Deterministic and Bayesian Orthogonal Nonnegative Matrix Factorization: Application to Blind Decomposition of Multispectral Document Images

by

Abderrahmane RAHICHE

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, FEBRUARY 7, 2022

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Mohamed Cheriet, Thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Mrs. Sylvie Ratté, Chair, Board of Examiners
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Stéphane Coulombe, Member of the Jury
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Abdessamad Ben Hamza, External Examiner
Concordia Institute for Information Systems Engineering (CIISE), Condordia University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON JANUARY, 2022

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

FOREWORD

The present dissertation is structured in the form of a compilation of papers that have been published at or submitted to prestigious top rank journals in the field of image and signal processing. The papers included in this dissertation are integrated with high fidelity to ensure compliance with the proposed and published articles' structure and shape. Still, only peripheral modifications (e.g., figures framing, repositioning, and rescaling) were made under Ecole de technologie tuperieure's thesis guidelines.

ACKNOWLEDGEMENTS

I am deeply indebted to my supervisor, Prof. Mohamed Cheriet, for his continued guidance and support with full encouragement and enthusiasm throughout my journey at his research laboratory, Synchromedia Lab.

Special thanks to the members of my Ph.D. committee, Prof. Sylvie Ratté and Prof. Stéphane Coulombe from Software and Information Technology Engineering Department, École de technologie supérieure, and Prof. Abdessamad Ben Hamza from Concordia Institute for Information Systems Engineering (CIISE), Condordia University, for accepting to evaluate this work and for their worthwhile feedback.

I would like to acknowledge people who encouraged and supported me during this PhD journey. Thank you for lending me your invaluable support at the time of need. I have to mention also my Synchromedia colleagues and my friends at the ETS.

My endless gratitude goes also to all my family members whose constant love and support kept me motivated and confident.

Méthodes déterministes et Bayésiennes de factorisation orthogonale de matrices non-négatives : Application à la décomposition aveugle des images multispectrales de documents

Abderrahmane RAHICHE

RÉSUMÉ

Cette thèse s'intéresse au problème de la décomposition aveugle des images spectrales de documents en couches de leurs matériaux constitutifs (encres, papier, pigments, etc.). L'objectif est de développer des outils de traitement avancés permettant d'utiliser efficacement les technologies d'imagerie spectrale pour l'analyse de documents et de simplifier leur traitement.

Comme les imageurs spectraux fournissent plus d'informations sur plusieurs longueurs d'onde d'acquisition, pour ces images, les observations des scènes de documents sont considérées comme des mélanges de certains signaux latents de leurs objets constitutifs, où ni l'opérateur de mélange ni les sources mélangées sont connus. Par conséquent, nous faisons appel aux techniques de séparation des sources (BSS) pour estimer ces deux opérateurs inconnus. Plus précisément, nous considérons la factorisation matricielle non négative (NMF), l'une des puissantes techniques BSS. Le problème est alors inversé, en se basant uniquement sur un ensemble d'observations, la NMF vise à identifier les deux opérateurs inconnus. Ce processus se transforme alors en une opération algébrique de la forme $\mathbf{X} \approx \mathbf{MA}$, où $\mathbf{X}$ représente les observations organisées en une matrice $m \times n$ de données, $\mathbf{M}$ est une matrice $m \times k$ de sources et $\mathbf{A}$ est une matrice $k \times n$ de coefficients.

La séparation aveugle des images spectrales est loin d'être une tâche simple de segmentation d'images. De nombreux défis dus à la haute dimensionnalité des données, au manque d'échantillons étiquetés et à d'autres ambiguïtés liées aux méthodes BSS, doivent être pris en compte pour traiter ce problème de manière appropriée. Pour relever ces défis, ce travail étudie la contrainte d'orthogonalité sur la variété Stiefel (Stiefel manifold) comme une caractéristique clé pour la NMF. Par conséquent, cette étude propose de nouveaux modèles NMF orthogonaux (ONMF), allant des approches déterministes aux modèles Bayésiens. Nous avons d'abord proposé une nouvelle formulation ONMF sur la variété Stiefel. Ainsi, en fonction de la matrice de facteurs qui doit être orthogonale, nous avons développé trois nouveaux modèles ONMF. Ensuite, pour tenir compte de la non-linéarité inhérente à l'espace des caractéristiques des données spectrales et la structure géométrique intrinsèque ignoreé par la vectorisation requise par NMF, nous avons proposé un nouveau modèle non linéaire avec une régularisation de la variation totale du graphe des données. Ce nouveau modèle est immunisé contre le problème de la pré-image et préserve la structure des données. Enfin, nous avons reformulé le problème d'un point de vue Bayésien afin d'aborder les limites des approches déterministes telles que l'incertitude des paramètres du modèle et la sélection de l'ordre du modèle. L'efficacité et les performances des algorithmes proposés ont été évaluées sur des ensembles de données synthétiques et réelles. Les résultats démontrent leur efficacité à traiter la séparation non supervisée des composantes dans les images spectrales de documents.

La contribution et l'impact de ce travail de recherche sont doubles. D'une part, il contribue plusieurs algorithmes pratiques pour la décomposition aveugle des images spectrales de documents basés sur la NMF orthogonale. D'autre part, il met en lumière les avantages des technologies d'imagerie spectrale pour l'analyse de documents à travers des cas d'utilisation concrets allant de la différenciation des encres à la séparation des matériaux en se basant seulement sur leurs propriétés optiques.

Le cadre de travail développé ouvre la porte à de nouvelles applications qui n'ont pas pu être ciblées par des approches classiques. L'identification et la différenciation des matériaux d'écriture, la détection des fraudes et des modifications de contenu sont les applications potentielles, pour n'en citer que quelques-unes. Ultimement, cela permettra une utilisation étendue et efficace de la technologie spectrale dans le domaine de l'analyse des images de documents.

Mots-clés: Documents historiques, décomposition aveugle des images, séparation aveugle de sources, images multispectrales, factorisation matricielle non-négative, variété Stiefel, modèles graphiques Bayesiens, variation totale.

Deterministic and Bayesian Orthogonal Nonnegative Matrix Factorization: Application to Blind Decomposition of Multispectral Document Images

Abderrahmane RAHICHE

ABSTRACT

This thesis addresses the problem of blind decomposition of spectral document images into layers of their constituent materials (e.g., inks, paper, pigments). The aim is to develop advanced processing tools that enable the effective use of spectral imaging technologies for document analysis and simplify their processing.

Since spectral imaging provides more information on several acquisition wavelengths, for such images, observations of document scenes are considered as a mixture of some latent signals of their constituent objects, where neither the mixing operator nor the mixed sources are known. Therefore, we resort to Blind Source Separation (BSS) techniques to estimate these two unknown operators. Specifically, we consider nonnegative matrix factorization (NMF), one of the powerful techniques for the analysis of high-dimensional data. NMF provides a lower-rank approximation of the data that is easier to interpret. The problem is then inverted, and NMF seeks to estimate the two unknown factors based only on a set of observations. Formally, this problem turns into an algebraic operation of the form $\mathbf{X} \approx \mathbf{MA}$, where $\mathbf{X}$ represents the observations reshaped into an $m \times n$ matrix, $\mathbf{M}$ is an $m \times k$ matrix of sources, and $\mathbf{A}$ is a $k \times n$ matrix of coefficients.

This problem is far from being a simple image segmentation task. Indeed, many challenges owing to the high-dimensionality of data, lack of labeled samples, and other ambiguities related to BSS methods should be considered to handle this problem appropriately. Hence, to tackle the aforementioned challenges, this thesis investigates the orthogonal constraint over the Stiefel manifold as a key feature for NMF. Consequently, this study proposes new orthogonal NMF (ONMF) models ranging from deterministic to probabilistic settings. We first proposed a new ONMF formulation over the Stiefel manifold. Therefore, according to which factor matrix is constrained to be orthogonal, we developed three new ONMF models. Then, to account for the non-linearity inherent in the feature space of spectral data and the intrinsic geometrical structure lost by the vectorization required by NMF, we proposed a new non-linear ONMF model with a graph-based total variation regularization. The new model is immune against the pre-image issue and preserves the structure of the data. Finally, we reformulated the problem from a Bayesian perspective to address limitations of deterministic frameworks, such as model parameters uncertainty and model order selection. The performance of the proposed algorithms has been evaluated on synthetic and real-world datasets. The results demonstrate their efficiency in handling unsupervised separation of components in spectral document images.

The contribution and impact of this research work are two folds. On the one hand, it provides several practical algorithms to handle spectral document images decomposition based on orthogonal NMF. On the other hand, it promotes the benefit of spectral imaging technologies for analyzing documents through concrete use-cases ranging from inks differentiation to materials separation based on their optical properties only.

