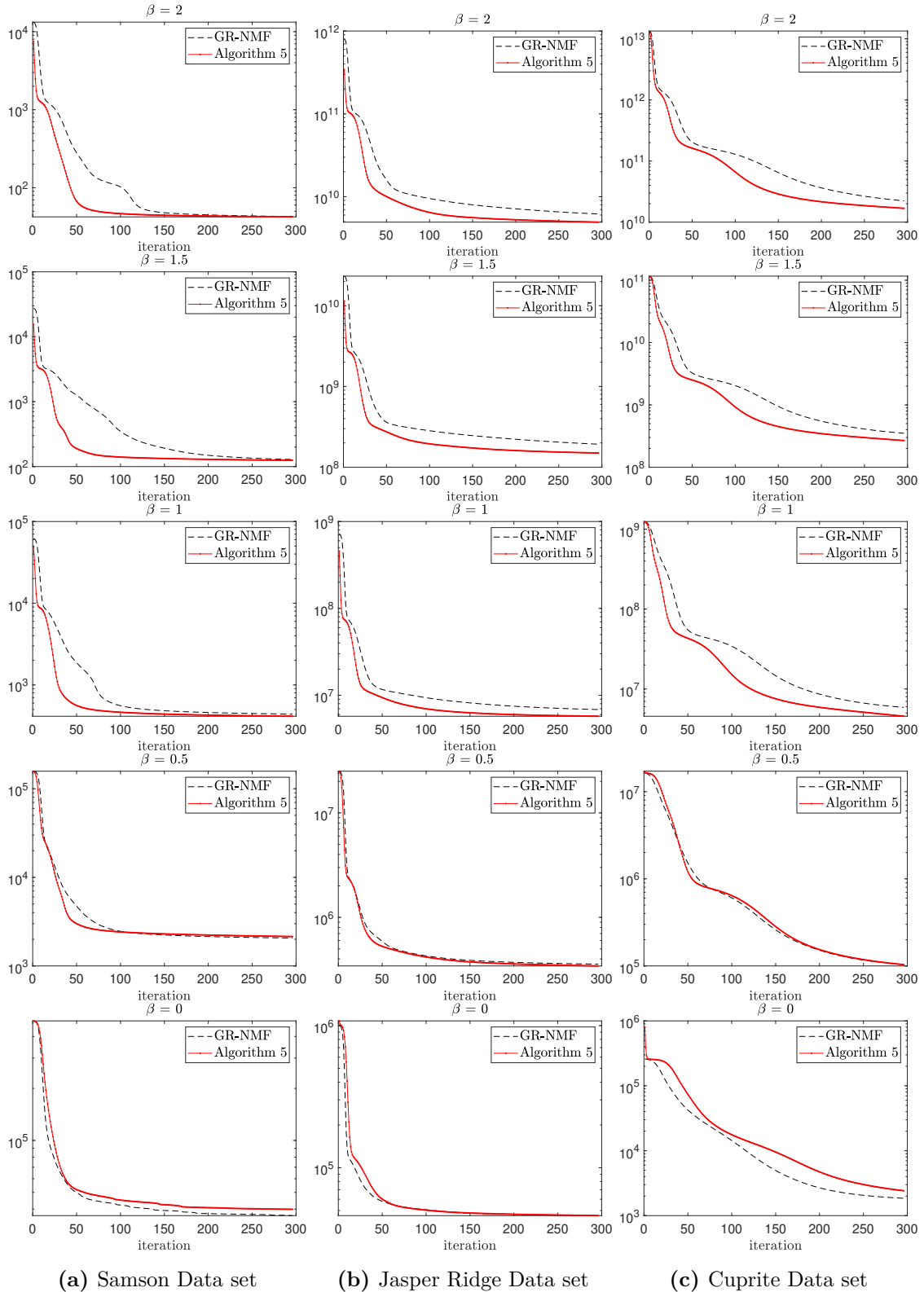
$\beta$ -SSNMF has shown itself as a powerful one to tackle BHU, hence this comparative study between Algorithm 5 and GR-NMF [44] focuses on the convergence aspects including the evolution of the objective function and the runtime. The algorithms are implemented and tested on a desktop computer with Intel Core i7-8700@3.2GHz CPU and 32GB memory. The codes are written in MATLAB R2018a. The code for Algorithm 5 applied to (5.28) is available from <https://bit.ly/31QWqz1>. The algorithms are compared for the following values for  $\beta \in \{0, \frac{1}{2}, 1, \frac{3}{2}, 2\}$ . For all simulations, the algorithms are run for 20 random initializations of  $W$  and  $H$  (each entry sampled from the uniform distribution in  $[0, 1]$ ). Table 5.2 reports the average and standard deviation of the runtime (in seconds) as the final value for the objective function over these 20 runs for a maximum of 300 iterations.

We observe that Algorithm 5 outperforms the GR-NMF in terms of runtime and final value for the objective functions for all test cases except when  $\beta = 0$  for the Samson and Cuprite data sets. In particular, for  $\beta = 1$ , Algorithm 5 is up to 2.5 times faster than the GR-NMF. For the Cuprite data set with  $\beta = 1/2$ , Algorithm 5 and GR-NMF perform similarly. We also observe that the standard deviations obtained with Algorithm 5 are in general significantly smaller for all  $\beta$ , except for  $\beta = 0$  for the Samson and Cuprite data sets.

In the following, we provide figures that show the evolution of the objective as a function of the iterations, and that confirm the observations above. Figure 5.1 displays the evolution of the objective function of  $\beta$ -SSNMF for Algorithm 5 and GR-NMF [44] on the experiments described above. We confirm that Algorithm 5 performs better than GR-NMF [44], except for  $\beta = 0$ .

<sup>1</sup><https://www.irit.fr/~Cedric.Fevotte/extras/tip2015/code.zip>

<sup>2</sup><http://lesun.weebly.com/hyperspectral-data-set.html>



**Fig. 5.1.** Averaged objective functions over 20 random initializations obtained for Algorithm 5 (red line with circle markers) and the GR-NMF (black dashed line) applied to the three data sets detailed in the text for 300 iteration. The comparison is performed for different values of  $\beta$ , from top to bottom:  $\beta = 2$ ,  $\beta = 3/2$ , and  $\beta = 1$ . Logarithmic scale for y axis.

Table 5.2: Runtime performance in seconds and final value of objective function  $F_{\text{end}}(W, H)$  for Algorithm 5 and the GR-NMF reported for  $\beta \in \{0, \frac{1}{2}, 1, \frac{3}{2}, 2\}$ . The table reports the average and standard deviation over 20 random initializations with a maximum of 300 iterations for three hyperspectral data sets.

Algorithms	Samson		Jasper Ridge		Cuprite	
	runtime (s.)	$F_{\text{end}}(W, H)$	runtime (s.)	$F_{\text{end}}(W, H)$	runtime (s.)	$F_{\text{end}}(W, H)$
$\beta = 2$						
Algorithm 5	<b>16.62</b> ±0.15	<b>42.37</b> ±0.92	<b>22.86</b> ±0.08	<b>(4.93 ± 0.41)</b> 10 <sup>9</sup>	121.04 ± 0.62	<b>(1.68 ± 0.10)</b> 10 <sup>10</sup>
GR-NMF	18.23±0.29	42.86±1.17	25.32±0.16	(6.19 ± 1.28)10 <sup>9</sup>	<b>114.27 ± 0.20</b>	(2.23 ± 0.12)10 <sup>10</sup>
$\beta = 3/2$						
Algorithm 5	<b>63.69</b> ±0.40	<b>124.82 ± 38.80</b>	<b>89.23 ± 0.30</b>	<b>(1.50 ± 0.08)</b> 10 <sup>8</sup>	<b>421.49 ± 2.79</b>	<b>(2.68 ± 0.12)</b> 10 <sup>8</sup>
GR-NMF	80.09±0.60	128.59 ± 31.22	112.72 ± 0.67	(1.93 ± 0.42)10 <sup>8</sup>	508.57 ± 3.50	(3.50 ± 0.16)10 <sup>8</sup>
$\beta = 1$						
Algorithm 5	<b>18.33 ± 0.08</b>	<b>413.76 ± 30.96</b>	<b>24.82 ± 0.35</b>	<b>(5.73 ± 0.20)</b> 10 <sup>6</sup>	<b>182.98 ± 14.14</b>	<b>(4.56 ± 0.21)</b> 10 <sup>6</sup>
GR-NMF	44.78 ± 0.18	439.40 ± 44.91	62.83 ± 0.76	(6.85 ± 1.41)10 <sup>6</sup>	370.25 ± 21.33	(5.89 ± 0.22)10 <sup>6</sup>
$\beta = 1/2$						
Algorithm 5	<b>89.80 ± 0.65</b>	(2.15 ± 0.22)10 <sup>3</sup>	<b>126.43 ± 0.61</b>	<b>(3.43 ± 0.30)</b> 10 <sup>5</sup>	682.80 ± 3.32	<b>(1.03 ± 0.05)</b> 10 <sup>5</sup>
GR-NMF	102.21 ± 0.72	<b>(2.06 ± 0.26)</b> 10 <sup>3</sup>	141.75 ± 0.69	(3.56 ± 0.39)10 <sup>5</sup>	<b>642.49 ± 1.22</b>	(1.04 ± 0.05)10 <sup>5</sup>
$\beta = 0$						
Algorithm 5	<b>52.89</b> ±0.54	(3.99 ± 0.57)10 <sup>4</sup>	<b>69.59</b> ±0.44	<b>(4.61 ± 0.14)</b> 10 <sup>4</sup>	479.84 ± 16.02	(2.42 ± 0.17)10 <sup>3</sup>
GR-NMF	55.61±0.47	<b>(3.67 ± 0.68)</b> 10 <sup>4</sup>	77.87±0.63	(4.61 ± 0.53)10 <sup>4</sup>	<b>354.65 ± 6.01</b>	<b>(1.85 ± 0.06)</b> 10 <sup>3</sup>

## 5.4 Showcase 2: minimum-volume KL-NMF

In this section, we showcase another important example of  $\beta$ -NMF with linear disjoint constraints, namely, the minimum volume NMF with  $\beta$ -divergences (min-vol  $\beta$ -NMF) optimization problem. This optimization problem was introduced in Section 3.2 by (3.1) and is based on the minimization of  $\beta$ -divergences including a penalty term promoting solutions with minimum volume spanned by the columns of the matrix  $W$ . Problem (3.1) is associated to the following approximate NMF model:

$$V \approx WH \text{ such that } W \in \mathbb{R}_+^{F \times K}, H \in \mathbb{R}_+^{K \times N}, W^T e = e, \text{ with } K \ll \min(F, N). \quad (5.29)$$

We recall the problem (3.1) here-under:

$$\min_{W(:,j) \in \Delta^F \forall j, H \geq 0} D_\beta(V|WH) + \lambda \log \det(W^T W + \delta I),$$

where  $\Delta^F = \{x \in \mathbb{R}_+^F \mid \sum_{i=1}^F x_i = 1\}$  is the unit simplex  $\lambda$  is a penalty parameter and  $\log \det(W^T W + \delta I)$  is a function that measures the volume spanned by the columns of  $W$  where  $\delta$  is a small positive constant that prevents  $\log \det(W^T W)$  to go to  $-\infty$  when  $W$  tends to a rank-deficient matrix (that is, when  $r = \text{rank}(W) < K$ ).

We showed in Chapter 3 that problem (3.1) is particularly powerful as it leads to identifiability which is crucial in many applications such as in hyperspectral imaging or audio source separation. Indeed, under some mild assumptions and in the exact case, we demonstrated that (3.1) is able to identify the groundtruth factors  $(W^\#, H^\#)$  that generated

the input data  $V$ , in the absence of noise, see Theorem 3.2.1. In Chapter 3, (3.1) is used for blind audio source separation. We have to mention that (3.1) is also well suited for hyperspectral imaging when  $\beta = 2$ . Indeed, [56] shows the efficiency of (3.1) to tackle the BHU problem.

In the next sections, we show that we can tackle the min-vol  $\beta$ -NMF optimization problem defined in (3.1) with the general framework presented in Section 5.2 in the case  $\beta = 1$ .

### 5.4.1 Problem formulation and algorithm

As the minimum-volume penalty of model (3.1) concerns matrix  $W$  only, the main challenge concerns the update of  $W$ . Indeed, the update of  $H$  is simply the one from [85]. Let us therefore consider the subproblem in  $W$  for  $H$  fixed:

$$\begin{aligned} \min_{W \in \mathbb{R}^{F \times K}} \quad & F(W) = D_\beta(V|WH) + \lambda \log \det(W^T W + \delta I) \\ \text{subject to} \quad & W \geq 0 \\ & e^T W(:, i) = 1 \text{ for } 1 \leq i \leq K, \end{aligned} \quad (5.30)$$

where  $e$  is the all-one column vector of appropriate dimension. Considering the general model (5.2), we have that:

- the subsets  $\mathcal{K}_i$  correspond to the columns of  $W$ , and there is no subset  $\mathcal{B}_j$ ,
- $\alpha_i^T = e$  and  $b_i = 1$  for  $1 \leq i \leq K$ .

To upper bound  $\log \det(W^T W + \delta I)$  as required by (5.6) in Assumption 5.2.1, we majorize it using a convex quadratic separable auxiliary function provided in (3.8) and (3.9) and which is derived as follows. First, the concave function  $\log \det(Q)$  for  $Q > 0$  can be upper bounded using the first-order Taylor approximation: for any  $\tilde{Q} > 0$ ,

$$\log \det(Q) \leq \log \det(\tilde{Q}) + \langle \tilde{Q}^{-1}, Q - \tilde{Q} \rangle = \langle \tilde{Q}^{-1}, Q \rangle + \text{cst},$$

where cst is some constant independent of  $Q$ . For any  $W, \tilde{W}$ , and denoting  $\tilde{Q} = \tilde{W}^T \tilde{W} + \delta I > 0$ , we obtain

$$\log \det(W^T W + \delta I) \leq \langle \tilde{Q}^{-1}, W^T W \rangle + \text{cst} = \text{Tr}(W \tilde{Q}^{-1} W^T) + \text{cst},$$

which is a convex quadratic and Lipschitz-smooth function in  $W$ . In fact, letting  $\tilde{Q}^{-1} = DD^T$  be a decomposition (such as Cholesky) of  $\tilde{Q}^{-1} > 0$ , we have  $\text{Tr}(W \tilde{Q}^{-1} W) = \|WD\|_F^2$ , from which (5.6) can be derived easily; see Section 3.3.2 for the details. With this and following our framework from Section 5.2, we obtain the Lagrangian function

$$G^\mu(W|\tilde{W}) = \sum_f G(w_f|\tilde{w}_f) + \lambda \left( \sum_f \bar{l}(w_f|\tilde{w}_f) + c \right) + \mu^T \sum_f \left( w_f - \frac{1}{F} e \right), \quad (5.31)$$

where  $w_f \in$  denotes the  $f$ -th row of  $W$ ,  $G$  is given by (3.7),  $\bar{l}$  by (3.9) and derived as explained above, and  $c$  is a constant. Let  $\mu$  is the vector Lagrange multipliers of dimension  $K$  associated to each linear constraint  $e^T w_i = 1$ .

Exactly as before (hence we omit the details here),  $G^\mu$  is separable, let us consider one specific row  $w \in \mathbb{R}^{K \times 1}$  of  $W$  and rewrite (5.31) for  $w$  as follows:

$$G^\mu(w|\tilde{w}) = G(w|\tilde{w}) + \lambda \bar{l}(w|\tilde{w}) + c + \bar{\mu}^T (w - e/F) \quad (5.32)$$

For  $\beta = 1$ , the derivative of  $G^\mu(w|\tilde{w})$  (5.32) w.r.t. a specific entry  $w_k$  is given by:

$$\begin{aligned} \nabla_{w_k} G^\mu(w|\tilde{w}) &= \sum_n h_{kn} - \sum_n h_{kn} \frac{\tilde{w}_k v_n}{w_k \tilde{v}_n} + 2\lambda [Y \tilde{w}]_k + 2\lambda \left[ \text{diag} \left( \frac{Y^+ \tilde{w} + Y^- \tilde{w}}{\tilde{w}} \right) \right]_k w_k \\ &\quad - 2\lambda \left[ \text{diag} \left( \frac{Y^+ \tilde{w} + Y^- \tilde{w}}{\tilde{w}} \right) \right]_k \tilde{w}_k + \mu_k. \end{aligned}$$

Due to the separability, canceling the derivative provides the following closed-form solution:

$$w_k^*(\mu_k) = \tilde{w}_k \frac{\sqrt{(\sum_n h_{kn} - 4\lambda [Y^- \tilde{w}]_k + \mu_k)^2 + 8\lambda [\text{diag}(Y^+ \tilde{w} + Y^- \tilde{w})]_k \sum_n h_{kn} \frac{v_n}{\tilde{v}_n} - \sum_n h_{kn} + 4\lambda [Y^- \tilde{w}]_k - \mu_k}}{4\lambda [\text{diag}(Y^+ \tilde{w} + Y^- \tilde{w})]_k} \quad (5.33)$$

which is non-negative. For clarity purpose, let us pose  $C = e_{F,N} H^T - 4\lambda (\tilde{W} Y^-)$ ,  $S = 8\lambda \tilde{W} (Y^+ + Y^-) \odot \left( \frac{[V]}{[\tilde{W}H]} H^T \right)$ ,  $D = 4\lambda \tilde{W} (Y^+ + Y^-)$ , then Equation (5.33) can be generalized in the following matrix form

$$W^*(\mu) = \tilde{W} \odot \frac{\left[ [C + e\mu^T]^2 + S \right]^{\frac{1}{2}} - (C + e\mu^T)}{[D]}, \quad (5.34)$$

where  $C = e_{F,N} H^T - 4\lambda (\tilde{W} Y^-)$ ,  $D = 4\lambda \tilde{W} (Y^+ + Y^-)$ , and  $S = 8\lambda \tilde{W} (Y^+ + Y^-) \odot \left( \frac{[V]}{[\tilde{W}H]} H^T \right)$  with  $Y = Y^+ - Y^- = (\tilde{W}^T \tilde{W} + \delta I)^{-1}$ ,  $Y^+ = \max(Y, 0) \geq 0$  and  $Y^- = \max(-Y, 0) \geq 0$ , and  $e_{F,N}$  is the  $F$ -by- $N$  matrix of all ones. As proved in Proposition 2, the constraint  $W^*(\mu)^T e = e$  is satisfied for a unique  $\mu$  in  $(-\infty, t)$  where  $t = \infty$  in this case. We can therefore use a Newton-Raphson method to find the  $\mu_i$  with quadratic rate of convergence, see Proposition 3. Algorithm 6 summarizes our method to tackle (3.1).

## 5.4.2 Numerical experiments

In this section we compare baseline KL-NMF (that is, the standard MU), Algorithm 3 from Chapter 3 that uses line search, and Algorithm 6 applied to the spectrogram of two monophonic piano sequences considered in Chapter 3. The first audio sample is the first measure of "Mary had a little lamb", a popular English song. The second audio sample corresponds to the first 30 seconds of "Prelude and Fugue No.1 in C major" from de Jean-Sebastien Bach played by Glenn Gould<sup>3</sup>.

We use the following three setups:

<sup>3</sup><https://www.youtube.com/watch?v=ZlbK5r5mBH4>

---

**Algorithm 6** Min-vol KL-NMF

---

**Require:** A matrix  $V \in \mathbb{R}^{F \times N}$ , an initialization  $H \in \mathbb{R}_+^{K \times N}$ , an initialization  $W \in \mathbb{R}^{F \times K}$ , a factorization rank  $K$ , and a maximum number of iterations, maxiter, the parameters  $\delta > 0$  and  $\lambda > 0$ .

**Ensure:** A rank- $K$  NMF  $(W, H)$  of  $V$  satisfying constraints in (3.1).

```

1: for  $it = 1 : \text{maxiter}$  do
2:     % Update of matrix  $H$ 
3:      $H \leftarrow H \odot \frac{W^T \left( \frac{[V]}{[WH]} \right)}{[W^T e_{F,N}]}$ 
4:     % Update of matrix  $W$ 
5:      $Y \leftarrow (W^T W + \delta I)^{-1}$ 
6:      $Y^+ \leftarrow \max(Y, 0)$ 
7:      $Y^- \leftarrow \max(-Y, 0)$ 
8:      $C \leftarrow e_{F,N} H^T - 4\lambda (W Y^-)$ 
9:      $S \leftarrow 8\lambda W (Y^+ + Y^-) \odot \left( \frac{[V]}{[WH]} H^T \right)$ 
10:     $D \leftarrow 4\lambda W (Y^+ + Y^-)$ 
11:     $\mu \leftarrow \text{root} (W^*(\mu)^T e = e) \text{ over } \mathbb{R}^K$  % see (5.34) for the expression of  $W^*(\mu)$ 
12:     $W \leftarrow W \odot \frac{\left[ \left[ [C + e\mu^T]^2 + S \right]^{\frac{1}{2}} - (C + e\mu^T) \right]}{[D]}$ 
13: end for

```

---

- Setup #1: sample "Mary had a little lamb" with  $K = 3$ , 200 iterations.
- Setup #2: sample "Mary had a little lamb" with  $K = 7$ , 200 iterations.
- Setup #3: "Prelude and Fugue No.1 in C major" with  $K = 16$ , 300 iterations.

For each setup, the algorithms are run for the same 20 random initializations of  $W$  and  $H$ . Table 5.3 reports the average and standard deviation of the runtime (in seconds) over these 20 runs. Table 5.4 reports the average and standard deviation of the final values for  $\beta$ -divergences (data fitting term) and the objective function of (3.1) over these 20 runs for Algorithm 3 and Algorithm 6. For this last comparison, the value for the penalty weight  $\lambda$  has been chosen so that KL-NMF leads to reasonable solutions for  $W$  and  $H$ . More precisely, the values for  $\lambda$  are chosen so that the initial value of  $\frac{\lambda \left| \log \det (W^{(0)T} W^{(0)} + \delta I) \right|}{D_\beta(V|WH)}$  is equal to 0.1, 0.1 and 0.022 for setup #1, setup #2 and setup #, respectively. The algorithms are implemented and tested on a desktop computer with Intel Core i7-8700@3.2GHz CPU and 32GB memory. The codes are written in MATLAB R2018a. The code for Algorithm 6 is available from <https://bit.ly/35J8Yth>.

We observe that the runtime of Algorithm 6 is close to the baseline KL-NMF algorithm which confirms the negligible cost of the Newton-Raphson steps to compute  $\mu^*$  as discussed

Table 5.3: Runtime performance in seconds of baseline KL-NMF, Algorithm 3 and Algorithm 6. The table reports the average and standard deviation over 20 random initializations.

Algorithms	runtime in seconds		
	setup #1	setup #2	setup #3
baseline KL-NMF	0.53±0.03	0.45±0.02	4.32±0.30
Algorithm 3	3.79±0.13	2.39±0.30	10.19±1.28
Algorithm 6	0.58±0.03	0.66±0.03	4.80± 0.38

Table 5.4: Final values for  $D_\beta$  and the penalized objective  $\Psi$  from (3.1) obtained with Algorithm 3 and Algorithm 6. The table reports the average and standard deviation over 20 random initializations for three experimental setups.

		Algorithm 3	Algorithm 6
setup #1	$D_{\beta,\text{end}}$	$(3.52 \pm 0.03)10^3$	<b>(2.31 ± 0.01)</b> $10^3$
	$\Psi_{\text{end}}$	$(4.17 \pm 0.03)10^3$	<b>(3.08 ± 0.01)</b> $10^3$
setup #2	$D_{\beta,\text{end}}$	$(3.54 \pm 0.03)10^3$	<b>(1.77 ± 0.02)</b> $10^3$
	$\Psi_{\text{end}}$	$(4.42 \pm 0.04)10^3$	<b>(2.87 ± 0.02)</b> $10^3$
setup #3	$D_{\beta,\text{end}}$	$(7.77 \pm 0.23)10^3$	<b>(4.67 ± 0.08)</b> $10^3$
	$\Psi_{\text{end}}$	$(9.14 \pm 0.20)10^3$	<b>(6.50 ± 0.06)</b> $10^3$

in Section 5.2.3. On the other hand, since no line search is needed, we have a drastic acceleration from 2x to 7x compared to the backtracking line-search procedure integrated in Algorithm 3. Moreover, we observe in Table 5.4 that Algorithm 6 outperforms Algorithm 3 in terms of final values for the data fitting term and objective function values with lower standard deviations.

## 5.5 Extension to quadratic disjoint constraints

Our general framework presented in Section 5.2 applies to penalized  $\beta$ -NMF with concave or  $L$ -smooth penalties and under disjoint linear equality constraints; see problem (5.2). We have showcased our approach on  $\beta$ -SSNMF in Section 5.3 and on min-vol KL-NMF under sum-to-one constraints on the columns of  $W$  in Section 5.4. In this section, we show that the same framework can be extended to other simple constraints, namely disjoint quadratic constraints. We consider sparse  $\beta$ -NMF for  $\beta = 1$  where the rows of  $H$  are penalized with the  $\ell_1$  norm and each column of  $W$  have a fixed  $\ell_2$  norm. We show that MU satisfying the set of constraints can be derived which we apply on BHU.

### 5.5.1 Problem formulation and algorithm

In this section we consider the following optimization problem involving quadratic disjoints constraints, that we refer to as hyperspheric-structured sparse  $\beta$ -NMF:

$$\min_{W \in \mathbb{R}_+^{F \times K}, H \in \mathbb{R}_+^{K \times N}} D_\beta(V|WH) + \sum_{k=1}^K \lambda_k \|H(k, :)\|_1 \quad \text{such that} \quad e^T w_j^{(2)} = \rho \text{ for } 1 \leq i \leq K, \quad (5.35)$$

where  $\lambda_k$  is a penalty weight to control the sparsity of the  $k$ -th row of  $H$ , and the quadratic constraints require the columns of  $W$  to lie on the surface of a hyper-sphere centered at the origin with radius  $\sqrt{\rho} > 0$ . Without this normalization, the  $\ell_1$ -norm regularization would make  $H$  tends to zero and  $W$  grows to infinity.

As done before, we update  $W$  and  $H$  alternatively. We tackle the subproblem in  $H$  with  $W$  fixed based on the MU developed in [43] and guaranteed to decrease the objective function:

$$H^* = \tilde{H} \odot \frac{\left[ W^T \left( V \odot [W\tilde{H}]^{(\beta-2)} \right) \right]}{\left[ W^T [W\tilde{H}]^{(\beta-1)} + \lambda e^T \right]}, \quad (5.36)$$

where  $\lambda \in \mathbb{R}_+^K$  is the vector of penalty weights. It remains to compute an update for  $W$ . To do so, we use the convex separable auxiliary function  $G$  from [45] constructed at the current iterate  $\tilde{W}$ , from which we obtain, as before, the Lagrangian function

$$G^\mu(W|\tilde{W}) = \sum_f G(w_f|\tilde{w}_f) + \sum_k \lambda_k \|H(k, :)\|_1 + \mu^T \sum_f \left( w_f^{(2)} - \frac{1}{F} \rho e \right), \quad (5.37)$$

where  $\mu \in \mathbb{R}^K$  is the vector of Lagrange multipliers associated to the constraint  $e^T W^{(2)} = \rho e^T$ . As  $G^\mu$  is separable, let us consider one specific row  $w \in \mathbb{R}^{K \times 1}$  of  $W$  and rewrite (5.37) for  $w$  as follows:

$$G^\mu(w|\tilde{w}) = G(w|\tilde{w}) + \sum_k \lambda_k \|H(k, :)\|_1 + \mu^T \left( w^{(2)} - \frac{1}{F} \rho e \right). \quad (5.38)$$

For  $\beta = 1$ , the derivative of  $G^\mu(w|\tilde{w})$  (5.38) w.r.t. a specific entry  $w_k$  is given by:

$$\nabla_{w_k} G^\mu(w|\tilde{w}) = \sum_n h_{kn} - \sum_n h_{kn} \frac{\tilde{w}_k v_n}{w_k \tilde{v}_n} + 2\mu_k w_k.$$

Due to the separability, canceling the derivative provides the following closed-form solution:

$$w_k^*(\mu) = \frac{\sqrt{(\sum_n h_{kn})^2 + 8\mu_k \sum_n h_{kn} \frac{\tilde{w}_k v_n}{\tilde{v}_n}} - \sum_n h_{kn}}{4\mu_k} \quad (5.39)$$

which is non-negative as soon as  $\mu_k > 0$ . Let, so (5.39) can be written in the following matrix form

$$W^*(\mu) = \frac{\left[ \left[ [C]^2 + 8(e\mu^T) \odot S \right]^{\frac{1}{2}} - C \right]}{[4e\mu^T]}, \quad (5.40)$$



where  $C = e_{F,N}H^T$  and  $S = \widetilde{W} \odot \left( \frac{[V]}{[WH]} H^T \right)$ . Let us now write the expression of the quadratic constraint  $\sum_f (W^*(\mu)_{f,i})^2 - \rho = 0$  for one specific column of  $W$ , say the  $i$ -th:

$$r_i(\mu_i) := \sum_f (W_{f,i}^*(\mu_i))^2 - \rho = \sum_f \left( \frac{\sqrt{(C_{f,i})^2 + 8\mu_i S_{f,i}} - C_{f,i}}{4\mu_i} \right)^2 - \rho = 0. \quad (5.41)$$

Computing the Lagrangian multiplier  $\mu_i$  to satisfy the constraint requires computing the roots of the functions  $r_i(\mu_i)$ . We can show that each  $W_{f,i}^*(\mu_i)$  (5.40) is a monotone decreasing, nonnegative convex function over  $(0, +\infty)$ . Therefore  $\sum_f (W_{f,i}^*(\mu_i))^2$  is also monotone decreasing and convex in  $\mu_i$  over  $(0, +\infty)$ . Indeed, let  $g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  be a monotone decreasing, nonnegative convex function. If  $g$  is twice-differentiable, then  $(g^2)'' = 2(g')^2 + 2gg'' \geq 0$  since  $g, g'' \geq 0$  and  $(g^2)' = 2g'g \leq 0$  since  $g \geq 0, g' \leq 0$  by hypothesis. Now we can conclude that  $r_i(\mu_i)$  is a monotone decreasing convex function over  $(0, +\infty)$ . Moreover, using Hospital's rule, we have:

$$\lim_{\mu_i \rightarrow 0^+} \sum_f (W_{f,i}^*(\mu_i))^2 - \rho = +\infty \quad \text{and} \quad \lim_{\mu_i \rightarrow +\infty} \sum_f (W_{f,i}^*(\mu_i))^2 - \rho = -\rho < 0,$$

since  $\rho > 0$ . Therefore, the root of  $r_i(\mu_i)$  is unique over  $(0, +\infty)$ . We use a Newton-Raphson method to solve the problem. Algorithm 7 summarizes our method.

---

**Algorithm 7** Hyperspheric-structured sparse KL-NMF
 

---

**Require:** A matrix  $V \in \mathbb{R}^{F \times N}$ , an initialization  $H \in \mathbb{R}_+^{K \times N}$ , an initialization  $W \in \mathbb{R}^{F \times K}$ , a factorization rank  $K$ , a maximum number of iterations, maxiter, a weight vector  $\lambda > 0$ .

**Ensure:** A sparse rank- $K$  NMF  $(W, H)$  of  $V$  satisfying constraints in (5.35).

```

1: for  $it = 1$  : maxiter do
2:   % Update of matrix  $H$ 
3:    $H \leftarrow H \odot \frac{W^T (V \odot [WH]^{(\beta-2)})}{W^T [WH]^{(\beta-1)} + \lambda e^T}$ 
4:   % Update of matrix  $W$ 
5:    $C \leftarrow e_{F,N} H^T$ 
6:    $S \leftarrow W \odot \left( \frac{[V]}{[WH]} H^T \right)$ 
7:   for  $j = 1$  :  $K$  do
8:      $\mu_i \leftarrow \text{root}(r_i(\mu_i))$  over  $(0, +\infty)$  % See Equation (5.41)
9:   end for
10:   $W \leftarrow \frac{[C]^2 + 8(e\mu^T) \odot S]^{1/2} - C}{4\mu e^T}$ 
11: end for
    
```

---

### 5.5.2 Numerical experiments

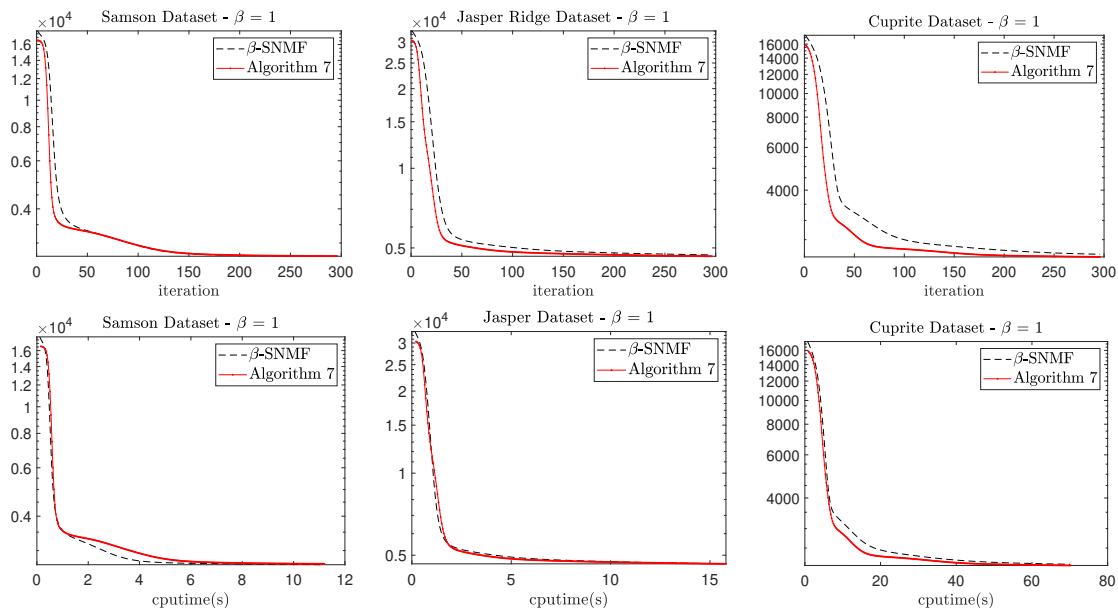
In this section, we perform numerical experiments to evaluate the effectiveness of Algorithm 7 on the HU problem. To the best of our knowledge, sparse  $\beta$ -NMF<sup>4</sup> from [119] is the most recent algorithm that is able to tackle problem (5.35) for the KL-divergence by integrating the  $\ell_2$ -normalization for each update of matrix  $W$ . This approach is similar to that of [44] for  $\beta$ -SSNMF, that is, it uses parametrization, and resort to a heuristic with no guarantee on the decrease of the objective function. We refer to this algorithm as  $\beta$ -SNMF. We apply Algorithm 7 and  $\beta$ -SNMF [119] to the three real hyperspectral datasets detailed in Section 5.3. This comparative study focuses on the convergence aspects including the evolution of the objective function and the runtime. All the algorithms are implemented and tested on a desktop computer with Intel Core i7-8700@3.2GHz CPU and 32GB memory. The codes are written in MATLAB R2018a. The code for Algorithm 7 is available from <https://bit.ly/35MWUYb>. For all simulations, the algorithms are ran for 20 random initializations of  $W$  and  $H$ , the entries of the penalty weight  $\lambda$  has been set to 0.1, 0.05 and 0.05 for Samson, Jasper Ridge and Cuprite data sets, respectively. In order to fairly compare both algorithms,  $\rho$  has been set to 1 as  $\beta$ -SNMF considers a  $\ell_2$ -normalization for the columns of  $W$ , and the entries of the weight vector  $\lambda$  in Algorithm 7 have the same values as  $\beta$ -SNMF requires to use the same values for all rows of  $H$ . Table 5.5 reports the average and standard deviation of the runtime (in seconds) as the final value for the objective function over these 20 runs for a maximum of 300 iterations. Figure 5.2 displays the objective function values.

Table 5.5: Runtime performance in seconds and final value of objective function  $\Phi_{\text{end}}(W, H)$  for Algorithm 7 and  $\beta$ -SNMF. The table reports the average and standard deviation over 20 random initializations with a maximum of 300 iterations for three hyperspectral data sets.

Algorithms	Samson data set		Jasper Ridge data set		Cuprite data set	
	runtime (sec)	$\Phi_{\text{end}}(W, H)$	runtime (sec)	$\Phi_{\text{end}}(W, H)$	runtime (sec)	$\Phi_{\text{end}}(W, H)$
Algorithm 7	11.07±0.19	(2.68±0.00)10 <sup>3</sup>	15.67±0.17	( <b>4.65 ± 0.00</b> )10 <sup>3</sup>	70.16 ± 0.85	( <b>2.12 ± 0.00</b> )10 <sup>3</sup>
$\beta$ -SNMF [119]	<b>7.63±0.13</b>	(2.68±0.00)10 <sup>3</sup>	<b>10.98±0.18</b>	(4.71 ± 0.00)10 <sup>3</sup>	<b>51.86 ± 0.74</b>	(2.18 ± 0.00)10 <sup>3</sup>

According to Table 5.5 (top row), we observe that Algorithm 7 outperforms the heuristic from [119] in terms of final value for the objective functions while  $\beta$ -SNMF shows lower runtimes. Additionally, based on Figure 5.2, we observe that Algorithm 7 converges on average faster than  $\beta$ -SNMF for all the data sets, in terms of iterations. However,  $\beta$ -SNMF has a lower computational cost per iteration. Thus, we complete the comparison between both algorithms by imposing the same computational time: we run Algorithm 7 for 300 iterations, record the computational time and run  $\beta$ -SNMF for the same amount of time. Table 5.6 reports the average and standard deviation of the final value for the objective function over 20 runs in this setting. Figure 5.2 (bottom row) displays the

<sup>4</sup><http://www.jonathanleroux.org/software/sparseNMF.zip>



**Fig. 5.2.** Averaged objective functions over 20 random initializations obtained for Algorithm 7 with 300 iterations (red line with circle markers), and the heuristic  $\beta$ -SNMF from [119] (black dashed line).

Table 5.6: Final value of objective function values  $\Phi_{\text{end}}(W, H)$  for Algorithm 7 and the heuristic from [119]. The table reports the average and standard deviation over 20 random initializations for an equal computational time that corresponds to 300 iterations of Algorithm 7.

Algorithms	Samson data set	Jasper Ridge data set	Cuprite data set
	$\Phi_{\text{end}}(W, H)$	$\Phi_{\text{end}}(W, H)$	$\Phi_{\text{end}}(W, H)$
Algorithm 7	$(2.68 \pm 0.00)10^3$	$(\mathbf{4.65} \pm \mathbf{0.00})10^3$	$(\mathbf{2.12} \pm \mathbf{0.00})10^3$
$\beta$ -SNMF [119]	$(2.68 \pm 0.00)10^3$	$(4.66 \pm 0.00)10^3$	$(2.15 \pm 0.00)10^3$

objective function w.r.t. time for the three data sets. On this comparison, Algorithm 7 and the heuristic from [119] perform similarly although Algorithm 7 has slightly better final objective function values. However, keep in mind that only Algorithm 7 is theoretically guaranteed to decrease the objective function. For all simulations, the algorithms are ran for 20 random initializations of  $W$  and  $H$ , the entries of the penalty weight  $\lambda$  has been set to 0.1, 0.05 and 0.05 for Samson, Jasper Ridge and Cuprite data sets, respectively. In order to fairly compare both algorithms,  $\rho$  has been set to 1 as  $\beta$ -SNMF considers a  $\ell_2$ -normalization for the columns of  $W$ , and the weight vector  $\lambda$  in Algorithm 7 have the same values as  $\beta$ -SNMF requires to use the same values for all rows of  $H$ .

In the following we report qualitative results obtained with Algorithm 7 applied to three HS real data sets, that are Samson, Jasper and Urban data sets. The first two data sets are detailed in Section 5.3. The Urban data set contains 162 spectral bands with  $307 \times 307$  pixels with mostly six endmembers. Note that Cuprite data set is replaced by the Urban data set

since endmembers for Cuprite correspond to chemical components which are more difficult to interpret visually while endmembers for Urban data sets are more easily interpretable.

As mentioned earlier,  $\lambda_k$  enables to control sparsity of the  $k$ -th row of  $H$ . Given a row  $H(k, :) \in \mathbb{R}_+^N$  of  $H$ , a meaningful way to measure its sparsity is to consider the following measure [72]:

$$\text{sp}(H(k, :)) = \frac{\sqrt{N} - \frac{\|H(k, :)\|_1}{\|H(k, :)\|_2}}{\sqrt{N} - 1} \in [0, 1]. \quad (5.42)$$

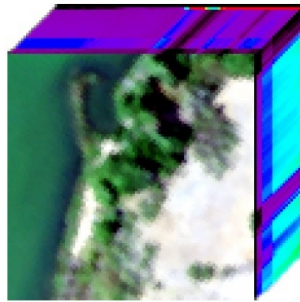
During the numerical experiments, we observed that Algorithm 7 gives better results when the initial values for  $\lambda$  are low and progressively increased. During a specified interval of iterations  $[it_{\min}, it_{\max}]$ , the sparsity of the current iterate is measured by using equation (5.42), and the entries of  $\lambda$  are dynamically updated (increased with a rate  $\alpha > 1$ ) to achieve a desired sparsity level  $sp$ . The dynamic update of the weight vector to reach the desired levels of sparsity has been activated in the iterations intervals  $[1, 150]$ ,  $[1, 150]$  and  $[1, 75]$  for Samson, Jasper and Urban, respectively. We report here the abundance maps of each end-member for two levels of average target sparsity that are 0.25 and 0.5. For all the simulations, the weight vector  $\lambda$  has been initialized to  $0.05e$ , and the algorithm was run for 300 iterations. We fix the number of endmembers to 3, 4 and 6 respectively for Samson, Jasper Ridge and Urban data sets, these values are commonly considered in the HS community [154]. Figures 5.3 to 5.5 picture the abundance estimation for the three data sets for the two levels of sparsity.

In order to validate the results obtained for the abundances of the endmembers, we display in Figures 5.6, 5.7 and 5.8 the ground truth results obtained in [154]. Note that the grayscale used in [154] is the complementary of the one used in Figures 5.3 to 5.5.

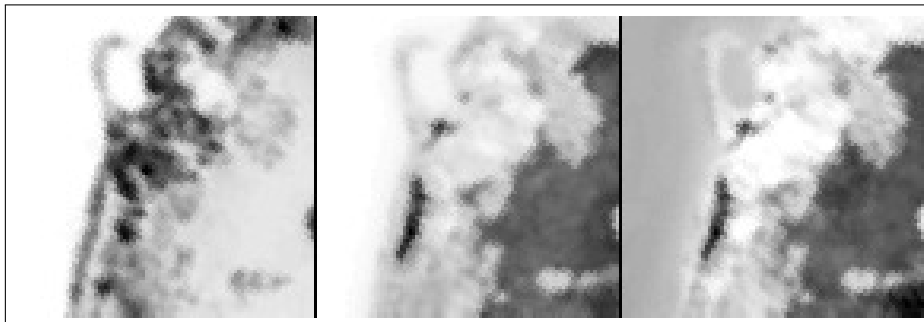
We observe that the abundance estimation gets significantly more accurate when the level of average sparsity is higher. For the Samson and Jasper Ridge data sets, the abundances for the endmembers are nicely estimated while five endmembers over six are well estimated for the Urban data set. The ‘‘Roof’’ is divided into ‘‘Roof1’’ and ‘‘Roof2/shadow’’ [115, 154]. In our simulations, it seems that the sixth endmember corresponds to some shadows with a small residual of ‘‘Grass’’, while the ‘‘Roof’’ is not split into two groups.

## 5.6 Conclusion

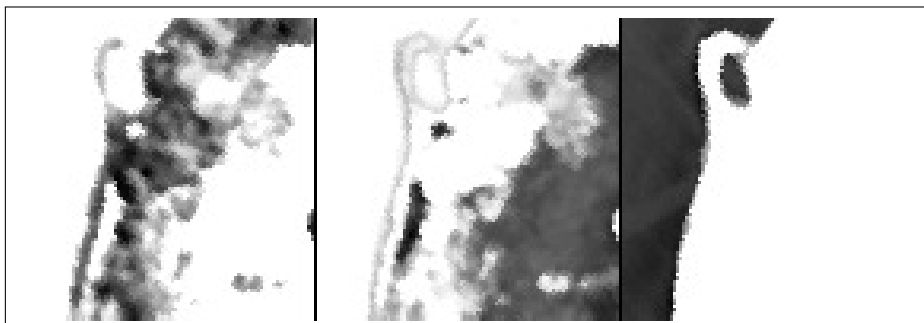
In this chapter we have presented a general framework to solve penalized  $\beta$ -NMF problems that integrates a set of disjoint constraints on the variables; see the general formulation (5.2). Using this framework, we showed that we can derive algorithms that compete favorably with the state of the art for a wide variety of  $\beta$ -NMF problems, such as the simplex-structured NMF and the minimum-volume  $\beta$ -NMF with sum-to-one constraints on the columns of  $W$ . We have also shown how to extend the framework to non-linear disjoint constraints, with application to a sparse  $\beta$ -NMF model for  $\beta = 1$  where each column of  $W$  lie on a hyper-sphere.