The developed framework opens the doors toward new applications that could not be handled using traditional document images processing approaches. Some of the potential applications include writing materials identification and differentiation, forgery detection, and content change detection, to name a few. Ultimately, this will enable extensive and efficient use of spectral technology in the field of document image analysis.

Keywords: Historical document, Blind image decomposition, Blind source separation, Multispectral image, Nonnegative matrix factorization, Stiefel manifold, Bayesian graphical models, total variation.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

ADMM	alternating direction method of multipliers
ARD	automatic relevance determination
BIC	Bayesian information criterion
BSS	blind source separation
DI	document image
HD	historical document
HS	Hyperspectral
ICA	independent component analysis
IR	infra-red
KKT	Karush-Kuhn-Tucker
KL	Kullback–Leibler
LMM	linear mixture model
MCMC	Monte Carlo Markov Chain
MS	Multispectral
NMF	nonnegative matrix factorization
ONMF	orthogonal nonnegative matrix factorization
PCA	principal component analysis
PGD	projected gradient descent
RGB	read, green, blue

SVD	singular value decomposition
UV	ultraviolet
VB	variational Bayes

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathbf{A}$, $\mathbf{a}$, a	Matrix, vector, and scalar of appropriate sizes
$\mathbf{A}^T$, $\mathbf{A}^{-1}$	Transpose and inverse of a matrix
$\mathbb{R}_+$	Set of positive real-valued matrices
b,n,k	Number of bands, pixels, and endmembers, respectively
$\mathbf{I}_k$	Identity matrix of k-by-k size
$\max(.,0)$	Element-wise maximum of two matrices
prox	Proximal operator
$\|.\|_F$	Frobenius norm
∇	Gradient (first order derivative) operator
$\langle \mathbf{A}, \mathbf{B} \rangle$	Euclidean inner product, $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$
Tr(.)	Matrix trace
$\mathcal{H}$	Hilbert space
$\mathbb{1}$	Vector of ones ($\mathbb{1} = [1, 1, 1, ..., 1]^T$)
$\mathcal{L}$	Lagrangian function
$\odot$	Hadamard product
$\mathcal{N}$	Normal distribution function
$\mathcal{E}$	Exponential distribution function
$\mathcal{TN}$	Truncated normal distribution function
$\mathcal{G}$	Gamma distribution function

$\mathcal{VMF}$	Von Misses-Fisher distribution function
$\mathbb{E}$	Expectation (1^{st} moment) of a random variable
$p(.)$	Probability distribution of a random variable
$q(.)$	Probability distribution of a variational variable
$\mathcal{D}_{\text{KL}}$	Kullback-Leibler (KL) divergence
ln	Natural logarithm

INTRODUCTION

This chapter outlines the context of this thesis, describes the rationale and the motivation behind this research work and highlights the problem statement alongside the questions addressed. At the end of this chapter, an overview of this dissertation is given to facilitate understanding it.

0.1 Context and background

Document writing and artwork painting are two incremental and piecemeal production processes. Starting from a blank page or media, several physical layers of inks or pigments (in terms of text, figures, drawings, etc.) are usually added (or superposed gradually on the support) over time until the final word (or touch) is put on the support. In the case of historical documents (HD), they may also contain other layers of materials and degradation added to the original ones over time. In order to preserve existing HD, facilitate their storage and archiving, and provide public access to explore them, several digitization projects have been launched around the world since the development of the first computers. The goal is to produce a digital representation of these physical documents. Despite these efforts, this dematerialization process can not guarantee that all characteristics are transformed into digital forms. Indeed, when creating a digital replica of a document or painting, especially with conventional imaging systems, two essential pieces of information are generally lost: materials' physical characteristics and chronological arrangement of content on the paper.

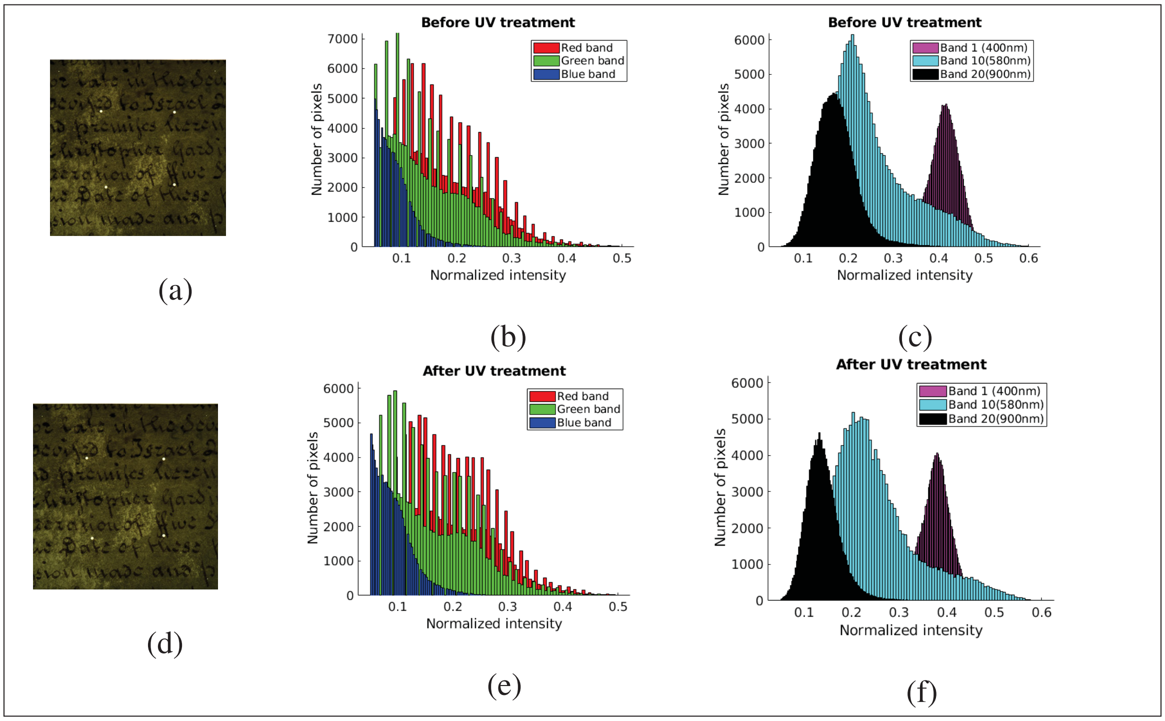
The pipeline of document image (DI) analysis consists of several steps involving the segmentation, which one of the crucial and challenging tasks. The goal behind separating the pixels of different objects (ultimately of different materials) present in a document image (DI) scene from each other is to make the scene more understandable for machines. DI segmentation has been extensively studied in the past decades, and several methods have been developed, which can be broadly divided into three main categories (Le, Nayef, Visani, Ogier & De Tran, 2015), namely, pixel-based methods, region-based techniques, and connected component methods. Nevertheless, due to the variability and complexity of DI scenes, it is not easy to design a generic approach that can be efficient for all types of documents (Jain & Yu, 1998).

From a practical perspective, the segmentation level (i.e., the objects that can be segmented) depends on the representation space of the image's pixels. It can vary from binary segmentation, i.e., text/non-text segmentation[1] (Le *et al.*, 2015), in the case of monochromatic representation, to multi-objects segmentation in the case of trichromatic representation (Ouji, Leydier & LeBourgeois, 2013). It is evident that there is a proportional relationship between the level of representation and the ability to distinguish between different objects in an image scene. Indeed, in nature, the ability to see ultraviolet (UV) light in some predators like Kestrels allows them to track urine trails left by small mammals and other prey species from the sky even with the presence of vegetation. Contrarily, other animals, such as dogs, are bichromatic and tend to mix between red and green colors, which limits their visual acuity.

Technically, due to the limited representation space, distinguishing between objects (materials) that look similar to the naked-eye, identifying if any changes or modifications have been made to a document, or inferring the chronological order of any modifications are quite challenging tasks that can not be handled with conventional gray-level or RGB (Red, Green, Blue) representation systems. Thus, obtaining more optical properties of materials (i.e., paper, inks, pigments) using higher spectral representation is of great importance for analyzing documents, especially historical ones. As illustrated in Fig. 0.1, by using appropriate imaging systems, even a fine change of the same image scene can be tracked between two states, i.e., before and after being exposed to ultraviolet (UV). While conventional RGB representation of the two images looks

[1] Non-text in document images means one category or more of the following: background, degradation, stamp, logo, signature, drawing, table, etc.

similar and the corresponding pixels intensity histograms show similar characteristics, the histograms of the UV representation show a clear difference between the two states[2].