(a) Samson Data set

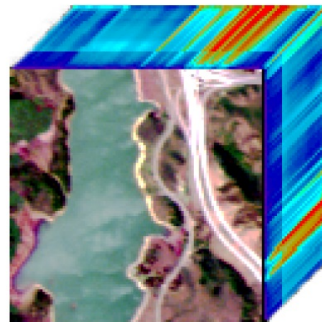


(b) Abundance map with average sparsity level set to 0.25

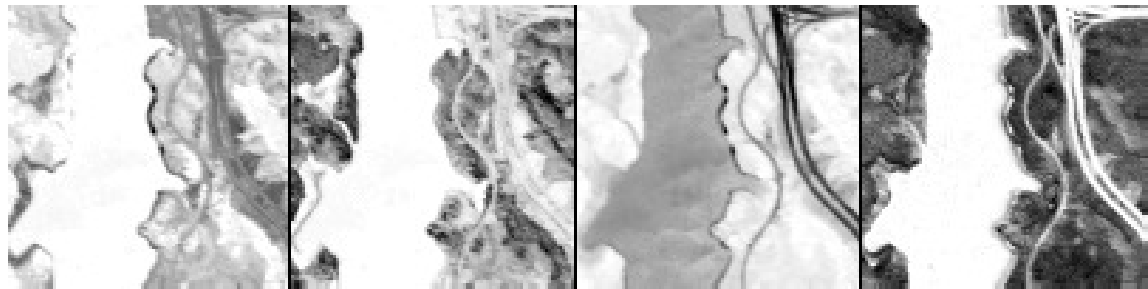


(c) Abundance map with average sparsity level set to 0.5

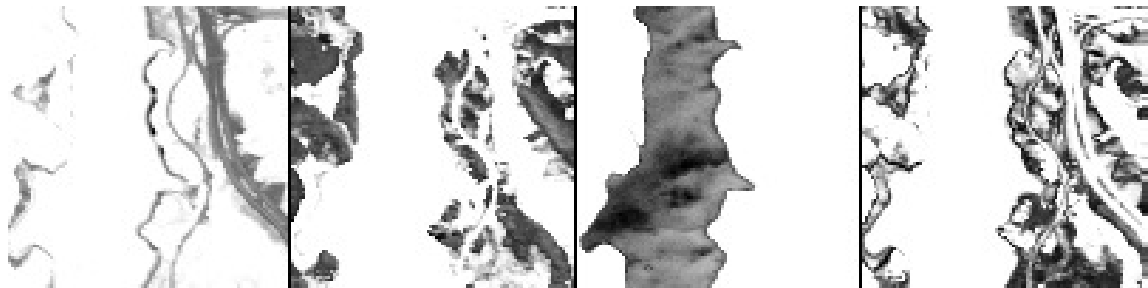
**Fig. 5.3.** Samson data set (a) and results ((b) and (c)) for the Abundance maps estimated using Algorithm 7 for the three endmembers: #1 Tree, #2 Soil and #3 Water. Two average sparsity levels considered: 0.25 (b) and 0.5 (c).



(a) Jasper Ridge Data set

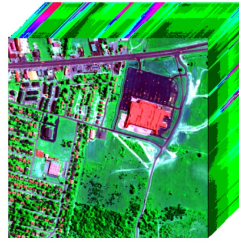


(b) Abundance map with average sparsity level set to 0.25



(c) Abundance map with average sparsity level set to 0.5

**Fig. 5.4.** Jasper Ridge data set (a) and results ((b) and (c)) for the Abundance maps estimated using Algorithm 7 for the four endmembers: #1 Road, #2 Tree, #3 Water and #4 Soil. Two average sparsity levels are considered: 0.25 (b) and 0.5 (c).



(a) Urban Data set

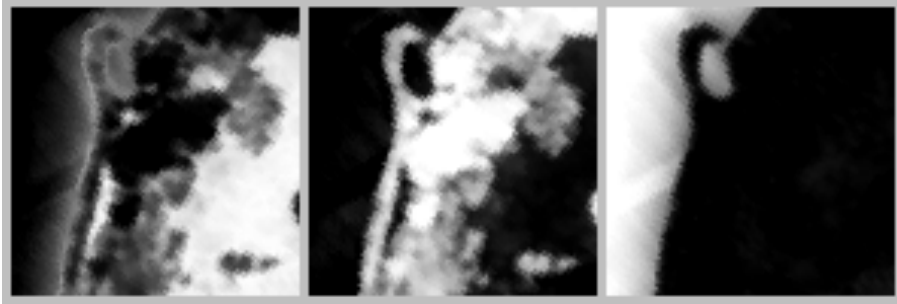


(b) Abundance map with average sparsity level set to 0.25

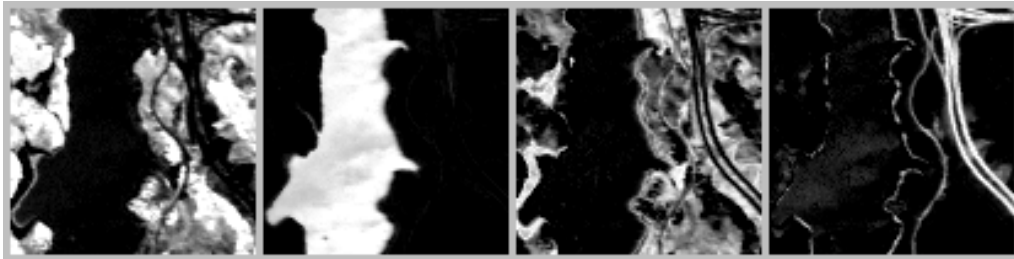


(c) Abundance map with average sparsity level set to 0.5

**Fig. 5.5.** Urban data set (a) and results ((b) and (c)) for the Abundance maps estimated using Algorithm 7 for the six endmembers: #1 Soil, #2 Tree, #3 Grass, #4 Roof, #5 Road/Asphalt and #6 Roof2/shadows. Two average sparsity levels are considered: 0.25 (b) and 0.5 (c).



**Fig. 5.6.** Baseline abundances for the endmembers obtained for Samson data extracted from [154]: #1 Soil, #2 Tree and #3 Water.



**Fig. 5.7.** Baseline abundances for the endmembers obtained for Jasper Ridge data extracted from [154]: #1 Road, #2 Soil, #3 Water and #4 Tree.



**Fig. 5.8.** Baseline abundances for the endmembers obtained for Urban data extracted from [154]: #1 Asphalt, #2 Grass, #3 Tree, #4 Roof1, #5 Roof2/Shadow and #6 Soil.

Further works will focus on the possible extension of the methods to non-disjoint constraints. The non-disjoint constraints will lead to roots finding problems of polynomial equations in the Lagrangian multipliers for which we hope to find conditions that ensure the uniqueness of the solution.

Another interesting direction of research would be to apply our framework to other NMF models. For example, in probabilistic latent semantic analysis/indexing (PLSA/PLSI), the model is the following: given a nonnegative matrix  $V$  such that  $e^T V e = 1$  (this can be assumed w.l.o.g. by dividing the input matrix by  $e^T V e$ ), solve

$$\max_{W \geq 0, H \geq 0, s \geq 0} \sum_{i,j} V_{i,j} \log(W \text{diag}(s)H)_{i,j} \text{ such that } W^T e = e, H e = e, s^T e = 1.$$

This model is equivalent to KL-NMF [38], with the additional constraint that  $e^T W H e =$



$e^T X e$ , and hence our framework is applicable to PLSA/PLSI. Such constraints have also applications in soft clustering contexts; see [144].

## 6 Exact NMF with conic programming

In this chapter, we introduce a novel framework that includes two approaches for computing an exact NMF. Exact NMF can be defined as follows: given an input nonnegative matrix  $V \in \mathbb{R}_+^{F \times K}$ , we search for two nonnegative matrices  $W \in \mathbb{R}_+^{F \times K}$  and  $H \in \mathbb{R}_+^{K \times N}$  such that  $V = WH$ . Each of the proposed approaches relies on the construction and the resolution of a specific optimization problem. For each optimization problem we introduce a particular change of variables that enables the use of two special cases of conic constraints, that are the exponential and second-order conic constraints.

In order to solve the two optimization problems, we propose a general algorithm with two key ingredients:

1. the original optimization problems are replaced by a sequence of easier optimization problems to solve, each successive problem being obtained by majorizing the objective functions by their linearization constructed at the current solution  $(W, H)$ . By doing so, we show that the successive approximated problems belong to two special cases of conic programming; namely the exponential programming and the second-order programming.
2. Interior-point methods are used to solve each successive problem with high accuracy.

We show that our algorithm once applied to tackle the two optimization problems is able to compute exact NMF for several classes of nonnegative matrices (namely, randomly generated, infinite rigid matrices and nested hexagons problem matrices) and as such demonstrate their competitiveness compared to recent methods from the literature. We finally show that the first approach relying on exponential programming is competitive compared to recent method for solving the so-called maximum-edge biclique problem for small size input matrices.

The following sections present some preliminaries required before we formally introduce the two formulations for computing an exact NMF. In particular, we present the main properties of the nonnegative rank of a matrix, the notion of conic programming and the special cases of cones we consider in this chapter.

### 6.1 Introduction and preliminaries

As introduced in Section 1.3, computing an NMF corresponds to finding good approximations of a given nonnegative matrix as a low-rank product of two nonnegative matrices.

Despite the fact that NMF is NP-hard in general as explained in Section 1.7, it has been used successfully in many practical situations, see Section 1.5. Many local optimization schemes have been developed to compute good factorizations and therefore try to identifying good local minima of the optimization problems associated to NMF models, see section 1.4. Most of the algorithms are based on iterative schemes such that at each iteration, they aim to improve the current solution. On a practical point of view, many state-of-the-art algorithms rely on a two-BCD scheme, see section 1.10 for more details. Comparatively, less attention has been given in the literature to the development of algorithms aimed at finding global minima of the optimization problems associated to NMF models. In this chapter, we are interested in computing high quality local minima for the NMF optimization problems without relying on the BCD framework; the optimization over  $W, H$  is performed jointly. In particular, our focus is on finding exact NMFs, that is, computing nonnegative factors  $W$  and  $H$  such that  $V = WH$  holds exactly. The minimum factorization rank for which such an exact NMF exists is called the nonnegative rank of  $V$  and is denoted  $\text{rank}_+(V)$ . In the following section, we briefly present some properties of the nonnegative rank.

### 6.1.1 Some properties of the nonnegative rank

Here-under we briefly present some properties of the nonnegative rank, most of these results presented here-under are extracted from the seminal paper by Cohen and Rothblum [31]. First, an upper bound and a lower bound for  $\text{rank}_+(V)$  can be easily computed.

**Lemma 6.1.1.** *Let  $V \in \mathbb{R}_+^{F \times N}$ , then*

$$\text{rank}(V) \leq \text{rank}_+(V) \leq \min(F, N)$$

*Proof.* The first inequality holds since it is not possible to find an exact factorization of lower rank than  $V$ . Indeed, the rank of  $V$  is the smaller integer  $K$  such that we can find two matrices  $W \in \mathbb{R}^{F \times K}$  and  $H \in \mathbb{R}^{K \times N}$ . The nonnegative rank is defined in the way with the requirement for  $W$  and  $H$  to be componentwise nonnegative. The second comes from one of the trivial factorizations  $I_F V$  or  $V I_N$ .  $\square$

In some particular cases, the first inequality is tight. For the rank-one nonnegative matrix  $V$ , we can easily find an exact NMF  $V = wh^T$  where  $w \in \mathbb{R}_+^F$  and  $h \in \mathbb{R}_+^N$ . This implies that  $\text{rank}(V) = \text{rank}_+(wh^T) = 1$ . This observation is still true for a rank-two nonnegative matrix, see the proofs in [31].

Let us cite here-under others well-known properties of the nonnegative rank proved in [31], let  $V$  and  $U$  be two nonnegative matrices  $\in \mathbb{R}_+^{F \times K}$  and  $\in \mathbb{R}_+^{K \times N}$  respectively:

- $\text{rank}_+(V^T) = \text{rank}_+(V)$  [31, Lemma 2.5],
- $\text{rank}_+(V + U) \leq \text{rank}_+(V) + \text{rank}_+(U)$  [31, Lemma 2.5],
- $\text{rank}_+(VU) \leq \min(\text{rank}_+(V), \text{rank}_+(U))$  [31, Lemma 2.6]

In the Appendix, we introduce more features of the nonnegative rank of a matrix, in particular we present the impact on the nonnegative rank of a matrix  $V$  under two types of perturbations, namely small continuous perturbations and rank-1 perturbations.

In all the experiments considered in this chapter, the nonnegative rank of the input matrices are known.

### 6.1.2 Conic programming

The key ingredients for CP are:

- a closed, convex cone  $\mathcal{K} \subseteq \mathbb{R}^n$  : if  $x_1, x_2 \in \mathcal{K}$ , then  $t_1x_1 + t_2x_2 \in \mathcal{K}$  for any  $t_1, t_2 \geq 0$ ,
- a linear operator  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , two vectors  $b \in \mathbb{R}^m$ ,  $c \in \mathbb{R}^n$  and a scalar  $d$ ,
- an inner product  $\langle \cdot, \cdot \rangle$  on  $\mathbb{R}^n$ .

The general form of a conic optimization problem is as follows:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \langle c, x \rangle + d \\ \text{subject to} \quad & Ax = b, \\ & x \in \mathcal{K}, \end{aligned} \tag{6.1}$$

where the last constraint is commonly referred to as convex conic constraint. Hence a conic optimization problem is an optimization problem in which a linear function is minimized over the intersection of an affine subspace and a convex cone. Thus, CP are convex optimization problems. By choosing a specific cone  $\mathcal{K}$  for the conic constraint, we choose a special case of CP. For instance, Linear programming (LP) is a special case of conic programming (CP) for which  $\mathcal{K} = \mathbb{R}_+^n$ . The important convex cones in optimization and the associated case of CP are detailed in Table 6.1.

Table 6.1: Important convex cones and the associated case of CP

Convex cone $\mathcal{K}$	Conic Programming
$\mathbb{R}_+^n$ (the nonnegative orthant)	Linear Programming (LP)
$\mathbb{S}_+^{n \times n}$ (the cone of Positive Semi-Definite matrices)	Semi-Definite Programming (SDP)
$\mathcal{K}_{exp}$ , the exponential cone	Conic Geometric Programming (CGP)
$\mathcal{Q}^n$ , the quadratic (ice cream) cone	Second-Order Conic Programming (SOCP)

In this chapter, we consider the exponential cones and the second-order cones which are discussed in the two following sections.

### Exponential cones

In the case  $n = 3$ , the (primal) exponential cone, denoted  $K_{exp}$ , is a convex subset of  $\mathbb{R}^3$  defined as follows:

$$K_{exp} = \left\{ (x, y, z) \mid x \geq y e^{\frac{z}{y}}, y > 0 \right\} \cup \left\{ (x, 0, z) \mid x \leq 0, z \geq 0 \right\}. \quad (6.2)$$

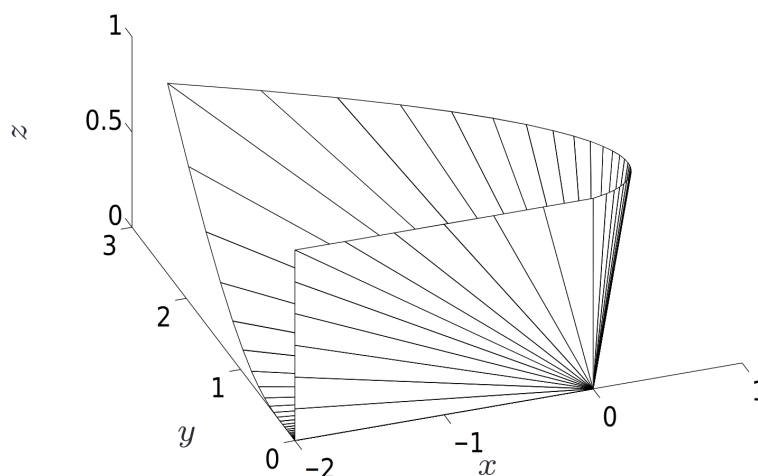
Thus the exponential cone is the closure in  $\mathbb{R}^3$  of the epigraph of  $e^x$  (non-proper cone). We can easily see that  $\mathcal{K}_{exp}$  defined by (6.2) is a convex cone since:

- for all  $x \in \mathcal{K}_{exp}$ , then  $\alpha x \in \mathcal{K}_{exp}$  for all  $\alpha \geq 0$ ,
- the convexity of  $\mathcal{K}_{exp}$  follows from the fact that the Hessian of  $f(z, y) = y \exp z/y$ , namely

$$\nabla^2 f = e^{\frac{z}{y}} \begin{pmatrix} y^{-1} & -zy^{-2} \\ -zy^{-2} & z^2 y^{-3} \end{pmatrix},$$

is positive semidefinite for  $y > 0$ .

Figure 6.1 displays the boundary of the exponential cone  $K_{exp}$  when  $n = 3$ :



**Fig. 6.1.** Boundary of the exponential cone  $\mathcal{K}_{exp}$  in the case  $n = 3$ .

The use of the exponential cone in CP leads to new types of constraint building blocks and new types of representable sets. We list here-under some useful modeling examples using the exponential cone:

- Exponential: the epigraph  $t \geq e^x$  is a section of  $K_{exp}$ , indeed:

$$t \geq e^x \iff (t, 1, x) \in K_{exp}.$$

- Log-sum-exp: the log-sum-exp (logarithm of sum of exponentials) expression  $t \geq \log(e^{x_1} + \dots + e^{x_n})$  is equivalent to the inequality  $e^{x_1-t} + \dots + e^{x_n-t} \leq 1$  and therefore

can be modeled as follows:

$$\begin{aligned} \sum_{i=1}^n \tau_i &\leq 1, \\ (\tau_i, 1, x_i - t) &\in K_{exp} \text{ for } i = 1, \dots, n \end{aligned} \tag{6.3}$$

These examples of modeling will be useful in Section 6.2.

### Second-order cones

The  $n$ -dimensional quadratic cone, denoted  $\mathcal{Q}^n$ , is defined as follows

$$\mathcal{Q}^n = \left\{ x \in \mathbb{R}^n \mid x_1 \geq \sqrt{x_2^2 + \dots + x_n^2} \right\}.$$

Further, we will use a variant of this quadratic cone which is usually referred to as the rotated quadratic cone. Mathematically, a  $n$ -dimensional rotated quadratic cone is defined as:

$$\mathcal{Q}_r^n = \{x \in \mathbb{R}^n \mid 2x_1x_2 \geq x_3^2 + \dots + x_n^2\}.$$

We pass from a quadratic to a rotated quadratic cone with an orthogonal transformation. Indeed, let us define the following orthogonal linear operator:

$$T_n = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & I_{n-2} \end{pmatrix}.$$

In the case  $n = 3$ , we easily verify that :

$$x \in \mathcal{Q}^3 \iff z = T_n x \in \mathcal{Q}_r^3.$$

Indeed,

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & I_{n-2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2}(x_1 + x_2) \\ 1/\sqrt{2}(x_1 - x_2) \\ x_3 \end{pmatrix},$$

therefore

$$2z_1z_2 \geq z_3^2, z_1, z_2 \geq 0 \implies (x_1^2 - x_2^2) \geq x_3^2, x_1 \geq 0.$$

The orthogonal transformation then corresponds to a rotation of  $\pi/4$  around axis  $x_3$ . As an illustration, the boundary of the 3-dimensional quadratic and the rotated 3-dimensional cone is depicted Figure 6.2.

Hence, one could argue that we only need vanilla second-order cone  $\mathcal{Q}^n$ , however there are many practical situations where it is more natural to use rotated second-order cones. For instance, some power-like inequalities can be modeled using rotated second-order cones:

$$|t| \leq \sqrt{x}, x \geq 0 \iff (x, 1/2, t) \in \mathcal{Q}_r^3.$$

We have now everything in hand to introduce in the next section the two optimization problems that we will try to solve for computing exact NMFs.

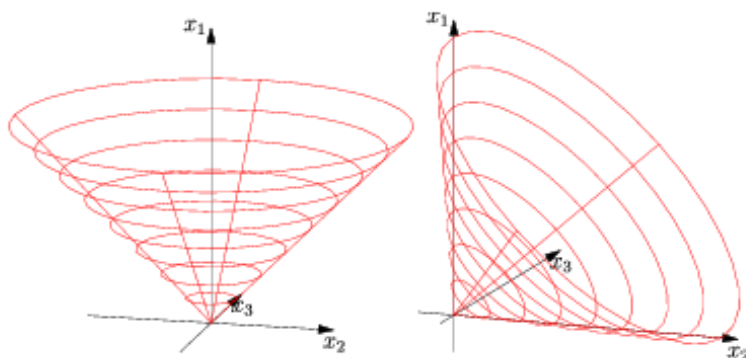


Fig. 6.2. Boundaries of  $Q^3$  and  $Q_r^3$ , reproduced from MOSEK doc.

## 6.2 Problem formulations for exact NMF

In this section we propose two optimization problems associated to the exact NMF model  $V = WH$ ,  $(W, H) \geq 0$ . Each optimization problem integrates a special case of conic constraints.