(a) (b) (c)

(d) (e) (f)

Figure 0.1 Optical properties of iron-gall ink before and after artificial ageing with UV light captured by RGB and MS cameras

Among a variety of imaging technologies used for the digitization of HD, spectral imaging is one of the most adopted digitization techniques for materials characterization of cultural heritage in the literature (Tonazzini *et al.*, 2019). Indeed, Spectral imagery is an advanced imaging modality and a non-destructive technique that allows extracting useful and high-dimensional spectral representations of materials from scanned objects. From a technical perspective, spectral cameras record reflected light (spectral reflectance) from objects at specific frequencies, i.e.,

[2] Figure is from (Rahiche, Hedjam, Al-maadeed & Cheriet, 2020) and the samples in (a) and (b) are adapted from (Giacometti *et al.*, 2017), a) represents the color image of the original non-treated document. b) histograms of ink pixels of the RGB channels of image (a), c) three histograms of ink pixels from MS images acquired in bands 1 (400nm), 10 (580nm), and 20 (900nm) in black. d) color image of the UV-treated document, e) histograms of ink pixels from the RGB channels of image (d), and f) corresponding histograms of the same three bands 1, 10, and 20 after UV treatment.

different wavelengths, across the electromagnetic spectrum, and provide several corresponding digital images (see Fig. 0.2).

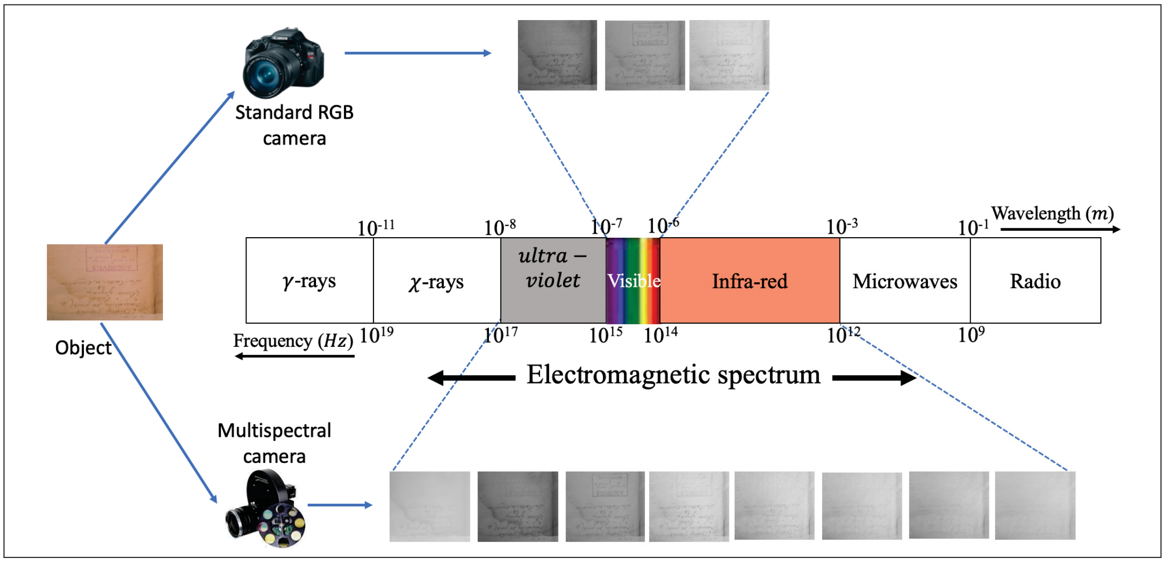


Figure 0.2 Conventional RGB cameras and Multispectral cameras

Spectral imaging systems were first developed in the early 1980s for airborne remote sensing. They were introduced to the field of cultural heritage, in the early 1990s, for the examination of paintings and artworks (Burmester *et al.*, 1992; Baronti, Casini, Lotti & Porcinai, 1998). Since then, they have become an attractive digitization technique in this field.

Depending on their spectral resolution, spectral imaging systems can be roughly divided into Multispectral (MS) and Hyperspectral (HS) systems[3]. The former can capture more than three bands with generally high spatial resolution, and the latter have a higher spectral resolution (up to several hundred channels). These two imaging systems measure, for every pixel of the image, the spectral response of objects at a number of wavelengths λ. Accordingly, they record a three-

[3] The main difference between MS vs. HS imaging is the number of bands and the technique used for data collected.

dimensional representation of a scene, usually expressed as a data-cube $I(hight, width, bands)$, see Fig. 0.3.

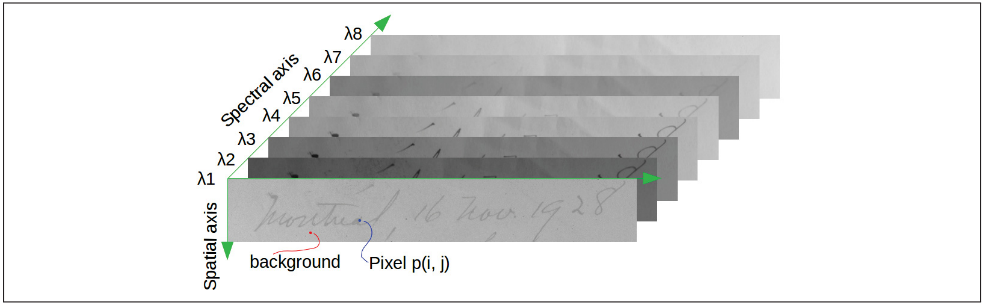


Figure 0.3 MS data-cube of a document image. Here # of bands = 8

Spectral imaging systems provide more spectral information than conventional panchromatic and trichromatic acquisition devices. They allow to obtain a unique spectral fingerprint of each material composing the scanned document. This characterization permits the identification and localization of these materials. Moreover, spectral imagery has opened up new horizons in the field of HD analysis and allowed reviving many non-exploitable documents. So far, it has been used for hidden writing revelation (Easton, Christens-Barry & Knox, 2011), erased writings recovering in Archimedes palimpsests (Easton & Kelbe, 2014; Salerno, Tonazzini & Bedini, 2007), historical documents enhancement and restoration (Hedjam & Cheriet, 2013b), inks mismatch detection (Khan, Shafait & Mian, 2013; Cosentino, 2014), forgery detection (Lyu, Liao, Li & Wu, 2016; Silva *et al.*, 2014), and document age estimation (Rahiche *et al.*, 2020).

0.2 Motivation

Although there is a growing interest in deploying MS/HS imaging techniques for digitizing HD and painting artworks in recent decades, processing tools still rely on methods developed for single-channel or RGB images. These techniques have shown their limitations in handling multichannel images. Moreover, most current studies target very specific tasks by investigating only selected channels, which does not benefit from the rich spectral representation. Therefore, in an attempt to simplify the low-level analysis of MS document images and expand their scope of applications, this research work aims to develop new and efficient unsupervised data factorization models for MS document images decomposition.

0.3 Problem statement

Our research work is based on the central hypothesis that spectral document image can be seen as a superposition of several layers of their constituent materials (papers, inks, pigments). Each material represents a layer added to the document and has its own spectral signature (fingerprint). Therefore, the spectral response at each pixel of the MS document image can be modelled as a mixture of some spectral signatures (sources) of materials present in the document. Based on that, we assume that the image decomposition (segmentation) issue can be rather seen as a BSS task.

The BSS problem can be defined as the problem of estimating the underlying sources from a given mixture without knowing neither the latent sources, their number, nor the mixing function. BSS methods, especially Nonnegative Matrix Factorization (NMF), have attracted significant attention within the signal and image processing communities over the past decades. NMF allows decomposing a data matrix into two (or three) nonnegative low-rank factors that, ultimately, approximate the latent sources and mixture model. Surprisingly, treating the problem of MS document image decomposition in the spirit of BSS is less widespread in the field of document image analysis.

Hence, in an attempt to fill in this literature gap, we propose to address the MS document images decomposition issue from a BSS perspective using NMF. Thus, the main problem this research investigates is **how to build an efficient BSS framework for the decomposition of MS document images based on constrained NMF algorithms under reasonable assumptions**

about the nature of data? More specifically, in order to address the main research problem formulated above, we focus on the following three cascading concerns, each formulated as a research question (RQ).

0.3.1 Research question RQ1

NMF is an ill-posed problem since the factorization may have many possible solutions (Donoho & Stodden, 2003; Laurberg, 2007; Laurberg, Christensen, Plumbley, Hansen & Jensen, 2008; Huang, Sidiropoulos & Swami, 2013). Therefore, in order to narrow down the solutions space, extra constraints and regularization terms are generally added to the standard NMF objective function such as orthogonality (Choi, 2008), sparsity (Hoyer, 2004; Pascual-Montano, Carazo, Kochi, Lehmann & Pascual-Marqui, 2006), and smoothness (Yokota, Zdunek, Cichocki & Yamashita, 2015). Among these constraints, the orthogonality condition has led to very interesting clustering and source separation performances in the literature. Therefore, the question that arises here is **how to design an efficient orthogonal NMF model to handle blind MS document image decomposition?**

The main challenges related to this question are:

1) How to efficiently incorporate the orthogonality condition into NMF?
2) How to solve the resulting model with the simultaneous nonnegativity and orthogonality constraints?
3) Theoretically, we can either place the orthogonality condition over the basis (endmember) matrix, the coefficient (abundance) matrix, or over both of them. Thus, which one of the three models should be adopted and based on which criteria?