### 6.2.1 Problem formulation via exponential cones

Given a non-negative matrix  $V \in \mathbb{R}_+^{F \times N}$  and a positive integer  $K \ll \min(F, N)$ , we want to compute an exact NMF. A natural way to compute such a factorization would be to solve the following problem:

$$\begin{aligned}
 & \max_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} \sum_{f,n} \left( \sum_k W_{fk} H_{kn} \right) \\
 & \text{subject to} \quad \sum_k W_{fk} H_{kn} \leq V_{fn} \text{ for all } f, n, \\
 & \quad \quad \quad W_{fk} \geq 0, H_{kn} \geq 0 \text{ for all } f, k, n.
 \end{aligned} \tag{6.4}$$

This formulation could be referred as an under-approximation formulation since we maximize  $\sum_k W_{fk} H_{kn}$  to equal  $V_{fn}$  for all  $f, n$ . In order to deal with such nonnegative constraints on the entries of  $W$  and  $H$ , we propose the following change of variables: we pose  $W_{fk} = G(U_{fk}) = e^{U_{fk}}$  and  $H_{kn} = G(V_{kn}) = e^{V_{kn}}$  with  $f = 1, \dots, F$ ,  $n = 1, \dots, N$  and  $k = 1, \dots, K$ . By applying a logarithm on top of this change of variable to the objective function and on both sides of the inequality constraints (it does not change the optimal solutions since a logarithm is a monotone increasing function), optimization problem (6.4) becomes:

$$\begin{aligned}
 & \max_{U \in \mathbb{R}^{F \times K}, V \in \mathbb{R}^{K \times N}} \log \left( \sum_{f,n} \sum_k e^{U_{fk} + V_{kn}} \right) \\
 & \text{subject to} \quad \log \left( \sum_k e^{U_{fk} + V_{kn}} \right) \leq \log(V_{fn}) \text{ for all } f, n,
 \end{aligned} \tag{6.5}$$

which corresponds to the maximization of a convex function (log-sum-exponentials) over a convex set  $Q$ . Indeed, we can easily check that any function  $f(x) = \log(\sum_i^n e^{x_i})$  is convex; by posing  $r_i = e^{x_i}$  for all  $i$ , the Hessian matrix has the following form:

$$\nabla^2 f = \frac{1}{e^T r} \text{diag}(r) - \frac{1}{(e^T r)^2} r r^T.$$

To show that  $\nabla^2 f \geq 0$ , we must verify that  $v^T \nabla^2 f v \geq 0$  for all  $v$ :

$$v^T \nabla^2 f v = \frac{(\sum_i r_i v_i^2)(\sum_i r_i) - (\sum_i r_i v_i)^2}{(\sum_i r_i)^2} \geq 0,$$

since  $(\sum_i r_i v_i)^2 \leq (\sum_i r_i v_i^2)(\sum_i r_i)$  from Cauchy-Schwarz inequality.

The convex set  $Q$  from (6.5) is defined as follows:

$$Q = \left\{ (U, V) \mid \log \left( \sum_k e^{U_{fk} + V_{kn}} \right) \leq \log(V_{fn}), f = 1, \dots, F, n = 1, \dots, N. \right\} \quad (6.6)$$

by using the log-sum-exp reduction from (6.3), we write (6.6) as explicit conic constraints as follows:

$$\begin{aligned} \sum_{k=1}^K t_{fkn} &\leq X_{fn} \text{ for all } f, n, \\ (t_{fkn}, 1, U_{fk} + V_{kn}) &\in K_{exp} \text{ for all } f, k, n. \end{aligned} \quad (6.7)$$

Thus, the optimization problem (6.5) becomes:

$$\begin{aligned} \max_{U \in \mathbb{R}^{F \times K}, V \in \mathbb{R}^{K \times N}} & \log \left( \sum_{fn} \sum_k e^{U_{fk} + V_{kn}} \right) \\ \text{subject to} & \sum_{k=1}^K t_{fkn} \leq X_{fn} \text{ for all } f, n, \\ & (t_{fkn}, 1, U_{fk} + V_{kn}) \in K_{exp} \text{ for all } f, k, n. \end{aligned} \quad (6.8)$$

This leads to  $F \times N$  inequality constraints and the introduction of  $F \times K \times N$  exponential cones. The strategy followed to tackle problem (6.8) is detailed in Section 6.3.

### 6.2.2 Problem formulation via rotated quadratic cones

In this section we present an optimization problem formulation to compute an exact NMF via (rotated) quadratic cones. Let us start by considering the following alternative formulation to (6.4) for computing an exact NMF:

$$\begin{aligned} \min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} & \sum_{f,n} \left( \sum_k W_{fk} H_{kn} \right) \\ \text{subject to} & \sum_k W_{fk} H_{kn} \geq V_{fn} \text{ for all } f, n, \\ & W_{fk}, H_{kn} \geq 0 \text{ for all } f, k, n. \end{aligned} \quad (6.9)$$



This formulation could be referred as an upper-approximation formulation since we minimize  $\sum_k W_{fk}H_{kn}$  to equal  $V_{fn}$  for all  $f, n$ . We consider the following change of variables: we pose  $W_{fk} = G(U_{fk}) = \sqrt{U_{fk}}$  and  $H_{kn} = G(V_{kn}) = \sqrt{V_{kn}}$  with  $f = 1, \dots, F$ ,  $n = 1, \dots, N$  and  $k = 1, \dots, K$ . Thus the optimization problem (6.9) becomes

$$\begin{aligned} & \min_{U \in \mathbb{R}^{F \times K}, V \in \mathbb{R}^{K \times N}} \sum_{fn} \left( \sum_k \sqrt{U_{fk}} \sqrt{V_{kn}} \right) \\ & \text{subject to} \quad \sum_k \sqrt{U_{fk}} \sqrt{V_{kn}} \geq V_{fn} \text{ for all } f, n, \end{aligned} \quad (6.10)$$

which corresponds to the minimization of a concave function over a convex set  $Q$ . We now write  $Q$  as explicit conic constraints as follows:

$$\begin{aligned} & \sum_{k=1}^K t_{fkn} \geq V_{fn}, \\ & (U_{fk}, 1/2V_{kn}, t_{fkn}) \in \mathcal{Q}_r^3 \text{ for all } f, k, n. \end{aligned} \quad (6.11)$$

Thus, the optimization problem (6.10) becomes

$$\begin{aligned} & \min_{U \in \mathbb{R}^{F \times K}, V \in \mathbb{R}^{K \times N}} \sum_{fn} \left( \sum_k \sqrt{U_{fk}} \sqrt{V_{kn}} \right) \\ & \text{subject to} \quad \sum_{k=1}^K t_{fkn} \geq V_{fn} \text{ for all } f, n, \\ & \quad \quad \quad (U_{fk}, 1/2V_{kn}, t_{fkn}) \in \mathcal{Q}_r^3 \text{ for all } f, k, n. \end{aligned} \quad (6.12)$$

which leads to  $F \times N$  inequality constraints and the introduction of  $F \times K \times N$  rotated quadratic cones. In Section 6.3, we present the algorithm developed to tackle (6.8).

### 6.3 Algorithm

In this section we present the methodology followed to tackle both problems (6.8) and (6.12). Let us first observe that both problems correspond to the minimization of a concave function  $\Phi$  over a convex set  $Q$ . Indeed, for problem (6.8), the maximization of the convex function  $g = \log \left( \sum_{fn} \sum_k e^{U_{fk} + V_{kn}} \right)$  is equivalent to minimizing the concave function  $-g = \Phi$ . There are three main building blocks for our proposed algorithm:

1. Initialization for  $U$  and  $V$ ; we chose to randomly initialize  $W$  and  $H$  (uniformly distributed random number) and apply the two changes of variables to compute the initializations for  $U$  and  $V$  for both problems.
2. The main algorithm which is detailed below. For problem (6.12), the main algorithm integrates a procedure that automatically updates the optimization problems in the case some of the entries of the current solutions tend to zero. This procedure is referred to as Sparsity Patterns Integration (SPI) and is detailed in Section 6.3.1.

3. A final refinement step that will try to further improve the output of the main algorithm as far as possible (ideally, until an exact NMF is found); we will use the accelerated HALS algorithm from [60]. The final refinement step will be applied to all solutions generated by the second building block of the algorithm. In this chapter, we use a tolerance for the relative error equal to  $10^{-6}$ , that is we assume that an exact NMF  $(W, H)$  is found for an input matrix  $V$  as soon as  $\frac{\|V-WH\|_F}{\|V\|_F} \leq 10^{-6}$ .

Let us give more insights about the second building block of our algorithm. The minimization of the objective functions from (6.8) and (6.12) over their respective convex set  $Q$  is replaced by a sequence of simpler problems in which the objective functions are replaced by their linearization constructed at the current solution  $(U, V)$ . By posing  $Z = (U, V)$  for clarity purpose, the algorithm is therefore based on a iterative scheme such that at each iteration  $i$  we update  $Z$  as follows:

$$\begin{aligned} Z^i &= \underset{Z \in Q}{\operatorname{argmin}} \Phi(Z^{i-1}) + \langle \nabla \Phi(Z^{i-1}), Z - Z^{i-1} \rangle_F \\ &= \underset{Z \in Q}{\operatorname{argmin}} \langle \nabla \Phi(Z^{i-1}), Z \rangle_F + d, \end{aligned} \quad (6.13)$$

where  $\Phi$  designates the objective function from (6.8) or from (6.12),  $d = \Phi(Z^{i-1}) - \langle \nabla \Phi(Z^{i-1}), Z^{i-1} \rangle_F$  is a constant. For illustration purpose, we explicit here-under the form of the successive optimization problems for (6.12):

$$\begin{aligned} &\min_{U \in \mathbb{R}^{F \times K}, V \in \mathbb{R}^{K \times N}, t \in \mathbb{R}_+^{F \times K \times N}} \langle \nabla_U \Phi(U^{i-1}, V^{i-1}), U \rangle_F + \langle \nabla_V \Phi(U^{i-1}, V^{i-1}), V \rangle_F + d \\ &\text{subject to} \quad \sum_{k=1}^K t_{fkn} \geq V_{fn} \text{ for all } f, n \\ &\quad (U_{fk}, 1/2V_{kn}, t_{fkn}) \in \mathcal{Q}_r^3 \text{ for all } f, k, n \end{aligned} \quad (6.14)$$

where  $\nabla_U \Phi$  and  $\nabla_V \Phi$  are respectively the gradients of  $\Phi$  w.r.t.  $U$  and  $V$  under matrix form defined as follows:

$$\begin{aligned} \nabla_U \Phi(U, V) &= \frac{1}{2} U^{\cdot \frac{-1}{2}} \odot \left( e \left[ \sum_n V_{1n}^{\frac{1}{2}} \dots \sum_n V_{Kn}^{\frac{1}{2}} \right] \right) \\ \nabla_V \Phi(U, V) &= \frac{1}{2} V^{\cdot \frac{-1}{2}} \odot \left( \left[ \sum_f U_{f1}^{\frac{1}{2}} \dots \sum_f U_{fK}^{\frac{1}{2}} \right]^T e^T \right) \end{aligned} \quad (6.15)$$

where  $e$  are the all-one column vectors of appropriate size. As we can see, each successive problem given by (6.14) is a particular case of the general conic optimization problem defined in (6.1), in other words, each successive problem is convex. In order to solve each successive problem, we use Interior Points Methods (IPM). IPM provide a general methodology via self-concordant barrier functions to obtain polynomial-time algorithms for general families of convex programs such as LP, SDP, CGP and SOCP. In a nutshell, let us illustrate the principle of IPM on a general constrained non-linear optimization problem that

includes a simple conic constraint, that is, the variable  $x$  must belong to the nonnegative orthant:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \Phi(x) \\ \text{subject to} \quad & h(x) = 0, \\ & x \geq 0, \end{aligned} \tag{6.16}$$

where the objective function and equality constraints,  $\Phi(x) : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $h(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$  are assumed to be twice continuously differentiable. The class of primal-dual path-following interior-point methods is considered the most successful and solves problem (6.16) through a sequence of barrier problems:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \Psi_\mu(x) = \Phi(x) - \mu \sum_i \log(x_i) \\ \text{subject to} \quad & h(x) = 0, \end{aligned} \tag{6.17}$$

with decreasing values of barrier parameter  $\mu > 0$  and where the violation of inequality constraints are prevented by augmenting the objective function with the self-concordant barrier term  $(-\mu \sum_i \log(x_i))$  that causes the optimal unconstrained value to be in the feasible space. Problem (6.17) is solved through the primal-dual equations:

$$\begin{aligned} \nabla \Phi(x) + \lambda \nabla h(x) - s &= 0 \\ h(x) &= 0 \\ X S e - \mu e &= 0 \end{aligned} \tag{6.18}$$

where  $\lambda \in \mathbb{R}^m$  is the vector of Lagrangian multipliers for the equality constraints, the dual variable is  $S = \mu X^{-1} e$ ,  $X$  and  $S$  are diagonal matrices generated respectively from vectors  $x$  and  $s$ . This set of equations (6.18) is iteratively solved using Newton-Raphson methods. We refer the reader to the seminal reference [109] for more details on IPM. Here we highlight the types of barrier functions that are employed in interior-point methods for conic constraints. For the second order cone as for the nonnegative orthant used in the simple example above, natural self-concordant barriers are specified by the associated logarithmic barriers [109]. For the exponential cone defined by (6.2), the following self-concordant barrier is considered [33]:

$$F(x, y, z) = -\log(y \log(x/y) - z) - \log(x) - \log(y).$$

In this thesis, we use MOSEK software to solve each successive problem (6.13) with IPM. Algorithm 8 summarizes our algorithm to tackle both problems (6.8) and (6.12). As mentioned earlier, the SPI procedure detailed in following section can be activated at stage 7 of Algorithm 8.

Finally, by posing  $\Psi(Z|Z^{i-1}) = \langle \nabla \Phi(Z^{i-1}), Z \rangle_F + d$  with  $d = \Phi(Z^{i-1}) - \langle \nabla \Phi(Z^{i-1}), Z^{i-1} \rangle_F$ , we observe that  $\Psi(Z|Z^{i-1})$  is an auxiliary function (tight upper-bound) since:

- $\Psi(Z|Z^{i-1}) \geq \Phi(Z)$  for all  $Z \in Q$ ; it follows the fact that  $\Psi(Z|Z^{i-1})$  is the linearization of a concave function  $\Phi(Z)$ .

**Algorithm 8** Successive Conic Convex Approximation for Exact NMF

**Require:** Input matrix  $V \in \mathbb{R}_+^{F \times N}$ , the factorization rank  $K$ , number of iterations  $maxiter$  and a threshold  $th$  for SPI procedure.

**Ensure:**  $(W, H) \geq 0$  is such that  $V = WH$  with  $\frac{\|V - WH\|_F}{\|V\|_F} \leq 10^{-6}$ .

- 1: % Block 1: Initialization
- 2:  $(W^0, H^0) \leftarrow$  nonnegative random initialization( $F, K, N$ ).
- 3:  $(U^0, V^0) \leftarrow G(W^0, H^0)$  with  $G$  defining the change of variable
- 4:  $Z^0 \leftarrow (U^0, V^0)$
- 5: % Block 2: iterative update of  $Z$
- 6: **for**  $i = 1, 2, \dots, maxiter$  **do**
- 7:      $Z^i \leftarrow \underset{Z \in Q}{\operatorname{argmin}} \langle \nabla \Phi(Z^{i-1}), Z \rangle_F$  with IPM available in MOSEK software and activation of SPI procedure if required.
- 8: **end for**
- 9:  $(W, H) \leftarrow G^{-1}(Z^i)$
- 10: % Block 3: Final Refinement
- 11:  $(W, H) \leftarrow$  Algo A-HALS( $W, H$ )

- $\Psi(Z^{i-1}|Z^{i-1}) = \Phi(Z^{i-1})$ .

Further each new iterate  $Z^i$  computed by IPM is the optimal solution  $Z^*$  of (6.13), since each successive problem (6.13) is convex. This guarantees  $\Phi$  to decrease at each iteration.

**Lemma 6.3.1.** *Let  $Z, Z^{(i-1)} \in Q$ , and let  $\Psi(Z|Z^{(i-1)})$  be an auxiliary function for  $\Phi$  at  $Z^{(i-1)}$ . Then  $\Phi$  is non-increasing under the update*

$$Z^{(i)} = \underset{Z \in Q}{\operatorname{argmin}} \Psi(Z|Z^{(i-1)}).$$

*Proof.* We have by definition that:

$$\Phi(Z^{(i-1)}) = \Psi(Z^{(i-1)}|Z^{(i-1)}) \geq \underset{Z \in Q}{\min} \Psi(Z|Z^{(i-1)}) = \Psi(Z^i|Z^{(i-1)}) \geq \Phi(Z^i).$$

□

In the next section we detail the SPI procedure.

### 6.3.1 Sparsity Pattern Integration

Due to nonnegative constraints on the entries of  $W$  and  $H$ , the sparsity for an input matrix  $V$  (many entries equal to zero) induces sparse patterns for the solutions  $(W, H)$ , as for the solutions  $(U, V)$  of (6.12) since  $W_{fk} = G(U_{fk}) = \sqrt{U_{fk}}$  and  $H_{kn} = G(V_{kn}) = \sqrt{V_{kn}}$ . One can observe that the objective function  $\Phi$  from (6.12) is not L-smooth on the interior of the domain (non-negative orthant). When an entry  $U_{fk}$  for the current solution  $U^{i-1}$  tends to zero, the corresponding entry in the gradient of  $\Phi$  w.r.t.  $U$  tends to  $\infty$  which

therefore ends the optimization process. As a first simplistic approach, we normalized the gradients of  $\Phi$  evaluated at  $(U^{i-1}, V^{i-1})$  w.r.t. Frobenius norm. However, based on preliminary numerical experiments, we have noticed that the solution obtained  $(W, H)$  for an input matrix with zero entries cannot reach the desired tolerance threshold to consider the solution as an exact NMF, even if the obtained solutions  $(W, H)$  seem to be close to an exact NMF. In order to tackle this issue and enables the solution to reach the desired tolerance of  $10^{-6}$  for the relative error, we integrated an additional stage within the second building block of Algorithm 8 when used to tackle (6.12). This additional stage is referred to as "Sparsity Pattern Integration" and can be summarized as follows: let us consider a simple case for which one entry  $W_{\bar{f}, \bar{k}}$  of the current solution  $W^{i-1}$  tends to zero, it implies that entry  $U_{\bar{f}, \bar{k}}$  tends to zero as well. Let us now fix this entry  $U_{\bar{f}, \bar{k}}$  to zero, it implies that this variable is dropped from the optimization process. Let us observe the impact on the constraints of (6.10) in which  $U_{\bar{f}, \bar{k}}$  is involved; the inequality constraints identified by index  $f = \bar{f}$  are:

$$\sqrt{U_{\bar{f}, 1}}\sqrt{V_{1, n}} + \dots + \sqrt{U_{\bar{f}, \bar{k}}}\sqrt{V_{\bar{k}, n}} + \dots + \sqrt{U_{\bar{f}, K}}\sqrt{V_{K, n}} \geq V_{\bar{f}, n} \text{ for } 1 \leq n \leq N.$$

First, since  $\sqrt{U_{\bar{f}, \bar{k}}} = 0$ , there is no more constraints on  $\sqrt{V_{\bar{k}, n}}$  for  $N$  inequalities identified by index  $f = \bar{f}$ . Second, for the problem (6.12) and the successive convex problems (6.14), it is then clear that  $N$  conic variables  $t_{\bar{f}, \bar{k}, n}$  (and hence the  $N$  associated conic constraints) can be dropped from the optimization process. Finally, the objective function is also automatically impacted by removing the linear term  $[\nabla_U \Phi(U^{i-1}, V^{i-1})]_{\bar{f}, \bar{k}} U_{\bar{f}, \bar{k}}$ .

The same rationale is followed for the case entries of the current solution for  $V$  tend to zero. To sum up, at each iteration, Algorithm 8 verifies if entries of the current solutions  $(W^{i-1}, H^{i-1}) = G^{-1}(U^{i-1}, V^{i-1})$  is below a threshold  $th$  defined by the user, then the corresponding entries of  $U$  and  $V$  are set to zero so that we determine a sparsity pattern, that are the indices of the entries set to zero. The optimization problem (6.14) is automatically updated based on the current sparsity pattern with the approach explained above. Let us illustrate the impact of triggering this SPI procedure on the solutions obtained for the factorization of the following input matrix  $V$  with zero entries:

$$V = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix}. \quad (6.19)$$

The nonnegative rank of (6.19) is known and is equal to 5. Algorithm 8 is used to compute an exact NMF of  $V$  with the following input parameters:

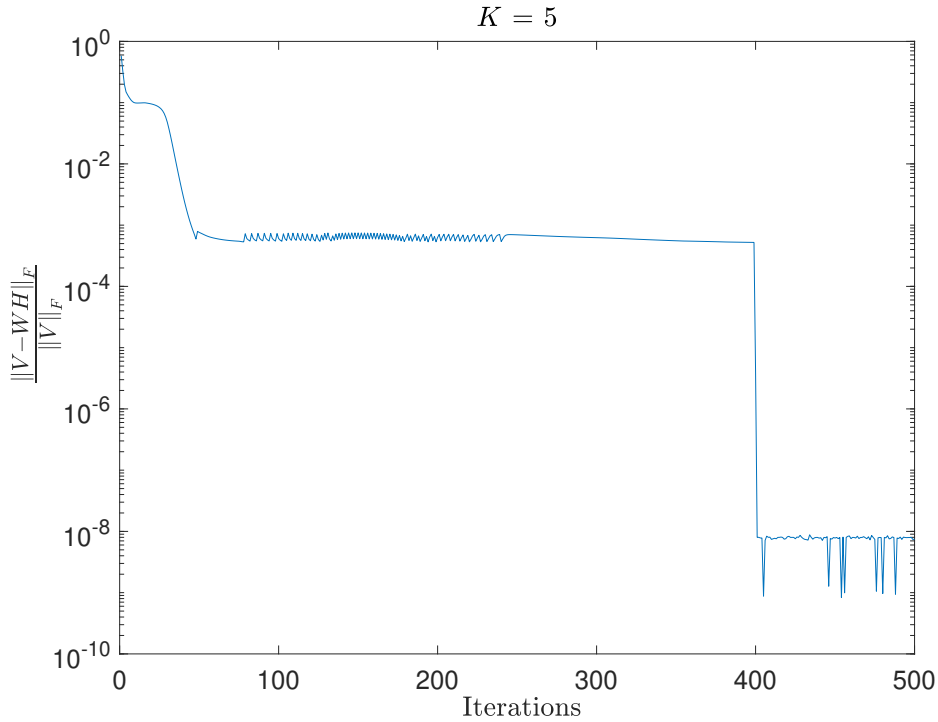
- $K = 5 = \text{rank}_+(V)$ ,
- $th = 10^{-3}$ ,

- the maximum number of iterations defined by parameter *maxiter* is set to 500 and the SPI procedure is activated in the iterations interval [400, 500].

Figure 6.3 displays the evolution of  $\frac{\|V-WH\|_F}{\|V\|_F}$  along iterations for  $V$  (6.19) with a factorization rank  $K = 5$ . One can observe that, once the SPI is activated, the relative Frobenius error drops from  $5 \cdot 10^{-4}$  to  $8 \cdot 10^{-9}$ , hence below the tolerance of  $10^{-6}$  such that we assume an exact NMF  $(W, H)$  is found. For this experiment, we obtain:

$$W = \begin{pmatrix} 0 & 1.4748 & 0.9259 & 0 & 0 \\ 0.7824 & 0 & 1.8517 & 0 & 0 \\ 0 & 0 & 0.9259 & 0 & 1.4716 \\ 0 & 0 & 0 & 0.6024 & 1.4716 \\ 0.7824 & 0 & 0 & 1.2049 & 0 \\ 0 & 1.4748 & 0 & 0.6024 & 0 \end{pmatrix},$$

$$H = \begin{pmatrix} 0 & 0 & 1.2781 & 0 & 0 & 1.2781 \\ 0 & 0.6780 & 1.3561 & 0.6780 & 0 & 0 \\ 0 & 0 & 0 & 1.0801 & 1.0801 & 0 \\ 1.6599 & 1.6599 & 0 & 0 & 0 & 0 \\ 0.6796 & 0 & 0 & 0 & 0.6796 & 1.3591 \end{pmatrix}.$$



**Fig. 6.3.** Evolution of  $\frac{\|V-WH\|_F}{\|V\|_F}$  along iterations; SPI is activated in the iterations interval [400, 500].