0.3.2 Research question RQ2

Due to its simplicity, the linear mixture model is widely used in the literature of source separation using NMF (Cichocki, Zdunek, Phan & Amari, 2009). However, abstracting the mixture as a linear model is not always suitable to model real-world data in practice. The LMM model tends to ignore the correlation between bands in the feature space. Several studies, including (Zhang, Zhou & Chen, 2006; Buciu, Nikolaidis & Pitas, 2008; Zafeiriou & Petrou, 2009; Pan, Lai & Chen, 2011), have demonstrated that nonlinear extensions of NMF can extract more useful latent features from the original data than linear ones.

Nevertheless, most existing nonlinear NMF models suffer from two major drawbacks. The first one is related to the low computational effeciency of some existing methods as in (Zhang *et al.*, 2006; An, Yun & Choi, 2011; Tolić, Antulov-Fantulin & Kopriva, 2018). These models require calculating memory expensive Gram matrices involved by the kernel model, which become very challenging in the large-scale setting. The second one is related to the so-called pre-image problem (Kwok & Tsang, 2004) in the literature of kernel methods. Several existing models can only provide basis factors in the mapped space (Zhu & Honeine, 2016; Tolić *et al.*, 2018) rather than the original space, which allows analysing only one of the two factors. Obviously, excluding the Gram matrix of the input data matrix from the NMF model can relax the model complexity. Moreover, obtaining a basis matrix, which consists of a set of endmembers signatures, in the original feature space as the input data matrix is indeed essential for results interpretation in many cases.

On the other hand, processing MS images by NMF involves the vectorization of MS images to form the required data matrix. However, this transformation alters the intrinsic geometrical structure between pixels of input images making the standard NMF incapable of preserving such structures.

Hence, the second research question we aim to answer is **how to develop a new NMF model that takes into account the correlation between channels and the intrinsic geometrical structure of data points while being more adapted to handle our data?**

The main technical challenges related to this research question can be listed as follows:

1) How to formulate a new NMF model that accounts for the non-linearity of features and incorporates the intrinsic geometrical structure of document image data?
2) How to bypass computing a huge Gram matrix?
3) How to efficiently solve the formulated problem?

0.3.3 Research question RQ3

Usually, the optimization scheme used to solve NMF models depends on certain parameters, which, together with those induced by kernel functions, have to be adequately tuned to achieve good results. However, this step is generally time consuming and computationally expensive.

On the other hand, in the standard NMF models, the noise inherent in many real-world datasets is ignored, making deterministic NMF models less robust. Moreover, due to the variability in document images' content, estimating the number of materials (sources) present in document image scenes is another challenging problem that should be addressed. This issue is still one of the open questions in this field. The number of sources is often manually selected, ultimately by an expert, or inferred using existing techniques such as trial and error or SVD decomposition. The latter gives the optimal solution without the positivity condition.

Therefore, to overcome the aforementioned limitations, the question we want to answer here is **how to formulate a new NMF model with the following features: data-driven parameters tuning, automatic model selection, and robust to noise while being computationally efficient?**

The main related challenges are:

1) How to incorporate the noise and rank estimation into the NMF problem?

2) How to enable data-driven parameters tuning in the new NMF model?

3) How to efficiently solve the resulting optimization problem with scalability aptitude?

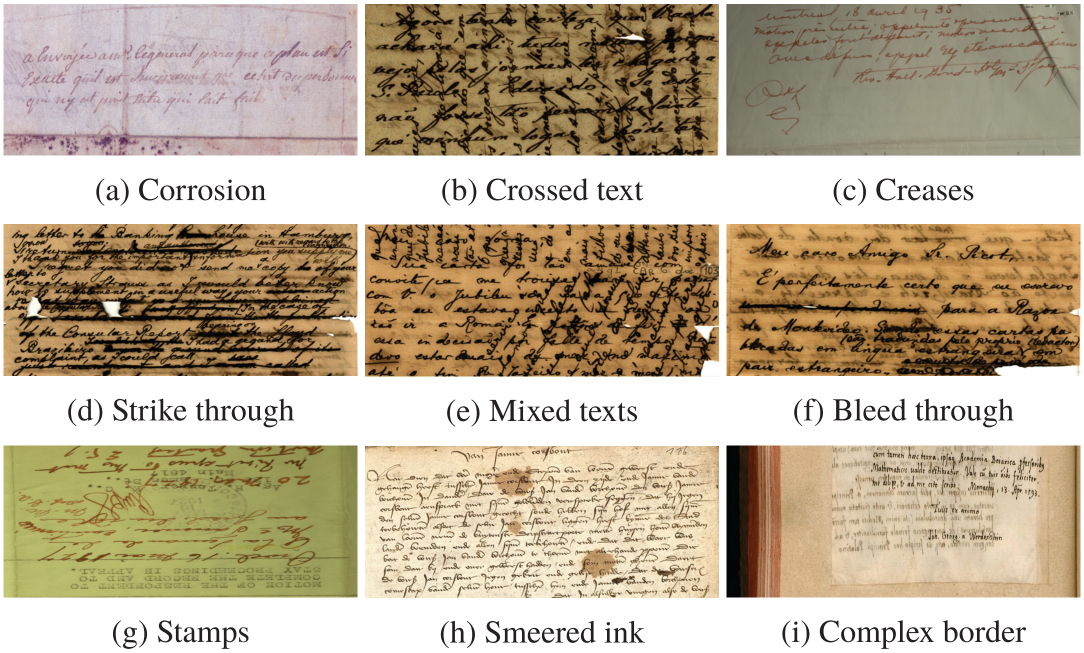
0.4 Challenges

Addressing the research questions mentioned above is not a trivial task due to several challenges that can be divided into three main parts:

0.4.1 Challenges related to ancient documents

The decomposition (segmentation) of HD images is challenging and more complicated than other types of digitized documents. The main challenges related to HD can be summarized into the following points:

1) Various degradation types and damages. Fig. 0.4 shows examples of some degradation types usually found in HD (images are taken from the MSTEx (Hedjam, Nafchi, Moghaddam, Kalacska & Cheriet, 2015) and the Document Image Binarization COmpetition (DIBCO series competition, (Pratikakis *et al.*, 2019)) datasets). As for the cause of these damages, many possible reasons could be the origin, such as the age of documents, the storing conditions, the chemical reaction between inks and papers, etc.



(a) Corrosion (b) Crossed text (c) Creases

(d) Strike through (e) Mixed texts (f) Bleed through

(g) Stamps (h) Smeered ink (i) Complex border

Figure 0.4 Example of challenging aspects in historical document images. Samples taken from the MSTEX and DIBCO datasets

2) Variability of writing materials and content.

3) Lack of labeled MS image datasets.

0.4.2 Challenges related to spectral imaging

Most challenges related to spectral imaging systems are summarized as follows:

1) The number of channels depends on the resolution of the spectral camera used, see Fig. 0.5. That means any processing system that relies on a specific number of bands would be impractical to handle another dataset acquired with a different spectral camera (with a different number of bands), making the generalization a big issue.

Figure 0.5 Examples of different number of bands in spectral imagery

2) The resulting large volume of collected data usually contains redundant and correlated information that need to be cleaned or pre-processed.
3) Depending on the targeted application, in most cases, not all the wavelength bands are required to solve a particular problem. The bands that are good for one application might not be helpful for another one.
4) MS imagery produces, in general, high-resolution images, which require high computational time to analyse them.

0.4.3 Challenges related to BSS methods

When it comes to developing a new NMF model to address a given BSS problem, several related challenges arise. The first and the most critical one is the non-uniqueness ambiguity for the decomposition solution. Among other challenges, we can also mention, for example, the number of factors considered or the factorization model, the constraints involved in the objective function, the regularization terms to be added to the optimization problem, the loss function employed, and the optimization method used to solve the formulated problem. A researcher has to make careful choices for each step to efficiently solve the problem at hand.

0.5 Thesis organization

This article-based dissertation is organized as a compilation of papers published by or submitted to prestigious and high-ranking peer-reviewed journals and conferences in image processing, computer vision, and pattern recognition to address the task of MS document images blind decomposition. Excluding this introductory chapter, the remaining of this thesis is organized as follows:

Chapter 1 summarizes the literature review about this research topic and previous works. A specific overview of related work is also given in each chapter with a focus on the related challenges.

Chapter 2 defines the objectives of this research work and presents the methodology adopted to address each research question.