## 6.4 Numerical experiments

Algorithm 8 is tested for two main applications (1) the computation of exact NMF for particular classes of matrices usually considered in the exact NMF literature and (2) the computation of the largest biclique in a bipartite graph. For the first application, both optimization problems (6.8) and (6.12) are considered and solved by Algorithm 8. For the second application, as we need to use under-approximation models (the reason will be detailed in Section 6.4.2), only optimization problem (6.12) will be considered. The algorithm will be benchmarked against recent algorithms available in the literature.

### 6.4.1 Benchmark Nonnegative Matrices for Exact NMF

Throughout this section, we will compare exact NMF algorithms on the following nonnegative matrices:

- Randomly Generated Matrices: It is standard in the NMF literature to use randomly generated matrices to compare algorithms (see, e.g., [38]), with the nice feature that the resulting nonnegative rank of these matrices can be specified. In this chapter, we have generated matrices  $V \in \mathbb{R}^{4 \times 15}$  as follows:  $W$  is taken as the matrix from (2.2):

$$W = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

so that  $\text{rank}(W) = 3 < K = 4$ , and each column of  $H$  is distributed using the Dirichlet distribution of parameter  $(0.1, \dots, 0.1)$ . Each column of  $H$  with an entry larger 0.8 is resampled as long as this condition does not hold. This guarantees that no data point is close to a column of  $W$ .

- Infinitesimally rigid factorizations: in this chapter, we consider four infinitesimally rigid factorizations for  $5 \times 5$  matrices with positive entries and of nonnegative rank

four extracted from [80]:

$$\begin{aligned}
 V_{inf1} &= \begin{pmatrix} 573705 & 806520 & 167622 & 246500 & 531659 \\ 397096 & 39600 & 299176 & 63720 & 274120 \\ 131646 & 403260 & 30269 & 226915 & 264510 \\ 9114 & 85160 & 311182 & 827468 & 851798 \\ 147857 & 3200 & 351037 & 599025 & 697755 \end{pmatrix}, \\
 V_{inf2} &= \begin{pmatrix} 30893 & 319912 & 149770 & 873 & 111428 \\ 383490 & 87990 & 5580 & 628440 & 587250 \\ 560076 & 1030324 & 331070 & 288045 & 350647 \\ 203830 & 305184 & 277512 & 264376 & 205933 \\ 90911 & 142936 & 500784 & 618842 & 609633 \end{pmatrix}, \\
 V_{inf3} &= \begin{pmatrix} 948201 & 723609 & 958755 & 591858 & 397953 \\ 222448 & 218040 & 30429 & 348793 & 15825 \\ 329588 & 7189 & 623001 & 12012 & 469185 \\ 467424 & 160704 & 115092 & 835504 & 343912 \\ 1114797 & 932972 & 975775 & 997164 & 636096 \end{pmatrix}, \\
 V_{inf4} &= \begin{pmatrix} 88076 & 294646 & 658787 & 902872 & 244559 \\ 2216 & 4216 & 596705 & 652698 & 250465 \\ 279360 & 180864 & 769506 & 1051380 & 391634 \\ 553284 & 826606 & 765406 & 293965 & 883775 \\ 696039 & 897917 & 148301 & 832169 & 169525 \end{pmatrix}.
 \end{aligned}$$

These matrices have shown to be challenging to factorize. We refer the reader to [80] for more details.

- Nested hexagons problem: as explained in Section 1.6 NMF has a nice geometric interpretation. In particular we have seen that computing an exact NMF with a factorization rank  $K$  is equivalent to finding a polytope,  $\text{conv}(\Pi_{\Delta^F}(W))$ , nested between two given polytopes,  $\text{conv}(\Pi_{\Delta^F}(V))$  and the unit simplex  $\Delta^F$ . The dimension of the inner polytope  $\text{conv}(\Pi_{\Delta^F}(V))$  is  $\text{rank}(V) - 1$ , while the dimension of the outer polytope  $\Delta^F$  is  $F - 1$ . The dimension of  $\text{conv}(\Pi_{\Delta^F}(W))$  is not known a priori but when we impose explicitly that  $\text{rank}(V) = \text{rank}(W)$ , we explained that the outer polytope can be restricted to  $\Delta^F \cap \text{col}(\Pi_{\Delta^F}(V))$  where  $\text{col}(A) = \{x \in \mathbb{R}^F | x = Ay, y \in \mathbb{R}^N\}$ . Since the dimension of  $\Delta^F \cap \text{col}(\Pi_{\Delta^F}(V)) = \text{rank}(V) - 1$  per [Lemma 2.5, [56]], then the inner, nested, and outer polytopes have the same dimension. This problem is well known in computational geometry and is referred to as nested polytope problem. Here we consider a family of input matrices whose computing an exact NMF corresponds to find a polytope nested between two hexagons; let  $a > 1$  and let  $V_a$  be



the matrix:

$$V_a = \frac{1}{a} \begin{pmatrix} 1 & a & 2a-1 & 2a-1 & a & 1 \\ 1 & 1 & a & 2a-1 & 2a-1 & a \\ a & 1 & 1 & a & 2a-1 & 2a-1 \\ 2a-1 & a & 1 & 1 & a & 2a-1 \\ 2a-1 & 2a-1 & a & 1 & 1 & a \\ a & 2a-1 & 2a-1 & a & 1 & 1 \end{pmatrix}$$

which has  $\text{rank}(V_a) = 3$  and therefore the dimension of the nested hexagons is  $\text{rank}(V) - 1 = 2$ . The inner hexagon is smaller than the outer one with a ratio of  $\frac{a-1}{a}$ . We consider three values for  $a$ :

- $a = 2$ : the inner hexagon is twiced smaller than the outer one and we can fit a triangle between the two, thus  $\text{rank}_+(V_a) = 3$ .
- $a = 3$ : the inner hexagon is  $2/3$  smaller than the outer one and we can fit a rectangle between the two, thus  $\text{rank}_+(V_a) = 4$ .
- $a = 4$ :  $\text{rank}_+(V_a) = 5$ .
- $a \rightarrow +\infty$ , which gives:

$$V = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix}$$

with  $\text{rank}_+(V) = 5$

For each of the 8 matrices above we ran Algorithm 8 and the algorithm from [133] with Multi-Start 1 heuristic "ms1" and the Rank-by-rank heuristic "rbr" ten times with SPARSE10 for the initialization as recommended in [133]. For these analyzes, the target precision is  $10^{-6}$ . Table 6.2 reports the number of success over 10 attempts for computing the exact NMF of the input matrices such that  $\frac{\|V-WH\|_F}{\|V\|_F}$  is below the target precision. One can observe that for three types of matrices (random, nested hexagons with  $a = 2$  and  $a = 3$ ), SCCAE-NMF algorithm with both formulations (6.8) and (6.12) perform equivalently with Algorithm from [133]. Further, Algorithm from [133] outperforms in average SCCAE-NMF algorithm for  $V_{inf2}$ ,  $V_{inf3}$  and the two last nested hexagons. However, for matrices  $V_{inf1}$  and  $V_{inf4}$ , none of the runs performed with Algorithm from [133] found a rank-4 factorization while SCCAE-NMF algorithm finds 2 times out of 10 attempts. For the nested hexagons with  $a \rightarrow +\infty$ , SCCAE-NMF with formulation (6.8) found none of the factorization while SCCAE-NMF with formulation (6.12) gives satisfactory results. We report that SCCAE-NMF with formulation (6.8) requires 50 attempts one average to find

Table 6.2: Comparison of Algorithm 8 with algorithm from [133] with "ms1" and "rbr" heuristic for 10 attempts to compute the factorizations of matrices described in the text. In bold we specify the matrices for which SCCAE-NMF is the only one to find exact NMF's.

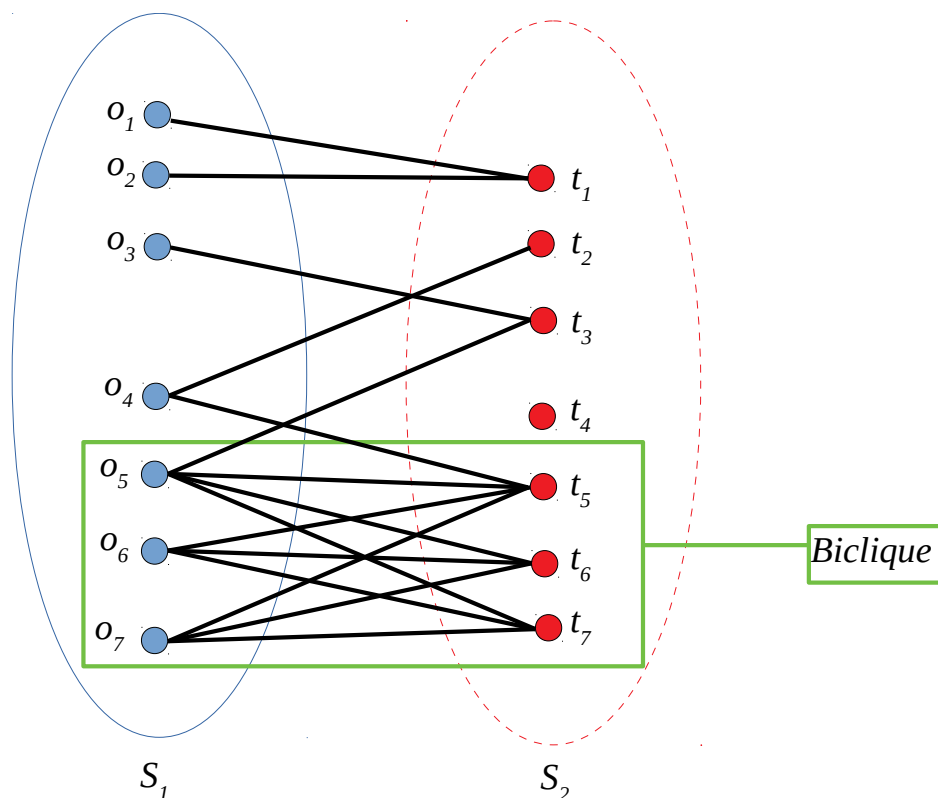
		SCCAE-NMF with Problem (6.8)	SCCAE-NMF with Problem (6.12)	Algorithm from [133] with "ms1"	Algorithm from [133] with "rbr"
<b>Matrices</b>		<b>/10</b>	<b>/10</b>	<b>/10</b>	<b>/10</b>
Random matrices		10	10	9	10
Inf. Rig. Fac.	$V_{inf1}$	1	2	0	0
	$V_{inf2}$	6	5	4	10
	$V_{inf3}$	2	1	3	10
	$V_{inf4}$	4	2	0	0
Nested hexagons	$a = 2$	10	10	10	10
	$a = 3$	10	10	10	10
	$a = 4$	3	7	3	10
	$a \rightarrow +\infty$	0	6	1	10

a rank-5 factorization. We observe similar behaviour for SCCAE-NMF with formulation (6.8) based on additional numerical tests involving input matrix whose have significant number of null entries; the number of attempts required to find an exact factorization significantly increases.

#### 6.4.2 The largest biclique in a bipartite graph

A bipartite graph  $G = (S, A)$  is a graph whose vertices can be divided into two disjoint and independent sets  $S_1$  and  $S_2$  such that  $S = S_1 \cup S_2$  and every edge connects a vertex in  $S_1$  to one in  $S_2$ , in other words, there is no edge that connects two vertices that belong to the same set  $S$ . A biclique is a subgraph of  $G$  where all the vertices are connected by an edge. The so-called maximum-edge biclique problem in a bipartite graph  $G$  is the problem of finding a biclique in  $G$  with maximum number of edges. The corresponding decision problem: *Given  $B$ , does  $G$  contain a biclique with at least  $B$  edges?* has been shown to be NP-complete [113]. Therefore, the problem of finding the biclique with maximum number of edges in  $G$  is at least NP-hard [59]. For such a problem, the input matrix  $V \in \{0, 1\}^{F \times N}$  is the so-called biadjacency matrix of the bipartite graph  $G = (S_1 \cup S_2, A)$  with  $S_1 = \{o_1, \dots, o_f, \dots, o_F\}$  and  $S_2 = \{t_1, \dots, t_n, \dots, t_N\}$  and  $V_{fn}=1$  if and only if  $(o_f, t_n) \in A$ . Figure 6.4 shows an illustration of the maximum-edge biclique problem for a bipartite graph that corresponds to the following biadjacency matrix:

$$V = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}. \quad (6.20)$$



**Fig. 6.4.** An illustration of the maximum-edge biclique problem for a graph that corresponds to the biadjacency matrix (6.20).

In [59], authors show that finding the best rank-one under-approximation of a biadjacency matrix is equivalent to finding the largest biclique in the corresponding bipartite graph, that is, finding the largest rectangles of all ones hidden in the binary matrix. We compare our approach, namely SCCAE-NMF, with formulation (6.12), with that of [59] on randomly generated binary matrices with  $F = N$  and values for  $N$  in the set  $\{8, 10, 14, 16, 18, 20, 24\}$ . The method that finds the largest rectangle is the better one. Each method is launched 10 times and the best result among these 10 attempts is kept. Table 6.3 displays the results obtained, namely the number of edges for the largest biclique extracted by the two methods from bipartite graphs defined by matrices  $V$ .

One can observe that for biadjacency matrices  $V$  such that  $F = N \in [8, 20]$ , methods find the same results. But when the size is above  $20 \times 20$ , then method from [59] outperforms

Table 6.3: Comparison of Algorithm 8 with algorithm from [59] for finding the maximum-edge biclique from a bipartite graph. The table reports the number of edges for the largest biclique extracted by the two methods from bipartite graphs defined by random binary matrices  $V$  of size  $N \times N$ .

	SCCAE-NMF with formulation (6.12)	Algorithm from [59]
$N = 8$	12	12
$N = 10$	12	12
$N = 14$	18	18
$N = 16$	24	24
$N = 18$	24	24
$N = 20$	24	24
$N = 24$	18	32

SCCAE-NMF. It implies that our method shows difficulty to scale up. We report that the number of random initializations required by SCCAE-NMF to find competitive solutions significantly increase with the size of the input matrix.

## 6.5 Conclusion

In this chapter, we introduced two formulations for computing exact NMFs. Each of the proposed formulation relies on the construction and the resolution of a specific optimization problem; namely problems (6.4) and (6.9) that can be respectively referred to as under-approximation and upper-approximation formulations for NMF. For each optimization problem we introduced a particular change of variables that enabled the use of two special cases of conic constraints, that are the exponential and second-order conic constraints. In order to solve the two optimization problems, we proposed a general algorithm, denoted SCCAE-NMF, that relies on the resolution of successive approximations of the objective functions based on interior-point methods. We showed that the successive approximated problems belong to two special cases of conic programming. We showed that our algorithm is able to compute exact NMFs for several classes of nonnegative matrices (namely, randomly generated, infinite rigid matrices and nested hexagons problem matrices) and as such demonstrate its competitiveness compared to recent methods from the literature. However, we have tested algorithm SCCAE-NMF on a limited number of nonnegative matrices. In the future we plan to test it on a larger library of nonnegative matrices at our disposal, in order to better understand the behavior of SCCAE-NMF along with the two formulations (6.8) and (6.12). In particular, we need to develop better strategies to deal with nonnegative matrices with many zero entries when SCCAE-NMF is used with formulation based on exponential cones. We have shown that our framework is flexible and can be used for other interesting applications of NMF such as the maximum-edge biclique

problem. In particular, we demonstrated that SCCAE-NMF along with formulation (6.8) is able to give competitive results compared to recent methods for small size biadjacency matrices. However, we highlighted a drawback of our approach; the number of random initializations required by SCCAE-NMF to find competitive solutions significantly increase with the size of the input matrix. In others words, the approach shows difficulties when the size of the input matrix increases.

Further works include:

- the theoretical guarantees for the convergence of SCCAE-NMF algorithm, in particular can we prove that SCCAE-NMF converges towards stationary points of both problems (6.8) and (6.12) ?
- The design of more advanced strategies for the initialization of  $(U, V)$ .
- The elaboration of alternative formulations for (6.8) and (6.12) to deal with the non-uniqueness of the NMF models, see Section 1.8.1. In particular, we plan to develop new formulations so that we remove most of alternative solution  $V = \tilde{W}\tilde{H} = (WE)(E^{-1}H)$  for a given solution  $(W, H)$  and for any invertible matrix  $E$  such that  $WE \geq 0$  and  $E^{-1}H \geq 0$ .
- The use of our framework to others applications such as the computation of symmetric NMFs. Symmetric NMF can be used for data analysis and in particular for various clustering tasks [81].

## 7 Conclusion

We conclude the thesis in this chapter. We first summarize our results, we recall the main contributions over four aspects and finally give some directions for further research.

### Summary

In this thesis, we have explored a famous problem from linear algebra, namely nonnegative matrix factorization (NMF). NMF is a linear dimensionality reduction technique for non-negative data, and requires factors of the corresponding low-rank matrix approximation to be nonnegative. These additional constraints enhance compression (through sparsity) and allow to extract easily interpretable and meaningful information from the input data. However, they make the problem much more difficult to solve (NP-hard).

**Chapter 1** We gave a brief introduction of the thesis and discuss some theoretical background required for the thesis purposes. In particular the geometric interpretation gave useful insights to understand one of the main issues about NMF models, that is, the nonuniqueness and the main motivation for the introduction of NMF models and associated (optimization) problems relying on the notion of minimum-volume, that is, the construction of identifiable NMF models and problems. By solving these problems, under some mild conditions such as the SSC condition on  $H$ , the solutions obtained  $(W^*, H^*)$  are permuted and scaled versions of the ground-truth factors  $(W^\#, H^\#)$  that gave rise to the data  $V$ . These degrees of freedom are unavoidable and, most importantly, inconsequential for the applications at hand.

**Chapter 2** We have shown that minimum-volume NMF can be used meaningfully for the rank-deficient scenario. We have provided an optimization problem for minimum-volume NMF that relies on the minimization of an objective function that integrates the Frobenius norm of the residual matrix  $V - WH$  and a function that measures the volume of the columns of  $W$ . We have provided a simple algorithm, referred to as min-vol NMF, to tackle this optimization problem and have illustrated the behaviour of the method on synthetic and real-world data sets. In particular, we solved the proposed minimum-volume NMF problem by transforming the objective function for the subproblem in  $W$  into a quadratic form (defined by matrix  $A$ ) which is a strongly convex upper approximation of the objective function and we used a PFGM optimization scheme so that we have a linear convergence method with rate  $1 - \sqrt{\kappa^{-1}}$  where  $\kappa$  is the condition number of  $A$  [107]. We

have developed and tested a faster algorithm referred to as fast-min vol. We have shown that fast-min vol NMF outperforms min-vol NMF algorithm in low-dimensional setting but tends to be more easily stuck in saddle points as the dimension of the problem increases.

**Chapter 3** We have presented a new NMF problem for audio source separation based on the minimization of a cost function that includes a  $\beta$ -divergence (data fitting term) and a penalty term that promotes solutions  $W$  with minimum volume. We have proved the identifiability of the problem in the exact case, under the sufficiently scattered condition for the activation matrix  $H$ . We have provided multiplicative updates to tackle this problem and have illustrated the behaviour of the method on real-world audio signals. On an application aspect, we briefly reviewed the audio BSS problem, and discussed how and why NMF can be applied to audio Blind source separation (BSS) problem. By solving the minvol  $\beta$ -divergence NMF problem, we demonstrated that it can be used to decompose single channel audio recording of piano music into components correspond to each of the musical notes. We highlighted in particular the capacity of minvol  $\beta$ -divergence NMF problem to deal with the case where  $K$  is overestimated by setting automatically to zero some components and give good results for the source estimates hence performing model order selection automatically.

**Chapter 4** We have presented a new NMF approach for blind spectral unmixing, called multi-resolution  $\beta$ -NMF (MR- $\beta$ -NMF). The estimation relies on the minimization of the  $\beta$ -divergence, a flexible family of measures of fit. MR- $\beta$ -NMF addresses the resolution trade-off between two adversarial dimensions by fusing the information coming from multiple data with different resolutions in order to produce a factorization with high resolutions for all the dimensions. We have provided multiplicative updates to tackle the minimization problem and we showed that MR- $\beta$ -NMF is flexible and can be successfully applied to various problems. In particular, we have showcased its efficiency on two instrumental examples. The first is the audio spectral unmixing for which the frequency-by-time data matrix is computed with the short-time Fourier transform and is the result of a trade-off between the frequency resolution and the temporal resolution. We highlighted the capacity of this approach to provide solutions that show high frequency and high temporal accuracy taking advantage from the input data. Based on these results, MR- $\beta$ -NMF seems to be well suited for audio applications such as transcription problems and performs in general better than baseline NMF methods. The second is BHU for which the wavelength-by-location data matrix is a trade-off between the number of wavelengths measured and the spatial resolution. We demonstrated the efficiency of MR- $\beta$ -NMF to tackle the HS-MS data fusion problem. Based on various quantitative quality assessments, the proposed method performs competitively with the state of the art.

**Chapter 5** We have presented a general framework to solve penalized  $\beta$ -NMF problems that integrates a set of disjoint constraints on the variables. Using this framework, we

showed that we can derive algorithms that compete favorably with the state of the art for a wide variety of  $\beta$ -NMF problems, such as the simplex-structured NMF and the minimum-volume  $\beta$ -NMF with sum-to-one constraints on the columns of  $W$ . We have also shown how to extend the framework to non-linear disjoint constraints, with application to a sparse  $\beta$ -NMF model for  $\beta = 1$  where each column of  $W$  lie on a hyper-sphere.