Chapters 3 to **5** represent our published and submitted works according to the thesis objectives. Each chapter consists of a paper that targets one of the three research questions of this thesis. Each chapter follows a specific structure dictated by the corresponding journal/conference.

Chapter 6 provides a general discussion, which elaborates on our findings, the strengths, and limits of our approaches.

Conclusion and future work Chapter summarizes the contributions of this research work and enumerates some recommendations for improvements and future directions to build upon this work.

CHAPTER 1

LITERATURE REVIEW

This chapter is divided into two parts. The first part is devoted to explaining key concepts and prerequisite information that are essential for understanding this dissertation's main body. The second part of the chapter provides a review of the relevant literature related to our research problem in order to position our research with respect to the literature and highlight gaps in existing studies.

1.1 Background

1.1.1 Multi- and Hyper-spectral document imaging

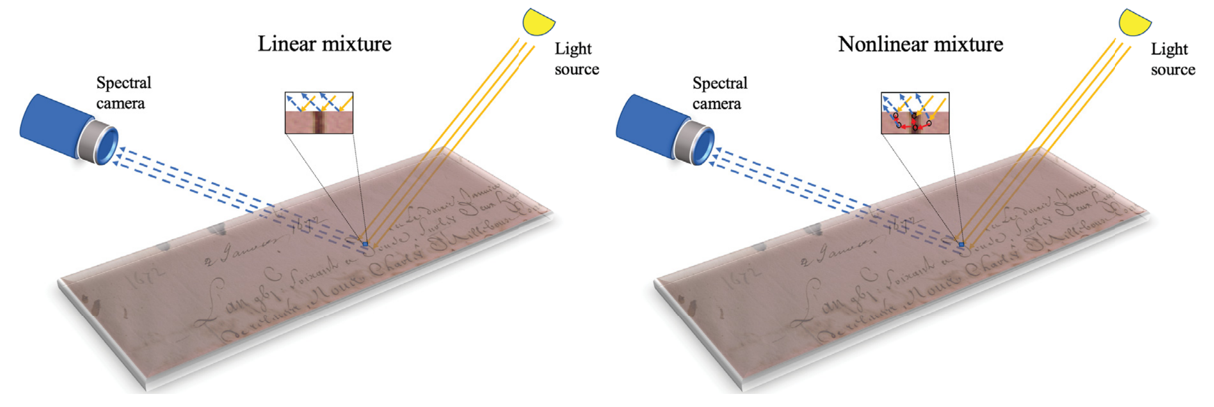
Spectral imagery refers to the simultaneous acquisition of images in several narrow and adjacent spectral bands through the electromagnetic spectrum (Goetz, Vane, Solomon & Rock, 1985). This technology estimates the physical properties of objects by measuring electromagnetic energy reflected from a scenes' surface. The modality of image acquisition and collection, as well as the spectral resolution (i.e., the number of spectral bands) of imagery devices, yields two main technologies: Multispectral and Hyperspectral imaging sensors (Shaw & Burke, 2003; Garini, Young & McNamara, 2006).

Spectral imaging technology has been first developed for remote sensing in space-borne and airborne systems (Shaw & Burke, 2003; Bioucas-Dias *et al.*, 2013), which can be tracked back to 1960s (Landgrebe, 1999). Since then, due to its versatility, it has gained growing interest in many other disciplines both in academia and industry. Since its introduction to the field of cultural heritage in the early 1990s, spectral imaging has emerged as an effective non-destructive imaging modality. This contact-less and non-invasive digitization technique provides much more information than conventional devices. It became the privileged and most effective technique for document analysis.

Archival documents and artworks (paintings) are the two main subjects of interest for most cultural heritage digitization projects. For paintings, previous studies turn mostly around material identification (Aldrovandi *et al.*, 1988; Toque, Sakatoku & Ide-Ektessabi, 2009; Grabowski, Masarczyk, Głomb & Mendys, 2018) and pigments dating (Melessanaki, Papadakis, Balas & Anglos, 2001). As for ancient documents, in addition to traditional use-cases such as text extraction and document segmentation, spectral images allowed targeting new applications ranging from text restoration and recovery (Balas *et al.*, 2003; Hedjam & Cheriet, 2013b), forensic document examination (Edelman, Gaston, Van Leeuwen, Cullen & Aalders, 2012), inks differentiation (Khan, Shafait & Mian, 2015), to inks dating (Rahiche *et al.*, 2020).

It is worth mentioning that MS/HS data is generally represented in the form of a 3D Cube, where each slice represents a 2D image obtained at a specific wavelength. Moreover, in the literature, it is common to consider spectral data as a mixture of some pure source signals (also called endmembers) whether they have been obtained in laboratory (Kopriva & Cichocki, 2009a,b; Gutierrez-Navarro, Campos-Delgado, Arce-Santana, Mendez & Jo, 2013) or not (Nuzillard & Bijaoui, 2000; Bioucas-Dias *et al.*, 2013). Following this assumption[4], two main models have been proposed in the literature to model spectral data (Bioucas-Dias *et al.*, 2012) based on the level of interaction between light photons and materials of an image scene, namely the linear and non-linear mixture models:

[4] In our work, we also consider this assumption. However, investigating the physical interactions between light and materials on documents and the nature of resulting spectral responses is out of the scope of this study.



(a) Macroscopic single scattering (b) Microscopic multiple scattering

Figure 1.1 Linear (a) versus nonlinear mixture models (b)

1) The linear mixing model (LMM) is the most used model in the literature due to its simplicity. The LMM assumes that the mixture can be expressed as a linear combination of the endmembers spectra. That is, for each pixel, the corresponding model can be expressed as (Iordache, Bioucas-Dias & Plaza, 2011)

$$x_i = \sum_{j=1}^{b} m_{ij} a_j + n_i \tag{1.1}$$

where x_i is the measured intensity of the reflectance, m_{ij} and a_j are the reflectance and the fractional abondance, that are positive and sum to one, of the jth endmember, b is the number of spectral band, and n_i denotes the noise.

2) The nonlinear mixing model (NLMM) assumes that every pixel spectrum is now a nonlinear function of the endmembers and the abundance coefficients (Heylen, Burazerovic & Scheunders, 2010). That is

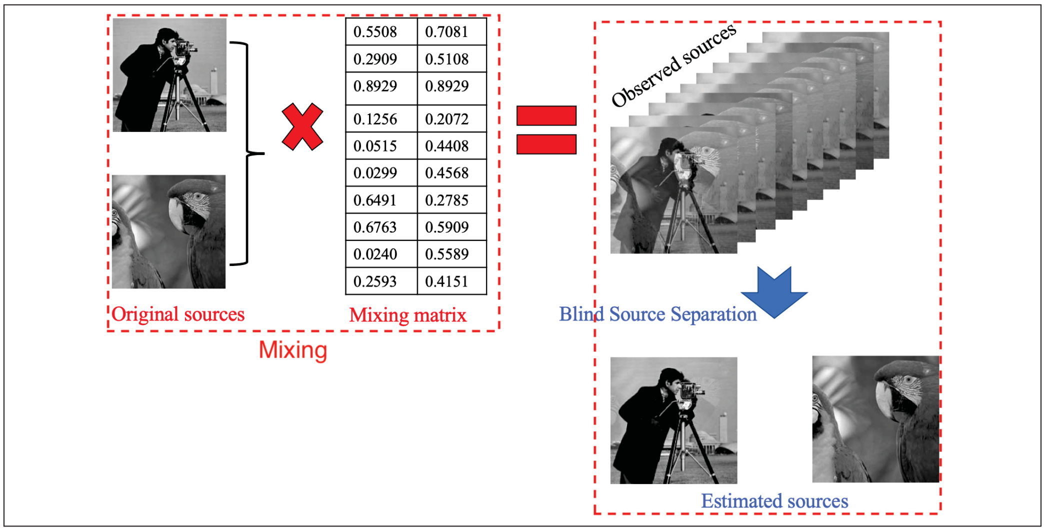
$$x = f(m, a) \tag{1.2}$$

The NLMM mixture can be expressed by numerous forms of non-linearity (Heylen *et al.*, 2010), e.g., multiple reflectance and scattering (Arai, 2008) and kernel-based models (Broadwater & Banerjee, 2009).

1.1.2 Blind Source Separation

Blind Source Separation (BSS) refers to the problem of extracting unknown source signals that have been mixed together in a number of observations without any information about the mixture model (Comon & Jutten, 2010) or the number of sources.

BSS techniques were initially developed for signal processing problems such as the cocktail party problem in speech signal processing (Hyvärinen & Oja, 2000; Bruna, Sprechmann & LeCun, 2015). They also found applications in image processing (see a blind image separation example in Fig. 1.2), remote sensing (Halimi, Altmann, Dobigeon & Tourneret, 2011), and medical image analysis (Boukouvalas, Levin-Schwartz & Adalı, 2017). In remote sensing, this problem is also known as the unmixing problem.