**Chapter 6** We introduced two formulations for computing exact NMFs. Each of the proposed formulation relies on the construction and the resolution of a specific optimization problem. For each optimization problem we introduced a particular change of variables that enabled the use of two special cases of conic constraints, that are the exponential and second-order conic constraints. In order to solve the two optimization problems, we proposed a general algorithm, denoted SCCAE-NMF, that relies on the resolution of successive approximations of the objective functions based on interior-point methods. We showed that the successive approximated problems belong to two special cases of conic programming. Unlike the majority of existing algorithms to tackle NMF problems, our algorithm updates both matrices  $W$  and  $V$  simultaneously. We showed that our algorithm is able to compute exact NMFs for several classes of nonnegative matrices and as such demonstrate its competitiveness compared to recent methods from the literature. We have shown that our framework is flexible and can be used for other interesting applications of NMF such as the maximum-edge biclique problem.

## Summary of contributions

The contributions of this thesis were centered on NMF over four aspects: models, optimization problems, algorithms and applications:

1. On the model and optimization problems aspect, we studied a specific class of NMF called minimum-volume NMF. This class of NMF generalizes another class of NMF called Separable NMF. We showed that our proposed model and associated optimization problem with determinant volume for minimum-volume NMF is identifiable under the SSC condition on  $H$ , and we argue that such model and optimization problem are highly relevant for real-life applications. Further, we proposed models and problems, referred to as multi-resolution NMF, to tackle a common issue for many input matrices; they are generally the result of a resolution trade-off between two adversarial dimensions. We addressed this issue by fusing the information coming from multiple data with different resolutions in order to produce a factorization with high resolutions for all the dimensions. Finally we proposed new models and problems to tackle a special case of NMF referred to as exact NMF by using conic programming.
2. On the algorithmic aspect, we proposed efficient algorithms to solve the optimization problems for minimum-volume NMF. We mainly focused on two classes of optimization problems for minimum-volume NMF: the first one integrates a Frobenius



norm for the data fitting term whereas the second one integrates the family of  $\beta$ -divergences, in particular we deal with the Kullback-Leibler divergence that is notorious hard to handle. Further we introduced a general framework to derive efficient algorithms to tackle penalized  $\beta$ -divergence NMF problems under disjoint equality constraints. Finally we proposed a general algorithm that is able to tackle problems to compute an exact NMF such that at each iteration, matrices  $W$  and  $H$  are simultaneously updated.

3. On the application aspect, we demonstrated the efficiency of models and algorithms compared to state-of-the-art methods on hyperspectral imaging and audio source separation problems. In particular, we showed that minvol  $\beta$ -divergence NMF was able to perform automatic MOS which is rare in the literature. Further we showed that MR- $\beta$ -NMF was able to give a factorization that show high resolution in adversarial dimensions, we showed that the Kullback-Leibler divergence was an efficient choice to deal with the HS-MS fusion problem in the case we have Poisson noise within the data.

## Summary of perspectives and further research

We conclude this section by listing the four major open problems related to the thesis.

- **Robustness for minimum-volume NMF:** Minimum-volume NMF is arguably the most versatile class of NMF models and optimization problems as it allows identifiability condition under weak requirements. However, as opposed to separable NMF, currently there is no theoretical robustness analysis on models and optimization problems for minimum-volume NMF. We will investigate the conditions on factors  $W$  and  $H$  under which minimum-volume NMF models and optimization problems are robust to bounded noise.
- **Theoretical analysis of the MOS:** We have seen in the numerical experiments carried out in Section 3.4 that, minvol KL-NMF automatically set to zero some components when the factorization rank  $K$  is overestimated regarding the number of rank-one sources present within the audio signal. Currently, the theoretical analysis of such phenomenon remains open. We will focus in developing the theoretical understanding of this behavior of the volume regularizer.
- **Theoretical guarantees of recoverability for MR- $\beta$ -NMF:** We have seen in Sections 4.4 and 4.5 that MR- $\beta$ -NMF approach was able to give a factorization with high resolutions for all the dimensions and stable results. The theoretical guarantees for recoverability or identifiability of the latent factors for the model and the associated problem remains open.
- **Theoretical convergence analysis of SCCAE-NMF:** Despite the promising results obtained with SCCAE-NMF algorithm in many experiments, the convergence

results are limited to the decreasingness of the objective function along iterations. The theoretical convergence analysis of SCCAE-NMF remains open. In particular can we prove that SCCAE-NMF converges towards stationary points of both problems (6.8) and (6.12) ?

*“He said that the root of education is bitter but the fruit is sweet.”*

- Attributed to Aristotle by Diogenes Laertius in his *Lives of the Eminent Philosophers*.

*“A lack of education is the mother of all suffering.”*

- Pythagoras, reported by Stobaeus 2.31.96.

## Bibliography

- [1] M. Abdolali and N. Gillis. “Simplex-Structured Matrix Factorization: Sparsity-based Identifiability and Provably Correct Algorithms”. In: *arXiv preprint arXiv:2007.11446* (2020).
- [2] Aeg. Menagius. *The Lives of the Ancient Philosophers*. 1702.
- [3] A. Aggarwal, H. Booth, J. O’Rourke, S. Suri, and C. Yap. “Finding minimal convex nested polygons”. In: *Information and Computation* 83.1 (1989), pp. 98–110.
- [4] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli. “Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis”. In: *IEEE Transactions on Geoscience and Remote Sensing* 40.10 (2002), pp. 2300–2312.
- [5] B. Aiazzi, S. Alparone, S. Baronti, A. Garzelli, and M. Selva. “MTF-tailored multiscale fusion of high-resolution MS and Pan imagery”. In: *Photogrammetric Engineering and Remote Sensing* 72.5 (2006), pp. 591–596.
- [6] B. Aiazzi, S. Baronti, and M. Selva. “Improving component substitution Pansharp-ening through multivariate regression of MS+Pan data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 45.10 (2007), pp. 3230–3239.
- [7] A. M. S. Ang and N. Gillis. “Algorithms and Comparisons of Nonnegative Matrix Factorizations With Volume Regularization for Hyperspectral Unmixing”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.12 (2019), pp. 4843–4853.
- [8] A.M.S Ang and N Gillis. “Accelerating nonnegative matrix factorization algorithms using extrapolation”. In: *Neural computation* 31.2 (2019), pp. 417–439.
- [9] A.M.S. Ang and N. Gillis. “Volume regularized Non-negative Matrix Factorizations”. In: *2018 Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2018.
- [10] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, and V. Visani. “The successive projections algorithm for variable selection in spectroscopic multicomponent analysis”. In: *Chemometrics and Intelligent Laboratory Systems* 57.2 (2001), pp. 65–73.
- [11] S. Arora, R. Ge, R. Kannan, and A. Moitra. “Computing a Nonnegative Matrix Factorization—Provably”. In: *SIAM Journal on Computing* 45.4 (2016), pp. 1582–1611.

- [12] N. Asgarian and R. Greiner. “Using Rank-One Biclusters to Classify Microarray Data”. In: 2007.
- [13] R.W. Basedow, D.C. Carmer, and M.E. Anderson. “HYDICE system: Implementation and performance”. In: *SPIEs 1995 Symp. OE/Aerospace Sens. Dual Use Photonics*. Vol. 2480. 1995, pp. 258–268.
- [14] S. Basu, R. Pollack, and M.-F. Roy. “On the Combinatorial and Algebraic Complexity of Quantifier Elimination”. In: 43.6 (1996).
- [15] A. Beck. *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017. DOI: 10.1137/1.9781611974997.
- [16] E. Benetos, S. Dixon, Z. Duan, and S. Ewert. “Automatic Music Transcription: An Overview”. In: *IEEE Signal Processing Magazine* 36.1 (2019), pp. 20–30.
- [17] S. Bergmann, J. Ihmels, and N. Barkai. “Iterative signature algorithm for the analysis of large-scale gene expression data”. In: *Phys Rev E Stat Nonlin Soft Matter Phys* 67 (2003).
- [18] D. P. Bertsekas. *Nonlinear Programming*. Second. Belmont, MA: Athena Scientific, 1999. ISBN: ISBN 1-886529-00-0.
- [19] M. Biggs, A. Ghodsi, and S. Vavasis. “Nonnegative matrix factorization via rank-one downdating”. In: *the 2008 International Conference on Machine Learning*. ICML. 2008.
- [20] J. Bioucas-Dias. “A variable splitting augmented Lagrangian approach to linear spectral unmixing”. In: *IEEE Workshop Hyperspectral Image and Signal Processing: Evolution Remote Sensing*. IEEE. 2009, 1–4.
- [21] J. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. “Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5.2 (2012), pp. 354–379.
- [22] C. Bocci, E. Carlini, and F. Rapallo. “Perturbation of Matrices and Nonnegative Rank with a View toward Statistical Models”. In: *SIAM Journal on Matrix Analysis and Applications* 32.4 (2011), pp. 1500–1512.
- [23] R. Bro. “Multi-way Analysis in the Food Industry: Models, Algorithms, and Applications”. PhD thesis. University of Amsterdam, 1998.
- [24] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. “Metagenes and molecular pattern discovery using matrix factorization”. In: *Proceedings of the National Academy of Sciences* 101.12 (2004), pp. 4164–4169.
- [25] W.J. Carper, T.M. Lillesand, and R.W. Kiefer. “The use of intensity-hue-saturation transform for merging SPOT panchromatic and multispectral image data”. In: *Photogramm. Eng. Remote Sens.* 56.4 (1990), pp. 459–467.

- [26] P.S. Chavez, S.C. Sides, and J.A. Anderson. “Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic”. In: *Photogramm. Eng. Remote Sens.* 57.3 (1991), pp. 265–303.
- [27] D. Chistikov, S. Kiefer, I. Marusic, M. Shirmohammadi, and J. Worrell. “Nonnegative Matrix Factorization Requires Irrationality”. In: *SIAM Journal on Applied Algebra and Geometry* 1.1 (2017), pp. 285–307.
- [28] D. Chistikov, S. Kiefer, I. Marusic, M. Shirmohammadi, and J. Worrell. “On Restricted Nonnegative Matrix Factorization”. In: *CoRR* abs/1605.07061 (2016).
- [29] A. Cichocki, R. Zdunek, and S. Amari. “Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization”. In: *Independent Component Analysis and Signal Separation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 169–176.
- [30] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley and Sons, 2009.
- [31] J. E. Cohen and U. G. Rothblum. “Nonnegative ranks, decompositions, and factorizations of nonnegative matrices”. In: *Linear Algebra and its Applications* 190 (1993), pp. 149–168.
- [32] M. D. Craig. “Minimum-volume transforms for remotely sensed data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 32.3 (1994), pp. 542–552.
- [33] J. Dahl and E. Andersen. *A primal-dual interior-point algorithm for nonsymmetric exponential-cone optimization*. Tech. rep. MOSEK ApS, Copenhagen, Denmark, 2019.
- [34] G. Das. “Approximation schemes in computational geometry”. PhD thesis. University of Wisconsin-Madison, 2002.
- [35] G. Das and M. Goodrich. “On the complexity of optimization problems for 3-dimensional convex polyhedra and decision trees”. In: *Computational Geometry* 8.3 (1997), pp. 123–137.
- [36] G. Das and D. Joseph. “The complexity of minimum convex nested polyhedra”. In: *Proceedings of the 2nd Canadian Conference on Computational Geometry*. 1990, pp. 296–301.
- [37] J. Dewez and F. Glineur. “Lower bounds on the nonnegative rank using a nested polytopes formulation”. In: *ESANN 2020 28th European Symposium on Artificial Neural Networks-Computational Intelligence and Machine Learning*. 2020.
- [38] C. Ding, T. Li, and W. Peng. “On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing”. In: *Computational Statistics & Data Analysis* 52.8 (2008), pp. 3913–3927.
- [39] J. Eggert and E. Korner. “Sparse coding and NMF”. In: *IEEE International Joint Conference on Neural Networks*. Vol. 4. 2004, 2529–2533 vol.4.

- [40] M. T. Eismann. “Resolution enhancement of hyperspectral imagery using maximum a posteriori estimation with a stochastic mixing model”. PhD thesis. Univ. Dayton, Dayton, OH, 2004.
- [41] M. Fazel. “Matrix rank minimization with applications”. PhD thesis. Stanford University, 2002.
- [42] M. Fazel, H. Hindi, and S. P. Boyd. “Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices”. In: *Proceedings of the 2003 American Control Conference*. Vol. 3. IEEE. 2003, pp. 2156–2162.
- [43] C. Févotte, N. Bertin, and J.-L. Durrieu. “Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis.” In: *Neural Comput* 21.3 (2009), pp. 793–830.
- [44] C. Févotte and N. Dobigeon. “Nonlinear Hyperspectral Unmixing With Robust Nonnegative Matrix Factorization”. In: *IEEE Transactions on Image Processing* 24.12 (2015), pp. 4810–4819.
- [45] C. Févotte and J. Idier. “Algorithms for nonnegative matrix factorization with the  $\beta$ -divergence”. In: *Neural computation* 23.9 (2011), pp. 2421–2456.
- [46] S. Fiorini, V. Kaibel, K. Pashkovich, and D. O. Theis. “Combinatorial bounds on nonnegative rank and extended formulations”. In: *Discrete Mathematics* 313.1 (2013), pp. 67–83.
- [47] *F.T. Tamas Janos, Geoinformatics (2008)*. [http://www.tankonyvtar.hu/en/tartalom/tamop425/0032\\_terinformatika/index.html](http://www.tankonyvtar.hu/en/tartalom/tamop425/0032_terinformatika/index.html). Accessed: 2020-09-05.
- [48] X. Fu, K. Huang, and N. D. Sidiropoulos. “On identifiability of nonnegative matrix factorization”. In: *IEEE Signal Processing Letters* 25.3 (2018), pp. 328–332.
- [49] X. Fu, K. Huang, and N. D. Sidiropoulos. “On Identifiability of Nonnegative Matrix Factorization”. In: *IEEE Signal Processing Letters* 25.3 (2018), pp. 328–332.
- [50] X. Fu, K. Huang, N. D. Sidiropoulos, and W. Ma. “Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications”. In: *IEEE Signal Processing Magazine* 36.2 (2019), pp. 59–80.
- [51] X. Fu, K. Huang, N. D. Sidiropoulos, Q. Shi, and M. Hong. “Anchor-free correlated topic modeling”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.5 (2019), pp. 1056–1071.
- [52] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N.-D. Sidiropoulos. “Robust Volume Minimization-Based Matrix Factorization for Remote Sensing and Document Clustering”. In: *IEEE Transactions on Signal Processing* 64.23 (2016), pp. 6254–6268.
- [53] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos. “Blind Separation of Quasi-Stationary Sources: Exploiting Convex Geometry in Covariance Domain.” In: *IEEE Transactions Signal Processing* 63.9 (2015), pp. 2306–2320.

- [54] A.R. Gillespie, A.B. Kahle, and R.E. Walker. “Color enhancement of highly correlated images—II Channel ratio and ‘chromacity’ transformation techniques”. In: *Remote Sens. Environ.* 22.3 (1987), pp. 343–365.
- [55] N. Gillis. “Introduction to Nonnegative Matrix Factorization”. In: *SIAG/OPT Views and News* 25.1 (2017), pp. 7–16.
- [56] N. Gillis. *Nonnegative Matrix Factorization*. SIAM, Philadelphia, to appear.
- [57] N. Gillis. “Successive nonnegative projection algorithm for robust nonnegative blind source separation”. In: *SIAM Journal on Imaging Sciences* 7.2 (2014), pp. 1420–1450.
- [58] N. Gillis. “The why and how of nonnegative matrix factorization”. In: *Regularization, Optimization, Kernels, and Support Vector Machines*. Ed. by J.A.K. Suykens, M. Signoretto, and A. Argyriou. Machine Learning and Pattern Recognition. Boca Raton, Florida: Chapman & Hall/CRC, 2014. Chap. 12, pp. 257–291.
- [59] N. Gillis and F. Glineur. “A continuous characterization of the maximum-edge biclique problem”. In: *J. of Global Optim* 58 (2014), pp. 439–464.
- [60] N. Gillis and F. Glineur. “Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization”. In: *Neural Computation* 24.4 (2012), pp. 1085–1105.
- [61] N. Gillis and F. Glineur. “On the geometric interpretation of the nonnegative rank”. In: *Linear Algebra and its Applications* 437.11 (2012), pp. 2685–2712.
- [62] N. Gillis, L. T. K. H. Hien, V. Leplat, and V. Y. F. Tan. *Distributionally Robust and Multi-Objective Nonnegative Matrix Factorization*. 2019. arXiv: 1901.10757 [cs.LG].
- [63] N. Gillis, D. Kuang, and H. Park. “Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization”. In: *IEEE Transactions on Geoscience and Remote Sensing* 53.4 (2015), pp. 2066–2078.
- [64] N. Gillis and Y. Shitov. “Low-rank matrix approximation in the infinity norm”. In: *Linear Algebra and its Applications* 581 (2019), pp. 367–382.
- [65] N. Gillis and S. A. Vavasis. “Fast and robust recursive algorithms for separable nonnegative matrix factorization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.4 (2014), pp. 698–714.
- [66] N. Gillis and S. A. Vavasis. “Fast and Robust Recursive Algorithms for Separable Nonnegative Matrix Factorization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.4 (2014), pp. 698–714.
- [67] G.H. Golub and C Van Loan. *Matrix Computations, 4th edn*. The Johns Hopkins University Press, 2013.

- [68] L. Grippo and M. Sciandrone. “On the convergence of the block nonlinear Gauss–Seidel method under convex constraints”. In: *Operations Research Letters* 26.3 (2000), pp. 127–136.
- [69] N. Guan, D. Tao, Z. Luo, and B. Yuan. “NeNMF: An optimal gradient method for nonnegative matrix factorization”. In: *IEEE Transactions on Signal Processing* 60.6 (2012), pp. 2882–2898.
- [70] M. Hong, M. Razaviyayn, Z. Luo, and J. Pang. “A Unified Algorithmic Framework for Block-Structured Optimization Involving Big Data: With applications in machine learning and signal processing”. In: *IEEE Signal Processing Magazine* 33.1 (2016), pp. 57–77.
- [71] R. A Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 1985.
- [72] P.O. Hoyer. “Non-Negative Matrix Factorization with Sparseness Constraints”. In: *J. Mach. Learn. Res.* 5 (Dec. 2004), 1457–1469. ISSN: 1532-4435.
- [73] K. Huang and N. D. Sidiropoulos. “Putting nonnegative matrix factorization to the test: a tutorial derivation of pertinent cramer—rao bounds and performance benchmarking”. In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 76–86.
- [74] K. Huang, N. D. Sidiropoulos, and A. P. Liavas. “A Flexible and Efficient Algorithmic Framework for Constrained Matrix and Tensor Factorization”. In: *IEEE Transactions on Signal Processing* 64.19 (2016), pp. 5052–5065.
- [75] K. Huang, N. D. Sidiropoulos, and A. Swami. “Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition”. In: *IEEE Transactions on Signal Processing* 62.1 (2014), pp. 211–224.
- [76] I. Joliffe. *Principal Component Analysis*. Springer, 2011.
- [77] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma. “Hyperspectral Super-Resolution: A Coupled Tensor Factorization Approach”. In: *IEEE Transactions on Signal Processing* 66.24 (2018), pp. 6503–6517.
- [78] Q. Ke and T. Kanade. “Robust  $L_{1/2}$  norm factorization in the presence of outliers and missing data by alternative convex programming”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. 2005, 739–746 vol. 1.
- [79] J. Kim, Y. He, and H. Park. “Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework”. In: *Journal of Global Optimization* 58 (2014), 285–319.
- [80] R. Krone and K. Kubjas. *Uniqueness of nonnegative matrix factorizations by rigidity theory*. 2020. arXiv: 1902.02868 [math.AG].
- [81] D. Kuang, C. Ding, and H. Park. “Symmetric Nonnegative Matrix Factorization for Graph Clustering”. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 106–117. DOI: 10.1137/1.9781611972825.10.



- [82] B. Lakshminarayanan and R. Raich. “Non-negative matrix factorization for parameter estimation in hidden Markov models”. In: *2010 IEEE International Workshop on Machine Learning for Signal Processing*. 2010, pp. 89–94.
- [83] C. Lanaras, E. Baltsavias, and K. Schindler. “Hyperspectral super-resolution by coupled spectral unmixing”. In: *IEEE Int. Conf. Computer Vision*. IEEE. 2015, 3586–3594.
- [84] 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).
- [85] D. Lee and H. Seung. “Algorithms for non-negative matrix factorization”. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*. NIPS. MIT Press Cambridge, 2000, pp. 535–541.
- [86] D. Lee and H.-S. Seung. “Algorithms for non-negative matrix factorization”. In: *Advances in neural information processing systems*. 2001, pp. 556–562.
- [87] D. D. Lee and H. S. Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), pp. 788–791.
- [88] A. Lefèvre. “Méthode d’apprentissage de dictionnaire pour la séparation de sources audio avec un seul capteur”. PhD thesis. Ecole Normale Supérieure de Cachan, 2012.
- [89] V. Leplat, A.M.S. Ang, and N. Gillis. “Minimum-volume Rank-deficient Nonnegative Matrix Factorizations”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3402–3406.
- [90] V. Leplat, N. Gillis, and A. M. S. Ang. “Blind Audio Source Separation With Minimum-Volume Beta-Divergence NMF”. In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 3400–3410.
- [91] V. Leplat, N. Gillis, and C. Févotte. *Multi-Resolution Beta-Divergence NMF for Blind Spectral Unmixing*. 2020. arXiv: 2007.03893 [eess.SP].
- [92] V. Leplat, N. Gillis, and J. Idier. *Multiplicative Updates for NMF with  $\beta$ -Divergences under Disjoint Equality Constraints*. 2020. arXiv: 2010.16223 [cs.LG].
- [93] V. Leplat, N. Gillis, X. Siebert, and A.M.S. Ang. “Séparation aveugle de sources sonores par factorisation en matrices positives avec pénalité sur le volume du dictionnaire”. In: *XXVIIeme Colloque francophone de traitement du signal et des images*. GRETSI. 2019.
- [94] C. Lin. “Projected Gradient Methods for Nonnegative Matrix Factorization”. In: *Neural Computation* 19.10 (2007), pp. 2756–2779.

- [95] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi. “Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case”. In: *IEEE Transactions on Geoscience and Remote Sensing* 53.10 (2015), pp. 5530–5546.
- [96] J. G. Liu. “Smoothing filter-based intensity modulation: a spectral preserve image fusion technique for improving spatial details”. In: *Int. J. Remote Sens.* 21.18 (2000), pp. 3461–3472.
- [97] L. Loncan, L. B. De Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G.A. Licciardi, and M. Simoes. “Hyperspectral pansharpening: A review”. In: *IEEE Geosci. Remote Sens. Mag.* 3.4 (2015), 27–46.
- [98] W.-K. Ma, J.M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A.-J. Plaza, A. Ambikapathi, and C.-Y. Chi. “A signal processing perspective on hyperspectral unmixing: Insights from remote sensing”. In: *IEEE Signal Processing Magazine* 31.1 (2014), pp. 67–81.
- [99] P. Magron. “Reconstruction de phase par modèles de signaux : application à la séparation de sources audio”. PhD thesis. TELECOM ParisTech, 2016.
- [100] X. Mao, P. Sarkar, and D. Chakrabarti. “On Mixed Memberships and Symmetric Nonnegative Matrix Factorizations”. In: ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia, 2017, pp. 2324–2333.
- [101] L. Miao and H. Qi. “Endmember Extraction From Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization”. In: *IEEE Transactions on Geoscience and Remote Sensing* 45.3 (2007), pp. 765–777.
- [102] L. Miao and H. Qi. “Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization”. In: *IEEE Transactions on Geoscience and Remote Sensing* 45.3 (2007), pp. 765–777.
- [103] K. Miller and G. Samko. “Completely monotonic functions”. In: *Integral Transforms and Special Functions* 12.3 (2001), 389–402.
- [104] D. Mond, J. Smith, and D. van Straten. “Stochastic factorizations, sandwiched simplices and the topology of the space of explanations”. In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 459.2039 (2003), pp. 2821–2845.
- [105] J.M.P. Nascimento and J. Bioucas-Dias. “Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors”. In: *Remote Sens. Environ.* 113.5 (2009), pp. 893–903.
- [106] J.M.P. Nascimento and J. Bioucas-Dias. “Vertex component analysis: A fast algorithm to unmix hyperspectral data”. In: *IEEE Trans. Geosci. Remote Sens.* 43.4 (2005), 898–910.

- [107] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.
- [108] Y. Nesterov. *Lectures on Convex Optimization, second edition*. Vol. 137. Springer International Publishing, 2018.
- [109] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994. DOI: 10.1137/1.9781611970791. eprint: <https://epubs.siam.org/doi/pdf/10.1137/1.9781611970791>. URL: <https://epubs.siam.org/doi/abs/10.1137/1.9781611970791>.
- [110] R. Nishii, S. Kusanobu, and S. Tanaka. “Enhancement of low spatial resolution image based on high resolution bands”. In: *IEEE Transactions on Geoscience and Remote Sensing* 34.5 (1996), pp. 1151–1158.
- [111] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. New York, NY: Academic Press, 1970.
- [112] P. Paatero and U. Tapper. “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values”. In: *Environmetrics* 5.2 (1994), pp. 111–126.
- [113] R. Peeters. “The maximum edge biclique problem is NP-complete”. In: *Discrete Applied Mathematics* 131.3 (2003), pp. 651–654.
- [114] J.C. Price. “Combining panchromatic and multispectral imagery from dual resolution satellite instruments”. In: *Remote Sens. Environ.* 21.2 (1987), pp. 119–128.
- [115] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly. “Hyperspectral Unmixing via  $L_{1/2}$  Sparsity-Constrained Nonnegative Matrix Factorization”. In: *IEEE Transactions on Geoscience and Remote Sensing* 49.11 (2011), pp. 4282–4297.
- [116] T. Ranchin and L. Wald. “Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation”. In: *Photogramm. Eng. Remote Sens.* 66.1 (2000), pp. 49–61.
- [117] M. Razaviyayn, M. Hong, and Z.-Q. Luo. “A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization”. In: *SIAM Journal on Optimization* 23.2 (2013), pp. 1126–1153.
- [118] E. Rechtschaffen. “Real roots of cubics: explicit formula for quasi-solutions”. In: *The Mathematical Gazette* 524 (2008), 268–276.
- [119] J. Le Roux, F. J. Weninger, and J. R. Hershey. *Sparse NMF – half-baked or well done?* Tech. rep. Mitsubishi Electric Research Laboratories (MERL), 2015.
- [120] A. Schneider and H. Feussner. “Chapter 5 - Diagnostic Procedures”. In: *Biomedical Engineering in Gastrointestinal Surgery*. Academic Press, 2017, pp. 87–220.

- [121] Y. Shitov. “The Nonnegative Rank of a Matrix: Hard Problems, Easy Solutions”. In: *SIAM Review* 59.4 (2017), pp. 794–800.
- [122] Silio. “An Efficient Simplex Coverability Algorithm in E2 with Application to Stochastic Sequential Machines”. In: *IEEE Transactions on Computers* C-28.2 (1979), pp. 109–120.
- [123] M. Simoes, J. Bioucas-Dias, L. Almeida, and J. Chanussot. “A convex formulation for hyperspectral image superresolution via subspace-based regularization”. In: *IEEE Trans. Geosci. Remote Sens.* 53.6 (2015), pp. 3373–3388.
- [124] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot. “A convex formulation for hyperspectral image superresolution via subspace-based regularization”. In: *IEEE Trans. Geosci. Remote Sens.* 5.2 (2015), pp. 3373–3388.
- [125] P. Smaragdis. “Convolutional Speech Bases and Their Application to Supervised Speech Separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1 (2007), pp. 1–12.
- [126] P. Smaragdis, C. Févotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman. “Static and Dynamic Source Separation Using Nonnegative Factorizations: A unified view”. In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 66–75.
- [127] O. Stojanovic and G. Pipa. “Predicting epileptic seizures using nonnegative matrix factorization”. In: *medRxiv* (2019). DOI: 10.1101/19000430.
- [128] H. Su, Q. Du, and P. Du. “Hyperspectral Image Visualization Using Band Selection”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7.6 (2014), pp. 2647–2658.
- [129] Y. Sun, P. Babu, and D. P. Palomar. “Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning”. In: *IEEE Transactions on Signal Processing* 65.3 (2017), pp. 794–816.
- [130] Y. Sun, P. Babu, and D.P. Palomar. “Majorization-minimization algorithms in signal processing, communications, and machine learning”. In: *IEEE Transactions on Signal Processing* 65.3 (2017), pp. 794–816.
- [131] L. B. Thomas. “Rank Factorization of Nonnegative Matrices (A. Berman)”. In: *SIAM Review* 16.3 (1974), pp. 393–394.
- [132] M. Udell and A. Townsend. “Why Are Big Data Matrices Approximately Low Rank?” In: *SIAM Journal on Mathematics of Data Science* 1.1 (2019), pp. 144–160.
- [133] A. Vandaele, N. Gillis, F. Glineur, and D. Tuytens. “Heuristics for exact nonnegative matrix factorization”. In: *J. of Global Optim* 65 (2016), 369–400.
- [134] G. Vane, R.O. Green, T.G. Chrien, H.T. Enmark, E.G. Hansen, and W.M. Porter. “The airborne visible/infrared imaging spectrometer (AVIRIS)”. In: *Remote Sens. Environ.* 44.2–3 (1993), pp. 127–143.

- [135] S. A. Vavasis. “On the complexity of nonnegative matrix factorization”. In: *SIAM Journal on Optimization* 20.3 (2010), pp. 1364–1377.
- [136] E. Vincent, R. Gribonval, and C. Févotte. “Performance Measurement in Blind Audio Source Separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1462–1469.
- [137] T. Virtanen. “Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007), pp. 1066–1074.
- [138] L. Wald. “Quality of high resolution synthesised images: Is there a simple criterion?” In: *Int. Conf. Fusion Earth Data*. 2000, pp. 99–105.
- [139] L. Wald. “Quality of high resolution synthesised images: Is there a simple criterion?” In: *Int. Conf. Fusion Earth Data*. 2000, 99–103.
- [140] Z. Wang and A.C. Bovik. “A universal image quality index”. In: *IEEE Signal Process. Lett.* 9.3 (2002), 81–84.
- [141] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J. Tourneret. “Hyperspectral and Multispectral Image Fusion Based on a Sparse Representation”. In: *IEEE Transactions on Geoscience and Remote Sensing* 53.7 (2015), pp. 3658–3668.
- [142] Q. Wei, J. Bioucas-Dias, N. Dobigeon, J.-Y. Tourneret, M. Chen, and S. Godsill. “Multiband Image Fusion Based on Spectral Unmixing”. In: *IEEE Transactions on Geoscience and Remote Sensing* 54.12 (2016), pp. 7236–7249.
- [143] Y. Xu and W. Yin. “A Block Coordinate Descent Method for Regularized Multi-convex Optimization with Applications to Nonnegative Tensor Factorization and Completion”. In: *SIAM Journal on Imaging Sciences* 6.3 (2013), pp. 1758–1789.
- [144] Z. Yang, J. Corander, and E. Oja. “Low-rank doubly stochastic matrix decomposition for cluster analysis”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 6454–6478.
- [145] S. Yaroslav. “Nonnegative rank depends on the field”. In: *Math. Program.* (2019).
- [146] S. Yaroslav. “On the complexity of Boolean matrix ranks”. In: *Linear Algebra and its Applications* 439.8 (2013), pp. 2500–2502.
- [147] N. Yokoya, C. Grohnfeldt, and J. Chanussot. “Hyperspectral and multispectral data fusion: a comparative review of the recent literature”. In: *IEEE Geoscience and Remote Sensing Magazine* 5.2 (2017), pp. 29–56.
- [148] N. Yokoya, N. Mayumi, and A. Iwasaki. “Cross-calibration for data fusion of EO-1/Hyperion and Terra/ASTER”. In: *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* 6.2 (2013), pp. 419–426.

- 
- [149] N. Yokoya, T. Yairi, and A. Iwasaki. “Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion”. In: *IEEE Transactions on Geoscience and Remote Sensing* 50.2 (2012), pp. 528–537.
- [150] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto. “Beyond NMF: Time-Domain Audio Source Separation without Phase Reconstruction”. In: *ISMIR*. 2013, pp. 369–374.
- [151] A. L. Yuille and A. Rangarajan. “The Concave-Convex Procedure”. In: *Neural Computation* 15.4 (2003), pp. 915–936. DOI: 10.1162/08997660360581958.
- [152] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie. “Nonnegative Matrix and Tensor Factorizations : An algorithmic perspective”. In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 54–65.
- [153] G. Zhou, S. Xie, Z. Yang, J.-M. Yang, and Z. He. “Minimum-volume-constrained nonnegative matrix factorization: Enhanced ability of learning parts”. In: *IEEE Transactions on Neural Networks* 22.10 (2011), pp. 1626–1637.
- [154] F. Zhu. *Hyperspectral Unmixing: Ground Truth Labeling, Datasets, Benchmark Performances and Survey*. 2017. arXiv: 1708.05125 [cs.CV].

# Appendix

## 1 Symbols

### Scalars, Vectors, Matrices

$\mathbb{R}, \mathbb{R}_+, \mathbb{R}_{++}$	The set of real, nonnegative and positive real numbers.
$\mathbb{R}^F, \mathbb{R}_+^F$	The set of real and nonnegative $F$ -vectors.
$\mathbb{R}^{F \times N}, \mathbb{R}_+^{F \times N}$	The set of real and nonnegative matrices of dimension $F$ -by- $N$ .

### Norms

$\ \cdot\ _1$	$\ell_1$ -norm, $\ x\ _1 = \sum_{i=1}^n  x_i $ , $x \in \mathbb{R}^n$
$\ \cdot\ _2$	vector $\ell_2$ -norm, $\ x\ _2 = \sqrt{\sum_{i=1}^n x_i^2}$ , $x \in \mathbb{R}^n$
	matrix $\ell_2$ -norm, $\ A\ _2 = \max_{x \in \mathbb{R}^N, \ x\ _2=1} \ Ax\ _2$ , $A \in \mathbb{R}^{F \times N}$
$\ \cdot\ _0$	$\ell_0$ -norm, $\ x\ _0 =  \{i   x_i \neq 0\} $ , $x \in \mathbb{R}^n$
$\ \cdot\ _F$	Frobenius norm, $\ A\ _F = \sqrt{\sum_{f=1}^F \sum_{n=1}^N A_{fn}^2}$ , $A \in \mathbb{R}^{F \times N}$

### Operators

$\mathbb{E}(\cdot)$	expected value of a random variable
$\mathcal{I}(\cdot)$	component-wise imaginary part of a vector, or a matrix
$\Pi_{\Delta^F}(\cdot)$	projection of a vector, or the columns of a matrix onto the unit simplex of dimension $F$
$\mathcal{R}(\cdot)$	component-wise real part of a vector, or a matrix
$S(\cdot)$	linear operator for the STFT computation of a vector
$\text{Var}(\cdot)$	variance of a random variable

## Functions on Matrices

$\text{Tr}(\cdot)$	trace of a matrix
$\langle A, B \rangle_F$	Frobenius inner product between two real matrices $A$ and $B$ , $\langle A, B \rangle_F = \sum_{pq} A_{pq} B_{pq} = \text{Tr}(A^T B)$
$\det(\cdot)$	determinant of a matrix
$\sigma_i(\cdot)$	$i^{\text{th}}$ singular value of a matrix
$\text{rank}(\cdot)$	rank of a matrix
$\text{rank}_+(\cdot)$	nonnegative rank of a matrix
$\text{cone}(\cdot)$	cone spanned by the columns of a matrix
$\text{conv}(\cdot)$	convex hull of the columns of a matrix
$\text{col}(\cdot)$	column space of a matrix
$A(f, :)$	$f^{\text{th}}$ row of $A$
$A(:, n)$	$n^{\text{th}}$ column of $A$
$A(f, n)$ or $[A]_{fn}$	entry at position $(f, n)$ of $A$
$A(\mathcal{K})$	submatrix of $A$ with row and column indices in $\mathcal{K}$
$A \odot B$	Hadamard product (component-wise multiplication), $[A \odot B]_{fn} = A_{ij} B_{fn}$
$\frac{[A]}{[B]}$	component-wise division, $\left[ \frac{[A]}{[B]} \right]_{fn} = \frac{A_{fn}}{B_{fn}}$
$(\cdot)^T$	transpose of a matrix, $[A^T]_{fn} = [A]_{nf}$
$A^{(\cdot, \alpha)}$	element-wise $\alpha$ exponent of $A$ , $[A^{(\cdot, \alpha)}]_{fn} = A_{fn}^\alpha$
$\text{vol}(W)$	function that measures the volume of the columns of a matrix,
$(\cdot)^+$	nonnegative part of a matrix, $(A)^+ = \max(0, A)$
$(\cdot)^-$	negative part of a matrix, $(A)^- = \max(0, -A)$
$\log \det(A)$	natural logarithm of the determinant of a matrix

## Sets

$\Delta^F$	unit simplex of dimension $F$
$\mathcal{S}^K$	convex hull of the unit simplex of dimension $K$ and the origin

## Miscellaneous

$e$	vector of all ones of appropriate dimension
$e_{F,N}$	$F$ -by- $N$ all ones matrix
$I_K$	identity matrix of dimension $K \times K$
$a : b$	set $\{a, a + 1, \dots, b - 1, b\}$ (for $a$ and $b$ integers with $a \leq b$ )
$\nabla f$	the gradient of the function $f$
$\nabla^2 f$	the hessian of the function $f$
$\setminus$	subtraction of two sets, $\mathcal{S} \setminus \mathcal{Q}$ is the set of elements in $\mathcal{S}$ and not in $\mathcal{Q}$



## 2 Acronyms

<b>3-SAT</b>	3-Satisfiability.
<b>ALS</b>	Alternating Least Squares.
<b>ANLS</b>	Alternating Nonnegative Least Squares.
<b>AO-ADMM</b>	Alternating Optimization-Alternating Direction for Multiplier.
<b>BCD</b>	Block-Coordinate Descent.
<b>BHU</b>	Blind Hyperspectral Unmixing.
<b>CGP</b>	Conic Geometric Programming.
<b>CP</b>	Conic Programming.
<b>DR-NMF</b>	Distributionally Robust Nonnegative Matrix Factorization.
<b>ERGAS</b>	Erreur Relative Globale Adimensionnelle de Synthèse.
<b>GR-NMF</b>	Group Robust Nonnegative Matrix Factorization.
<b>HALS</b>	Hierarchical Alternating Least Squares.
<b>HSI</b>	Hyperspectral Image.
<b>IPM</b>	Interior Point Methods.
<b>IS</b>	Itakura-Saito.
<b>KL</b>	Kullback-Leibler.
<b>LDR</b>	Linear Dimensionality Reduction.
<b>LP</b>	Linear Programming.
<b>LRMA</b>	Low Rank Matrix Approximation.
<b>LS</b>	Line Search.
<b>min-vol NMF</b>	minimum-volume Nonnegative Matrix Factorization.
<b>MM</b>	Majorization-Minimization.
<b>MOS</b>	Model Order Selection.
<b>MR</b>	Multi Resolution.
<b>MSI</b>	Multispectral Image.
<b>MU</b>	Multiplicative Updates .
<b>NMF</b>	Nonnegative Matrix Factorization.
<b>NPP</b>	Nested Polytope Problem.
<b>PCA</b>	Principal Component Analysis.
<b>PFGM</b>	Projected Fast Gradient Method.
<b>RE-NMF</b>	Restricted Exact Nonnegative Matrix Factorization.
<b>RMSE</b>	Root-Mean-Square Error.
<b>SAM</b>	Spectral Angle Mapper.
<b>SAR</b>	Signal to Artefacts Ratio.
<b>SCCAE-NMF</b>	Successive Conic Convex Approximation for Exact NMF.
<b>SDP</b>	Semi-Definite Programming.
<b>SDR</b>	Signal to Distortion Ratio.
<b>SIR</b>	Signal to Interference Ratio.
<b>SNPA</b>	Successive Nonnegative Projection Algorithm.

<b>SNR</b>	Signal-to-Noise Ratio.
<b>SOCP</b>	Second-Order Conic Programming.
<b>SPA</b>	Successive Projection Algorithm.
<b>SPI</b>	Sparsity Pattern Integration.
<b>SR</b>	Super Resolution.
<b>SSC</b>	Sufficiently Scattered Condition.
<b>SSMF</b>	Simplex-Structured Matrix Factorization .
<b>SSNMF</b>	Simplex-Structured Nonnegative Matrix Factorization .
<b>STFT</b>	Short Time Fourier transform.
<b>SVD</b>	Singular Value Decomposition.
<b>UIQI</b>	Universal Image Quality Index.

### 3 A brief introduction to convergence theory of popular BCD schemes

In this section we give a brief introduction to the convergence theory of popular BCD schemes used to tackle NMF optimization problems. Let recall that Algorithm 1 from Section 4.3 presents the general structure of the BCD schemes. First, we develop the first-order optimality conditions of standard NMF problem 1.4.1, then we present the Majorization-Minimization framework which is used to tackle each subproblem of a BCD scheme and finally we cite recent results about the convergence theory of popular BCD schemes.

#### First-order optimality condition for standard NMF problems

As explained several times in the thesis, the standard NMF problem 1.4.1 is symmetric in variables  $W$  and  $H$ , as long as  $D(V|WH) = D(V^T|H^T W^T)$  which holds for most error measures. Then we focus at the subproblem in  $H$  defined as follows:

$$\begin{aligned} \min_{H \in \mathbb{R}^{K \times N}} \quad & D(V|WH) = \sum_{fn} d(V_{fn}|[WH]_{fn}) \\ \text{subject to} \quad & H \geq 0, \end{aligned} \tag{1}$$

For the following, we assume that the objective function  $D(V|WH)$  is differentiable. Let us denote by  $\nabla_H D(V|WH) \in \mathbb{R}^{K \times N}$  the gradient of  $D(V|WH)$  w.r.t. matrix  $H$  such that:

$$[\nabla_H D(V|WH)]_{k,n} = \frac{\partial D(V|WH)}{\partial H_{k,n}}$$

In the case we choose for the scalar divergence  $d(V_{fn}|[WH]_{fn}) = d_\beta(V_{fn}|[WH]_{fn})$  (see Section 1.9.1), then the gradient  $\nabla_H D(V|WH)$  in matrix form is as follows:

$$\nabla_H D(V|WH) = W^T \left( (WH)^{(\beta-2)} \odot (WH - V) \right)$$

The point  $(W, H)$  is called a (first-order) stationary point of problem (1) if it satisfies the first-order optimality conditions, also known as the Karush-Kuhn-Tucker (KKT) conditions, given here-under:

$$H \geq 0, \nabla_H D(V|WH) \geq 0, \nabla_H D(V|WH) \odot H = 0_{K \times N},$$

where  $0_{K \times N}$  is matrix full of zeros of size  $K \times N$ . The last condition  $\nabla_H D(V|WH) \odot H = 0_{K \times N}$  implies that for all  $k, n$  we have

$$H_{k,n} = 0 \text{ or } [\nabla_H D(V|WH)]_{k,n} = 0,$$

which are the complementary slackness conditions. Most NMF algorithms are first-order methods, meaning that they only use the information from the gradient to find the minimizer of (1) (and similarly for  $W$ ). For the first-order methods, only convergence to stationary points can be achieved (such methods are stuck at stationary points). A stationary point of a differentiable function is either a local minimum, a local maximum or a saddle point, which is a point for which there exists a direction in which the objective function decreases and a direction in which the objective function increases, hence the term "saddle". As explained in Section 1.10, the large majority of NMF algorithms resort to BCD schemes, the variables  $W$  and  $H$  are updated alternatively until a stationary point of the standard NMF problem 1.4.1 is reached. Because  $D(V|WH)$  is jointly non-convex in  $W$  and  $H$ , the stationary point may be not a global minimum (and possibly not even a local minimum). In the next section, we present a standard approach, referred to as Majorization-Minimization, to update  $W$  and  $H$  at each iteration of the BCD scheme.

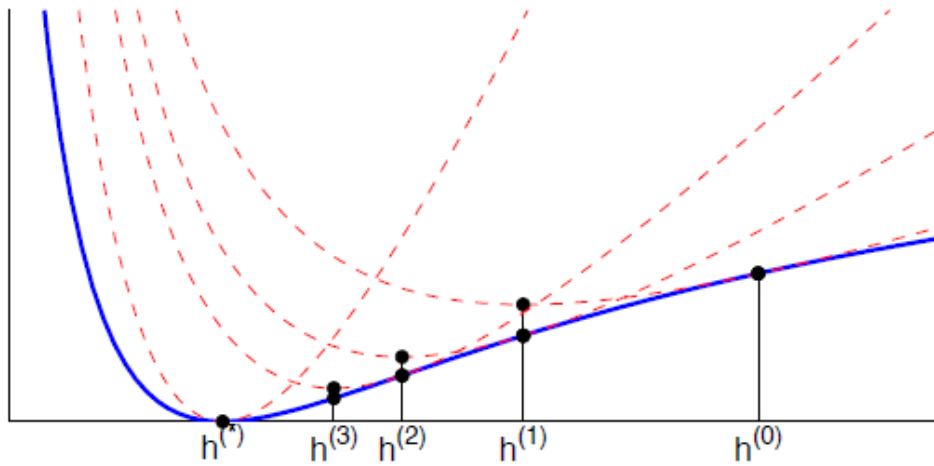
## Majorization-Minimization framework

In this section, we give a brief overview to the Majorization-Minimization (MM) framework for the minimization of the standard NMF problem 1.4.1. Generally speaking, MM consists in minimizing iteratively an easier-to-minimize tight upper bound of the original objective function in (1). Let us denote by  $\tilde{H}$  the current iterate and by  $C(H)$  the objective function for the supproblem in  $H$  (1), at each iteration, MM includes two steps:

1. "Majorization": we build an upper bound  $G(H|\tilde{H})$  of  $C(H)$  which is tight for  $H = \tilde{H}$ . In others words, were are looking for  $G(H|\tilde{H})$  such that:
  - a)  $G(H|\tilde{H}) \geq C(H)$  for all  $H \geq 0$ ,
  - b)  $G(\tilde{H}|\tilde{H}) = C(\tilde{H})$ .