0.5508	0.7081
0.2909	0.5108
0.8929	0.8929
0.1256	0.2072
0.0515	0.4408
0.0299	0.4568
0.6491	0.2785
0.6763	0.5909
0.0240	0.5589
0.2593	0.4151

Figure 1.2 Example of BSS principle. The two images on the top left were mixed using a random mixing matrix. The estimated images on the down right were estimated using BSS

From the BSS point of view, the set of images delivered by a spectral sensor can be represented as a mixture of small latent sources. In the case of a linear, instantaneous, and invariant mixing model (Settle & Drake, 1993), which is the most studied BSS model, the problem is formulated as follows:

$$\mathbf{X} = \mathbf{MA} + \mathbf{N}, \tag{1.3}$$

where $\mathbf{X} \in \mathbb{R}^{b \times n}$ represents the set of measured sources, $\mathbf{M} \in \mathbb{R}^{b \times k}$ denotes the matrix of endmember spectra or just endmembers as adopted in the literature, $\mathbf{A} \in \mathbb{R}^{k \times n}$ is the matrix of coefficient, $\mathbf{N} \in \mathbb{R}^{b \times n}$ represents the noise, b is the number of measured sources, n the number of samples, and k the number of latent sources. Here, both $\mathbf{M}$ and $\mathbf{A}$ are assumed to be unknown. Therefore, the main idea of BSS techniques is to estimate them given only $\mathbf{X}$ and k.

In the literature, several BSS models have been proposed, which can be divided into four main classes (Cichocki *et al.*, 2009), namely Independent Component Analysis (ICA), Nonnegative Matrix Factorization (NMF), Sparse Component Analysis (SCA), and Morphological Component Analysis (MCA). This taxonomy depends on the assumptions made on the mixing and sources matrices.

1.1.3 Nonnegative matrix factorization (NMF)

Basically, NMF seeks to approximate a given nonnegative data matrix by the product of two (or three) low-rank matrices. More formally[5], given the data matrix $\mathbf{X} \in \mathbb{R}_+^{b \times n}$ and an integer $k \in \mathbb{N}^*$ with $k < min\{b, n\}$, NMF factorizes $\mathbf{X}$ into two low-rank matrices $\mathbf{M} \in \mathbb{R}_+^{b \times k}$ and $\mathbf{A} \in \mathbb{R}_+^{k \times n}$, such that: $\mathbf{X} \approx \mathbf{MA}$. The underlying numerical problem is generally cast as an optimization problem of the form:

$$(\mathbf{M}^*, \mathbf{A}^*) \in \arg \min_{\substack{\mathbf{M} \in \mathbb{R}_+^{b \times k} \\ \mathbf{A} \in \mathbb{R}_+^{k \times n}}} D(\mathbf{X}, \mathbf{MA}) + R(\mathbf{M}, \mathbf{A}) \tag{1.4}$$

where the term $D(\mathbf{X}, \mathbf{MA})$ denotes a cost function that measures the goodness of the factorization, and $R(\mathbf{M}, \mathbf{A})$ is a regularization term added to the problem to include additional constraints. In

[5] For consistency, we kept the same notation and indices used in Eq.1.3.

the standard NMF formulation, the two terms D and R are given as follows:

$$D(\mathbf{X}, \mathbf{MA}) = ||\mathbf{X} - \mathbf{MA}||_F^2, \quad R(\mathbf{M}, \mathbf{A}) = 0,$$

where $||.||_F^2$ is the Frobenious norm. Fig. 1.3 illustrates the principle of NMF, where the matrix **X** is approximated by the two smaller matrices **M** and **A**. The new representation requires only three components.

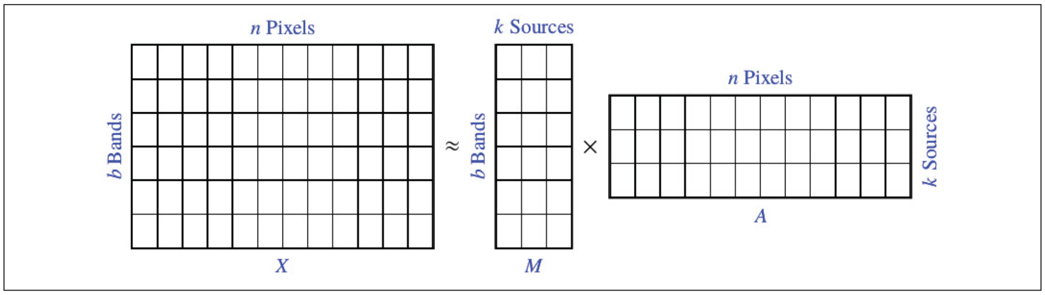


Figure 1.3 Illustration of the nonnegative matrix factorization principle. Here, $b = 6$, $n = 12$, and $k = 3$

The matrix factorization problem has been investigated by Paatero and Tapper (Paatero & Tapper, 1994). This topic got much more attention after the publication of the Lee and Seung work (Lee & Seung, 1999). Since then, several variants have been proposed as extensions to the standard NMF. A good but non-updated taxonomy of the basic NMF models can be found here (Wang & Zhang, 2012)

Nevertheless, NMF is generally an ill-posed problem (Gillis, 2014) due to the non-uniqueness of the solutions provided (Laurberg, 2007; Huang *et al.*, 2013). To better explain this problem, Fig. 1.4 illustrates a geometrical visualization of NMF approximation for a synthetic 3×13 data matrix.

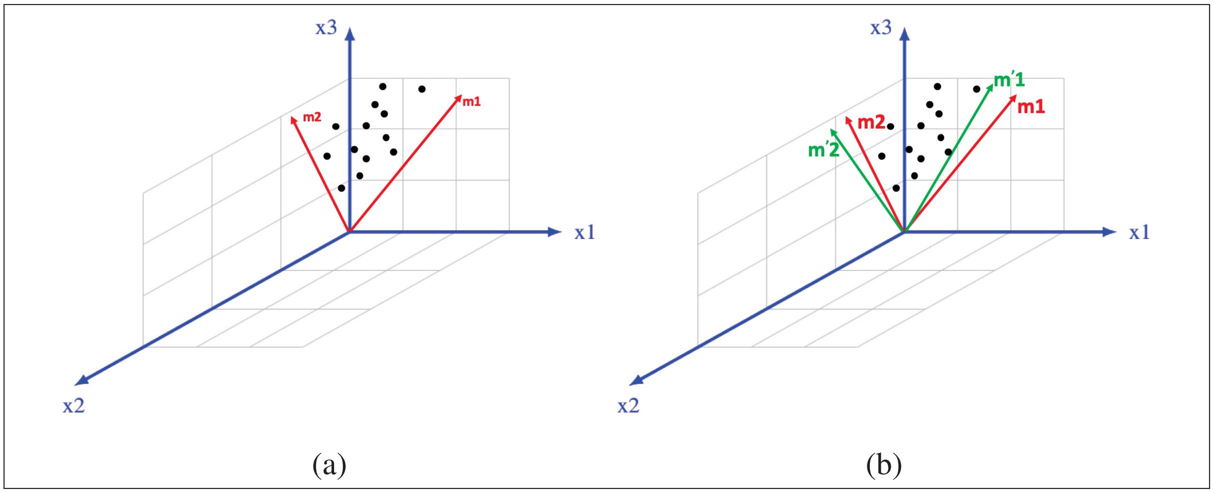


Figure 1.4 Geometric illustration of NMF representation of a 3 × 13 matrix. In red, a new representation with a rank $k = 2$. In b) another equivalent representation in green

In Fig. 1.4-a, the data can be represented by only the two axes (in red) obtained by NMF instead of the initial three axes (in blue). In Fig. 1.4-b, another equivalent representation for the same data points can be obtained by NMF (green axes).

Moreover, the initialization of NMF (Boutsidis & Gallopoulos, 2008) and the rank selection are two open problems in NMF that should also be taken into account when developing a new NMF model.

1.1.4 Stiefel manifold

The Stiefel manifold is defined as the set $St(p, n)$ of all $n \times p$ orthogonal matrices with $p < n$, that is

$$St(p, n) := \{\mathbf{X} \in \mathbb{R}^{n \times p} : \mathbf{X}^T\mathbf{X} = \mathbf{I}_p\}, \tag{1.5}$$

where $\mathbf{I}_p$ is the identity matrix of size $p \times p$.

1.1.5 Kernel function

A kernel k is any function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that corresponds to the inner (dot) product between two output vectors of some non-linear mapping function ϕ. It can be defined as

$$k(x, y) = \langle \phi(x), \phi(y) \rangle = \phi(x)^T \phi(y), \tag{1.6}$$

where $\phi : \mathbb{R}^n \to \mathbb{R}^d$ denotes a non-linear mapping function that maps an input vector $x \in \mathbb{R}^n$ to a higher dimentional vector $\phi(x) \in \mathbb{R}^d$ with $d \gg n$.