The function  $G(H|\tilde{H})$  is called an auxiliary function of  $C$  at  $\tilde{H}$ .

2. "Minimization": we minimize the auxiliary function w.r.t.  $H$  to derive a valid descent algorithm. MM ensures the decreasingness of the objective function at each iteration  $i$  since  $C(H^{(i)}) \leq G(H^{(i)}|H^{(i-1)}) \leq G(H^{(i-1)}|H^{(i-1)}) = C(H^{(i-1)})$ .



**Fig. 1.** Illustration of the MM principle on a 1-dimensional problem. For a current solution, the MM approach consists in minimizing an auxiliary function (dashed red curves) to the objective function  $C$  (blue curve) build at each iterate. The minimizer of the current auxiliary function is used to build the next auxiliary function, and so on until convergence. This figure has been reproduced from [45].

The principle of MM is illustrated in Figure 1. Typically, we build  $G$  such that it has convenient properties:

1.  $G(H|\tilde{H})$  is convex,
2. the minimizer of  $G(H|\tilde{H})$  can be computed in closed form. A common manner to ensure this property is to design  $G$  as a separable function, that is,

$$G(H|\tilde{H}) = \sum_{k,n} G_{k,n}(H_{k,n}|\tilde{H}),$$

for some functions  $G_{k,n}$  so that minimizing  $G(H|\tilde{H})$  requires solving  $K \times N$  independent univariate subproblems.

The main challenge for designing MM algorithms is the construction of the function  $G$  which should be a good approximation of  $C$  while being easy to optimize thanks to the aforementioned properties. For the NMF problems that include  $\beta$ -divergences as objective functions, [45] introduces a powerful framework to design auxiliary functions for  $C(H)$ . The trick is to decompose  $C(H)$  into the sum of a convex part and a concave part and to upper-bound each part separately. The convex part is upper-bounded by using Jensen's inequality and the concave part is upper-bounded using the tangent inequality. The two upper-bounds are finally summed and the resulting convex auxiliary function turns out to have a closed form minimizer, see [45] for more details.

## Convergence to stationary points

In this section, we present a few convergence results for BCD schemes. We only focus on presenting general results that concern the convergence of some of the methods presented in this thesis. Most of the algorithms discussed in this thesis are monotonically decreasing the objective function which is bounded below. As explained in [56], this implies the convergence of the objective function values

$$C^{(i)} = D(V|W^{(i)}H^{(i)}) \text{ for } i = 1, 2, 3, \dots$$

where  $(W^{(i)}, H^{(i)})$  is the  $i$ -th iterate produced by the algorithm. In the case the feasible set is compact, Bolzano-Weierstrass theorem ensures that there exists at least one converging subsequence of the iterates. However,

1. because of the scaling ambiguities of an NMF we may have that  $W^{(i)}H^{(i)}$  converge but not  $(W^{(i)}, H^{(i)})$ .
2. The feasible set of the standard NMF problem, namely  $\mathcal{S} = \{(W, H) | W \geq 0, H \geq 0\}$  is not compact since it is not bounded.

In [56], the author explains that fixing the scaling degree of freedom by adding constraints such as  $\|W(:, k)\|_1 = 1$  for all  $k$  can be used to guarantee compactness.

In many cases, BCD schemes fail to converge rapidly when they get close to stationary points of  $C$ , typically because of their zigzagging behavior (same behaviour can be observed for gradient-descent methods). However, when the blocks of coordinates are rather large and can be optimized efficiently (possibly up to global optimality), BCD schemes have shown to be a powerful technique. At least, they often exhibit a relatively fast initial convergence to the neighborhood of a stationary point.

It is sometimes possible to perform an exact BCD descent, in others words, the subproblems in  $W$  and  $H$  (steps 4 and 6 of Algorithm 1) are solved exactly, that is, an optimal solution is used for  $W^{(i)}$  and  $H^{(i)}$ . In that particular case, we have the following convergence guarantees:

**Theorem 3.1.** [18, Proposition 2.7.1] *The limit points of the iterates of an exact BCD algorithm are stationary points provided that the following two conditions hold:*

1. *each block of variables is required to belong to a closed convex set,*
2. *the minimum computed at each iteration for a given block of variables is uniquely attained.*

**Theorem 3.2.** [68, Corollary 2] *The limit points of the iterates of an exact two-BCD algorithm are stationary points provided that the following two conditions hold:*

1. *the objective function is continuously differentiable, and*
2. *each block of variables is required to belong to a closed convex set.*

Hence exact two-BCD does not require the minimum of the subproblems to be uniquely attained to guarantee convergence to a stationary point. For NMF problems and Algorithm 1, the second condition is satisfied since the nonnegative orthant is a closed convex set. The first condition is met for many objective functions such as the Frobenius norm. However, the condition is not met by  $\beta$ -divergences. Indeed, for  $x > 0$ :

$$\frac{\partial d_\beta(x, y)}{\partial y} = y^{\beta-1} - xy^{\beta-2}$$

Table 1 extracted from [56] provides the domain of  $\frac{\partial d_\beta(x, y)}{\partial y}$  depending on the values of  $x$  and  $\beta$ .

Table 1: domain of  $\frac{\partial d_\beta(x, y)}{\partial y}$  depending on the values of  $x$  and  $\beta$

	$\beta \leq 0$	$\beta \in (0, 1)$	$\beta \in [1, 2)$	$\beta \geq 2$
$x = 0$	$\emptyset$	$\mathbb{R}_{++}$	$\mathbb{R}_+$	$\mathbb{R}_+$
$x > 0$	$\mathbb{R}_{++}$	$\mathbb{R}_{++}$	$\mathbb{R}_{++}$	$\mathbb{R}_+$

Hence, for  $\beta < 1$ ,  $d_\beta(x, y)$  is not continuously differentiable at zero. Moreover, for  $\beta < 2$ , the derivative of  $d_\beta(x, y)$  w.r.t.  $y$  is not defined at zero when  $x > 0$ .

In the case MM framework is used to tackle each subproblem in  $W$  and  $H$  (steps 4 and 6 of Algorithm 1), convergence guarantee has been established by Razaviyayn et al. [117]. The authors introduce the so-called Block Successive Upper-bound Minimization (BSUM) framework. In this framework, for each block of variables, a majorizer is constructed which has additional properties than in the MM framework. In the MM framework, the majorizer is a global upper bound of the objective function which is tight at the current iterate. On top of that, BSUM requires that the directional derivatives of the majorizers and of the objective function coincide at the current iterate for each block of variables (intuitively, their tangent need to exist and coincide), while the majorizers should be continuous functions in all the variables. Moreover, in the BSUM framework, the majorizers are minimized exactly, while an essentially cyclic block update is used. We refer the reader to [117] for more details.

**Theorem 3.3.** [117, Theorem 2] *Convergence of BSUM can be guaranteed in the following two scenarios:*

1. *If the majorizers are quasi-convex<sup>1</sup> and their minimum is uniquely attained, then every limit point of the sequence of iterates generated by BSUM is a stationary point.*
2. *If the level sets of the majorizers are compact and the subproblems have a unique solution for all blocks but one, then the sequence of iterates generated by BSUM converges to the set of stationary points.*

<sup>1</sup>A function  $g$  is quasi-convex if the level sets  $\{x|g(x) \leq c\}$  are convex, for any constant  $c$ .

For interesting applications of Theorem 3.3, we refer the reader to [56] and to the survey [70]. Let briefly mention that in the case we use the framework proposed by [45] to tackle the subproblems in  $W$  and  $H$ , it is possible to have the convergence results of Theorem 3.3 by using a small lower bound for the entries of  $W$  and  $H$ , and by considering the following modified NMF problem:

$$\begin{aligned} \min_{W \in \mathbb{R}^{F \times K}, H \in \mathbb{R}^{K \times N}} D(V|WH) &= \sum_{f_n} d(V_{f_n} | [WH]_{f_n}) \\ \text{subject to} \quad W &\geq \epsilon, H \geq \epsilon, \end{aligned} \quad (2)$$

where  $\epsilon \geq 0$  is a parameter. Using this modification, [56] obtains the following results:

**Theorem 3.4.** [56, Theorem 8.9] *Let  $\epsilon > 0$  and let us define the modified update for  $H$  as:*

$$H \leftarrow \max \left( \epsilon, H \odot \left( \frac{\left[ W^T \left( (WH)^{\cdot(\beta-2)} \odot V \right) \right]^{\cdot\gamma(\beta)}}{\left[ W^T (WH)^{\cdot(\beta-1)} \right]} \right) \right)$$

where  $\gamma(\beta)$  is given in Table 4.1, and the update of  $W$  is obtained by symmetry. Then,

- The modified updates do not increase the objective function of (2), given that  $W \geq \epsilon$  and  $H \geq \epsilon$ .
- For any initial matrices  $(W, H)$ , every limit point of the modified updates that alternatively update  $H$  and  $W$  converge to a stationary point of (2).
- For  $\beta \geq 2$ , we may take  $\epsilon = 0$  (that is, the standard MU) and the same convergence result as in the case  $\epsilon > 0$  given that the initial iterate has only positive entries.

## 4 Behaviour of the nonnegative rank under perturbations

### Small continuous perturbations

In [22], the authors study how perturbing a matrix changes its nonnegative rank, they particularly prove that the nonnegative rank can only increase in a neighborhood of a matrix with no zero columns. Mathematically, they demonstrate that the nonnegative rank is lower semicontinuous in the topology given by the Frobenius norm. Let us recall the notion of lower semicontinuous function for a simple real function:

**Definition 4.1.** *A real-valued function  $f(x)$  is lower semicontinuous at a point  $x_0$  if, for any small positive number  $\epsilon$ , there exists  $r(\epsilon)$  such that  $f(x) > f(x_0) - \epsilon$  for all  $x \in B(x_0, r(\epsilon))$  where  $B(x_0, r(\epsilon))$  designates a ball centered at  $x_0$  of radius  $r(\epsilon)$ .*

Roughly speaking, the function values for arguments near  $x_0$  are not much lower than  $f(x_0)$ . In [22], authors use the following topological definition of lower semicontinuity, which is commonly used in algebraic geometry

**Definition 4.2.** A function  $f : \mathbb{R}^N \implies \mathbb{Z}$  is said to be lower semicontinuous if the set

$$\{x \in \mathbb{R}^N \mid f(x) \geq r\}$$

is open in  $\mathbb{R}^N$  for all  $r \in \mathbb{Z}$ .

Given a nonnegative matrix  $V \in \mathbb{R}_+^{F \times N}$  and  $\epsilon > 0$ , we define the ball of center  $V$  and radius  $\epsilon$  as:

$$B(V, \epsilon) = \{U \in \mathbb{R}_+^{F \times N} \mid \|U - V\|_F < \epsilon\}$$

**Theorem 4.1.** [22, Theorem 3.1] Let  $V$  be a  $F$  by  $N$  nonnegative matrix, without zero columns, such that  $\text{rank}_+(V) = K$ ; then there exists a ball  $B(V, \epsilon)$  such that  $\text{rank}_+(N) \geq K$  for all  $N \in B(V, \epsilon)$ .

*Proof.* We refer the reader to [22] for the detailed proof. The key ingredients are based on the geometric interpretation of NMF. More specifically, they consider the problem of finding a nested polytope between an inner and an outer polytope, perturbing slightly the inner and outer polytopes cannot lead to a nested polytope with fewer vertices.  $\square$

### Rank-one perturbations

In this section we show the impact of adding a rank-one matrix to a nonnegative matrix  $V$  onto its nonnegative rank. Let us first recall a well-known property for the rank; given a matrix  $V \in \mathbb{R}^{F \times N}$  and two vectors  $x \in \mathbb{R}^F$  and  $y \in \mathbb{R}^N$ , we have:

$$\text{rank}(V) - 1 \leq \text{rank}(V + xy^T) \leq \text{rank}(V) + 1$$

The question naturally arises; do we have similar intervals for  $\text{rank}_+(V + xy^T)$  for  $V \in \mathbb{R}_+^{F \times N}$  and two vectors  $x \in \mathbb{R}_+^F$  and  $y \in \mathbb{R}_+^N$ ? We can easily derive a similar upper-bound by using the properties of the nonnegative ranks presented in section 6.1.1:

$$\begin{aligned} \text{rank}_+(V + xy^T) &\leq \text{rank}_+(V) + \text{rank}_+(xy^T) \\ &= \text{rank}_+(V) + 1 \end{aligned}$$

In [56], the author shows an interesting result for the lower bound of  $\text{rank}_+(V + xy^T)$ , that is the nonnegative rank of  $V + xy^T$  can be smaller than the nonnegative rank of  $V$  minus one. This result is the consequence the following theorem.

**Theorem 4.2.** [56, Theorem 3.3] For any nonnegative matrix  $V \in \mathbb{R}_+^{F \times N}$ , there exists  $x \in \mathbb{R}_+^F$  and  $y \in \mathbb{R}_+^N$  such that:

$$\text{rank}_+(V + xy^T) = \text{rank}(V)$$

*Sketch of the proof.* Given an input matrix  $V$  whose columns sum to one and such that the rank of  $V$  equal to  $K$ , the proof begins with the construction of a first unconstrained factorization  $V = QP$  with  $Q \in \mathbb{R}^{F \times K}$  and  $P \in \mathbb{R}^{K \times N}$  where  $Q$  is built up by picking  $K$



linearly independent column of  $V$ . The key transformations are  $x = Qe \geq 0$  and  $y = \alpha e$  where  $\alpha = |\min_{kn} P(k, n)|$  hence  $V + xy^T = QP + Qee^T\alpha = Q(P + ee^T\alpha) = QP'$  and  $Q, P' \geq 0$  by construction. Then from an unconstrained factorization, an exact NMF with a factorization rank  $K$  is found allowing to upper-bound  $\text{rank}_+(V + xy^T) \leq K$ . Using the fact that for any matrix the nonnegative rank is lower-bounded by the rank, hence  $\text{rank}_+(V + xy^T) \geq \text{rank}(V + xy^T)$ . They conclude the proof by showing that  $\text{rank}(V + xy^T) = K$ .  $\square$

Let us illustrate the consequence of such theorem by considering a nonnegative matrix whose columns have  $\ell_1$  norm and for which there is a significant gap between the rank and the nonnegative rank. We showcase the results from theorem (4.2) on a simple example extracted from [104]; let  $a > 1$  and  $V_a$  be the matrix:

$$V_a = \frac{1}{6a} \begin{pmatrix} 1 & a & 2a-1 & 2a-1 & a & 1 \\ 1 & 1 & a & 2a-1 & 2a-1 & a \\ a & 1 & 1 & a & 2a-1 & 2a-1 \\ 2a-1 & a & 1 & 1 & a & 2a-1 \\ 2a-1 & 2a-1 & a & 1 & 1 & a \\ a & 2a-1 & 2a-1 & a & 1 & 1 \end{pmatrix}.$$

Matrix  $V_a$  has rank three and for any  $a > 3$ ,  $\text{rank}_+(V_a) = 5$ . Also, the matrix has columns with  $\ell_1$  norm thanks to the scaling factor  $\frac{1}{6a}$ . We choose here-under  $a = 4$  and we first built up an unconstrained exact factorization  $V_4 = QP$  such that

$$Q = \frac{1}{24} \begin{pmatrix} 1 & 4 & 7 \\ 1 & 1 & 4 \\ 4 & 1 & 1 \\ 7 & 4 & 1 \\ 7 & 7 & 4 \\ 4 & 7 & 7 \end{pmatrix},$$

where the columns of  $Q$  correspond to the first three columns of  $V_4$  which are linearly independent. We easily compute  $P$  as follows:

$$P = (Q^T Q)^{-1} Q^T V_4 = \begin{pmatrix} 1 & 0 & 0 & 1 & 2 & 2 \\ 0 & 1 & 0 & -2 & -3 & -2 \\ 0 & 0 & 1 & 2 & 2 & 1 \end{pmatrix}.$$

We have  $\alpha = |\min_{kn} P(k, n)| = 3$ , then we obtain  $x = Qe = (\frac{1}{2} \frac{1}{4} \frac{1}{4} \frac{1}{2} \frac{3}{4} \frac{3}{4})^T$  and  $y = \alpha e =$

$(333333)^T$ . Hence we can compute an exact NMF for  $V_4 + xy^T$  as follows:

$$V_4 + xy^T = QP' \text{ with } P' = P + ee^T\alpha,$$

$$\frac{1}{24} \begin{pmatrix} 37 & 40 & 43 & 43 & 40 & 37 \\ 19 & 19 & 22 & 25 & 25 & 22 \\ 22 & 19 & 19 & 22 & 25 & 25 \\ 43 & 40 & 37 & 37 & 40 & 43 \\ 61 & 61 & 58 & 55 & 55 & 58 \\ 58 & 61 & 61 & 58 & 55 & 55 \end{pmatrix} = \frac{1}{24} \begin{pmatrix} 1 & 4 & 7 \\ 1 & 1 & 4 \\ 4 & 1 & 1 \\ 7 & 4 & 1 \\ 7 & 7 & 4 \\ 4 & 7 & 7 \end{pmatrix} \begin{pmatrix} 4 & 3 & 3 & 4 & 5 & 5 \\ 3 & 4 & 3 & 1 & 0 & 1 \\ 3 & 3 & 4 & 5 & 5 & 4 \end{pmatrix}$$

which shows that  $V_4 + xy^T$  has a nonnegative rank equal to  $\text{rank}(V_4) = 3$  which is smaller than  $\text{rank}_+(V_4)$  minus one. Actually, we can easily prove that, in the case of a matrix  $V$  such that  $\text{rank}_+(V) \geq \text{rank}(V) + d$  with  $d$  an arbitrarily large positive integer, we can reduce at least by  $d$  the nonnegative rank of  $V$  by adding the adequate rank-one matrix. This directly follows Theorem 4.2; since for any nonnegative matrix  $V$  we can find a rank-one matrix  $xy^T$  such that  $\text{rank}_+(V + wy^T) = \text{rank}(V)$ , we can write:

$$\begin{aligned} \text{rank}_+(V) &\geq \text{rank}(V) + d = \text{rank}_+(V + wy^T) + d \\ \iff \text{rank}_+(V + wy^T) &\leq \text{rank}_+(V) - d \end{aligned}$$

This result is enunciated in [56, Corollary 3.5].

## 5 Convexity, concavity and complete monotonicity for a convex-concave decomposition of the discrete $\beta$ -divergence

The discrete  $\beta$ -divergence can always be expressed as the sum of convex, concave, and constant terms. In Table 2 we introduce a convex-concave decomposition of the  $\beta$ -divergence which slightly differ from the one given in [45, Table 1] (by the fact that ours contains no constant term  $\bar{d}$ ) as given in Table 2.

Decomposition $d_\beta = \tilde{d} + \hat{d}$	$\beta \in (-\infty, 1) \setminus \{0\}$	$\beta = 0$	$\beta = 1$	$\beta \in (1, 2)$	$\beta \in [2, +\infty)$
$\tilde{d}(x y)$	$\frac{1}{1-\beta} x y^{\beta-1}$	$\frac{x}{y}$	$-x \log y$	$\frac{1}{\beta} y^\beta - \frac{1}{\beta-1} x y^{\beta-1}$	$\frac{1}{\beta} y^\beta$
$\hat{d}(x y)$	$\frac{1}{\beta} y^\beta - \frac{1}{\beta(1-\beta)} x^\beta$	$\log \frac{y}{x} - 1$	$y + x \log x - x$	$\frac{1}{\beta(\beta-1)} x^\beta$	$-\frac{1}{\beta-1} x y^{\beta-1} + \frac{1}{\beta(\beta-1)} x^\beta$

Table 2: Proposed concave-convex decomposition of the discrete  $\beta$ -divergence.

In Table 2,  $y \in (0, \infty)$ ,  $\beta$  is real valued and  $x \in (0, \infty)$ . Further,  $\beta$  and  $x$  are considered as parameters,  $d_\beta$ ,  $\hat{d}$  and  $\tilde{d}$  being handled as univariate functions of  $y$ .

Let us now recall the definition of a complete monotonic function  $f$ :

**Definition 5.1.** A function  $f$  is said to be completely monotonic (c.m.) on an interval  $I$  if  $f$  has derivatives of all orders on  $I$  and  $(-1)^n f^{(n)}(x) \geq 0$  for  $x \in I$  and  $n \geq 0$ .

We can now introduce the properties of concavity, convexity and monotonicity for our convex-concave formulation of the discrete  $\beta$ -divergence:

**Proposition 4.** *Given  $\check{d}(\cdot|\cdot)$  and  $\hat{d}(\cdot|\cdot)$  as defined above, we have that*

1.  $\check{d}(x|y)$  is  $C^\infty$  and strictly convex on  $(0, \infty)$  for  $x > 0$  and  $\beta \in \mathbb{R}$ ;
2.  $\hat{d}(x|y)$  is concave for  $x > 0$  and  $\beta \in \mathbb{R}$ ;
3. for all  $\beta < 2$ ,  $\check{d}''(x|y)$  and  $\hat{d}''(x|y)$  are c.m.

*Proof.* The proof is straightforward, given that  $\check{d}(x|y)$  and  $\hat{d}(x|y)$  linearly combine  $C^\infty$  functions on  $(0, \infty)$ , and that in the same interval,

- $\log y$  is strictly concave;
- $y^\nu$  is strictly convex for all  $\nu \in (-\infty, 0) \cup (1, \infty)$ , and strictly concave for all  $\nu \in (0, 1)$ ;
- $y^\nu$  is c.m. for all  $\nu < 0$ .

□

According to the first two items of Proposition 4,  $\check{d}$  and  $\hat{d}$  indeed yield a convex-concave decomposition of the  $\beta$ -divergence, which is a variant of [45, Table 1]. Let us remark that the successive minimization of an upper approximation of this convex-concave decomposition following the methodology presented in [45] yields to the usual multiplicative update scheme.