1.1.6 Gram matrix

The Gram matrix (or Gramian matrix) is a matrix whose entries are given by

$$G_{ij} = k(x_i, y_j) = \langle \phi(x_i), \phi(y_j) \rangle = \phi(x_i)^T \phi(y_j) \tag{1.7}$$

1.1.7 Hilbert space

A *Hilbert* space $\mathcal{H}$ is a an inner product space.

1.2 Literature review on spectral document image processing

In general, binarization and text extraction are the two main tasks that dominate the spectral document image analysis topic. We can broadly divide the main approaches into the following three main categories:

1.2.1 Conventional approaches

These techniques generally focus on processing individual spectral images and use hand-crafted image processing strategies and traditional approaches primarily devoted to binary decomposition (i.e., text and non-text classes) of gray-scale or color document images that are generally not suitable for such kind of data. Thanks to the organizers of the first contest on multispectral text extraction (MS-TEx2015) (Hedjam *et al.*, 2015) for drawing the attention of the document image community to these kinds of challenges.

For instance, in the first step of the ink mismatch detection framework proposed in (Khan *et al.*, 2013), the authored resorted to the adaptive thresholding Sauvola's (Sauvola & Pietikäinen, 2000) method to extract text from a single channel of the hyperspectral image (640 nm) visually selected among 33 bands.

A two-steps semi-automatic binarization approach is proposed in (Hedjam & Cheriet, 2013a) to generate the ground truth of the MS-TEx dataset. In the first step, the method performs a foreground/background separation using a phase-based method developed in (Nafchi, Moghaddam & Cheriet, 2014) for each image. In the second step, all outputs are combined to produce a rough binary image and manually post-processed to obtain the final binary outcome.

The binarization method proposed by (Hollaus, Diem & Sablatnig, 2015) uses a single image taken from one of the visible channels of the stack of multispectral images, on which a state-of-the-art binarization algorithm is applied. This binary output is then combined with the output of an Adaptive Coherence Estimator (ACE) to improve the binarization performance. The authors in (Diem, Hollaus & Sablatnig, 2016) proposed a similar framework, in which they added the GrabCut technique for spatial segmentation to generate the final binary image.

The common drawbacks of thresholding methods lie in:

1) Considering only a single band or few images from the whole stack of spectral images.
2) Not leveraging the benefits of extra spectral bands.
3) Targeting only binary separation.
4) Depending on a specific spectral setting.

1.2.2 Segmentation-based approaches

Unlike the methods that process individual spectral images, segmentation-based methods employ several strategies to extract useful features from the set of spectral images to separate them into the targeted objects. These methods involve handcrafted feature schemes as well as automatic feature extraction techniques. They include dimensionality reduction and bands selection approaches to reduce the amount of data to be analyzed. They also resort to pattern recognition techniques to classify images pixels into desired categories in a supervised or unsupervised fashion.

For instance, in (Lettner & Sablatnig, 2009, 2010), the authors proposed a Markov Random Field (MRF) based model for the binarization of MS image of an ancient Slavonic missal written on parchment. They incorporated the stroke properties and spectral information of MS images to classify the pixels into foreground and background classes. Their approach was applied to a single band image, corresponding to the wavelength $450nm$, representing the best contrast within their dataset.

In the second step of the inks mismatch detection framework proposed in (Khan *et al.*, 2013), the authors resorted to the K-means algorithm, a generic data clustering method for grouping ink pixels into a predefined number of classes. In (Morales, Ferrer, Diaz-Cabrera, Carmona & Thomas, 2014), a supervised classification system is proposed for pen/ink identification in handwritten documents. The authors built a database of HS curves of 25 pens of different ink types. They employed the Support Vector Machine (SVM) technique to classify the ink in the test phase. Clearly, both methods rely on training data and can not identify any type of inks that have not been seen before during the training phase.

In (George & Hardeberg, 2015), the authors investigated the task of ink separation using two spectral classification algorithms. They employed the Spectral Angle Mapper (SAM) and the Spectral Information Divergence (SID) to compare the similarity between the spectrum of image pixels and the ink reference spectrum. Despite the promising results, limited experiments have been conducted.

Despite the great success of deep learning methods in various fields, the lack of datasets prevented them from achieving state-of-the-art performance. Recently, an end-to-end Convolutional Neural Network (CNN) model was proposed in (Hollaus, Brenner & Sablatnig, 2019) for the segmentation of MS historical document images. The authors extended the generic architecture introduced in (Oliveira, Seguin & Kaplan, 2018), which was initially developed for RGB images, to handle multi-bands images.

In summary, segmentation based approaches suffer from several limitations that can be summarized as follows:

1) They are application-driven.
2) They require significant training data to be effective.
3) They rely on the same spectral setting of the training phase, where any system trained on data with a specific number of bands cannot handle another data with a different number of bands without retraining it.

1.2.3 BSS-based approaches

Despite the successful use of BSS in several signals and image processing fields, tackling the problem of document image segmentation from a source separation point of view is not widespread in the document image area. Existing applications of BSS methods have almost focused on conventional images to address old problems such as show-through and bleed-through removal. In (Tonazzini, Salerno & Bedini, 2007; Cheriet & Moghaddam, 2008), using the Independent Components Analysis (ICA), the authors formulated the problem as a source separation problem. They explored how to extract clean text from overlapping texts in the gray-scale recto and verso scans of documents affected by bleed-through or show-through distortion. In (Tonazzini, Gerace & Martinelli, 2009b), a Bayesian Convolutive BSS has been proposed for the separation and deconvolution of overlapped text in RGB images document. The same problem has been investigated in a recent study (Hanif, Tonazzini, Savino, Salerno & Tsagkatakis, 2018) using a sparse dictionary learning method. All these approaches are limited to gray-scale images and also target only images with overlapped texts.

As for multichannel document images, very limited attempts have addressed the issue of document images analysis from the BSS point of view. For instance, (Tonazzini, Bedini & Salerno, 2004a; Tonazzini, Salerno, Mochi & Bedini, 2004b) applied the ICA extensively for the restoration and the enhancement (Tonazzini, Bianco & Salerno, 2009a; Legnaioli *et al.*, 2013) of an MS historical manuscript.

In (Salerno *et al.*, 2007), the authors proposed a two-step processing technique consisting of cascading the principal component analysis (PCA) then ICA methods. They focused only on the extraction of the faint and highly degraded underwritten text from the MS images of the Archimedes palimpsest. In (Hollaus, Gau & Sablatnig, 2012) the enhancement of the readability and the visibility of palimpsest texts are addressed with two operators, the PCA technique is applied first to reduce the number of bands, and then the ICA is used to extract statistically independent outputs from the principal components already obtained.

Authors in (Mitianoudis & Papamarkos, 2014) proposed a three-stages binarization method. In the first step, a modified ICA fusion system combines MS images into two outputs. Then, the background and text components are separated from the two fusion outputs of the previous stage using FastICA. In the last stage, for generating the final binary image, they used the adapted Spatial kernel k-harmonic means clustering approach (Li, Mitianoudis & Stathaki, 2007) proposed to overcome the sensitivity of the standard K-means to center initialization.

Our literature review shows that the ICA was the method of choice in many previous studies in this field due to its success in other areas. In ICA, the sources are assumed to be non-normally distributed (non-Gaussian), which can not be generalized to all data types.

In the same spirit of mixture modelling, recently, (Hollaus, Diem & Sablatnig, 2018a) proposed a Gaussian Mixture Model (GMM) based binarization approach, in which the raw MS images from the MSTEx dataset (Hedjam *et al.*, 2015) are filtered with a two-dimensional median filter. The preprocessed images are then clustered with GMM. The k-means clustering algorithm was used to initialize the parameters used by the Expectation-Maximization (EM) algorithm. The proposed method suffers from main limitations: i) it is designed to work with the images of the MSTEx dataset only as it takes exactly eight input images with the exact ordering, and ii) it targets only text extraction.

As it can be noticed, apart from few attempts limited to statistical methods, BSS techniques are still less popular in document image processing.

1.3 Conclusion

As we mentioned earlier, the key ingredient of an efficient document analysis system is a rich level of representation obtained by cutting-edge high-dimensional digitization devices combined with advanced image processing tools that allow adequate analysis. From what we have shown in this chapter, we can observe that there is a significant need to develop efficient image processing approaches to analyze MS document images appropriately. Another concern out of this study's scope is the lack of available datasets, which calls for the need to build and release more datasets to develop the research in this field.

CHAPTER 2

METHODOLOGY & CONTRIBUTIONS

This chapter highlights the research objectives we defined for this research work and describes the methodology followed to achieve each research objective we set.

2.1 Research objective

Having emphasized the need for developing new methods for efficient multichannel document images processing, the main objectives of this research work were to develop a unified NMF based framework for the blind (unsupervised) decomposition of MS document images (see Fig. 2.1). This framework should: exploit the spectral richness of MS images, not rely on a particular setting (i.e., flexible and generalize without regard to the number of bands and their corresponding wavelength), not target a specific application, and do not require labelled datasets.

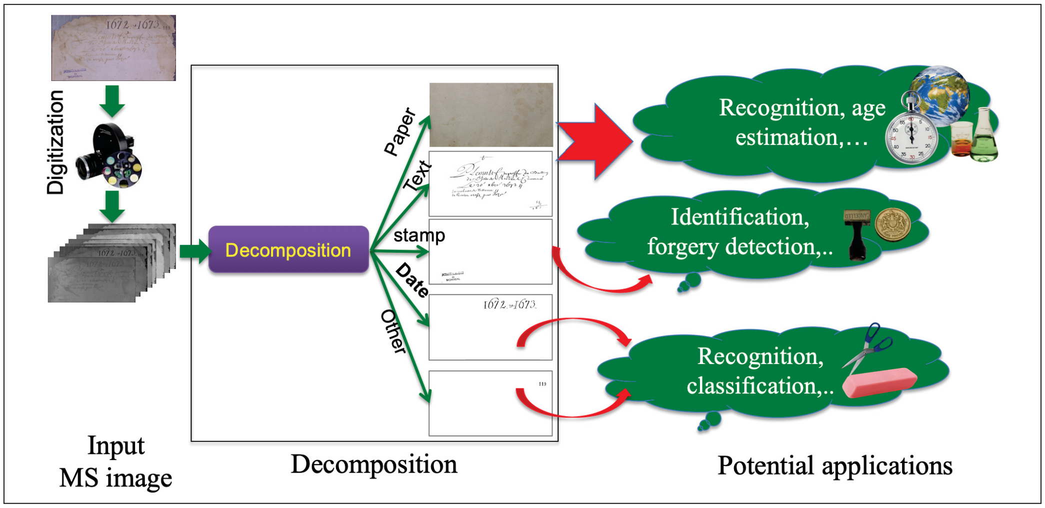


Figure 2.1 The main objective of our research problem

To provide a frame to this study and give achievable goals, we have further refined this main objective into more specific ones.

2.1.1 Objective 1: to develop an efficient orthogonal NMF model for source separation in MS document images.

As discussed in the previous chapter, NMF has found successful applications in many research areas but has not attracted much attention in the field of MS document images analysis. Nonnegativity is a quite important condition in NMF for the interpretation of results. However, due to the ill-posed nature of the NMF problem, this condition is not sufficient alone to obtain unique solutions.

Besides being one of the solutions to overcome the non-uniqueness limitation, incorporating the orthogonality constraint has demonstrated interesting clustering properties in the literature (Ding, Li, Peng & Park, 2006). From a physical point of view, imposing the orthogonality on the columns of the basis matrix is equivalent to assuming spectral non-correlation of the latent sources. Imposing orthogonality on the rows of the coefficient matrix amounts to their independence.

A geometric illustration of the orthogonality condition, when applied to NMF, is shown in Fig. 2.2. Based on this criterion, only the red solution is retained, and the green one is rejected. The red solution also implies linear independence of the red set.

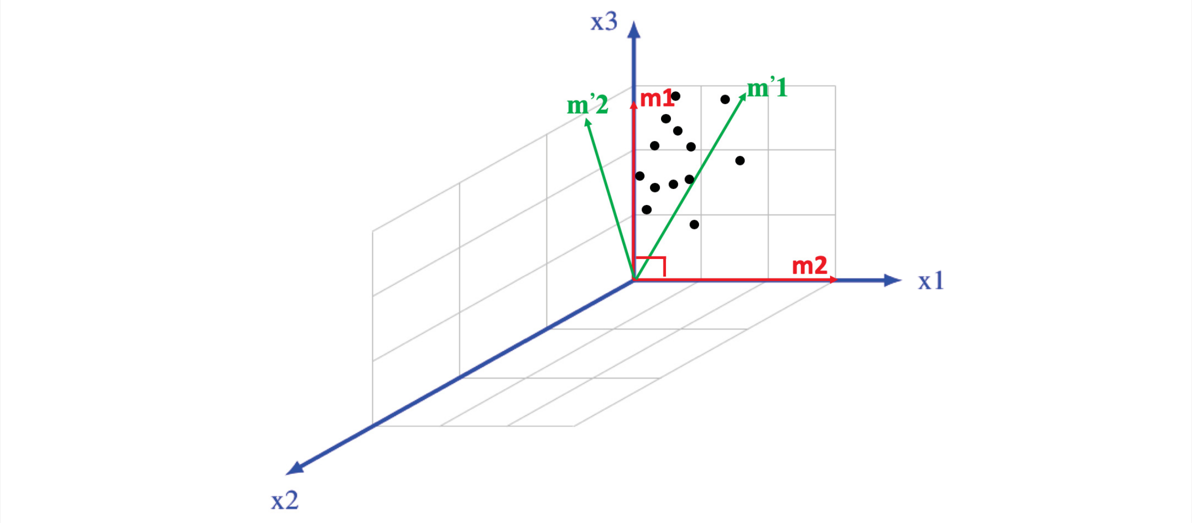


Figure 2.2 Geometric illustration of the NMF problem under orthogonal constraints

Therefore, in this research work, the first objective we defined is to develop an efficient orthogonal NMF based framework for the blind separation of MS document images. Unlike most previous studies that focus on applying conventional analysis on selected channels and for specific tasks, this framework should allow decomposing MS images components into several layers of materials (components) based only on their spectral signatures. In this stage, we focus only on performing good components separation and achieving better decomposition results. Therefore, the number of components is not investigated but rather set manually.

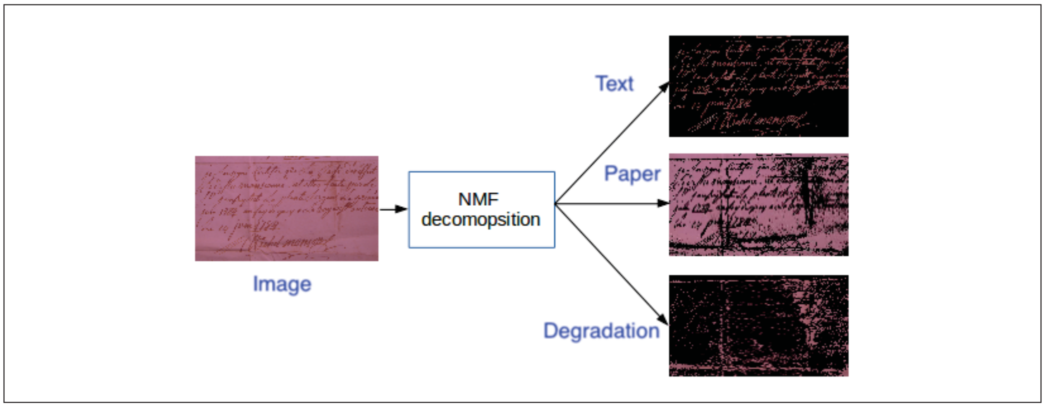
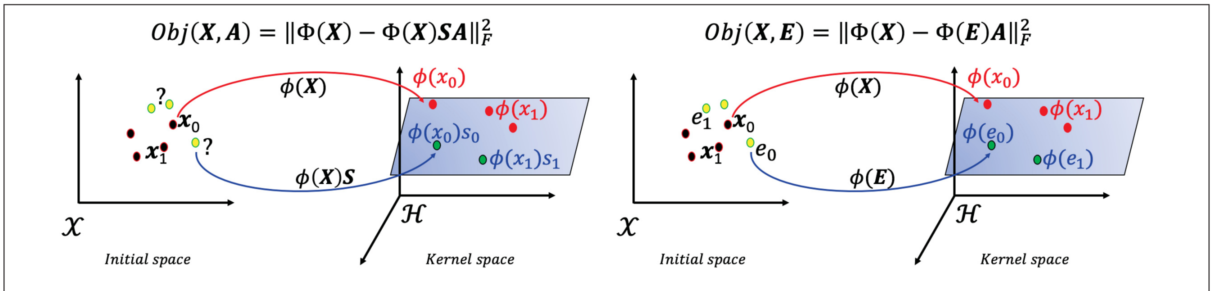


Figure 2.3 An example illustrating the 1st objective of our research problem. The input image is decomposed into three layers of components, i.e., text, paper, and degradation

2.1.2 Objective 2: to develop a new kernel orthogonal NMF model with graph-based regularization.

NMF is a powerful source separation method that requires an adequate input data representation to achieve good separation results. In the first objective, we adopted a linear mixture model to model the data. However, this model doesn't account for the non-linearity existing between the different bands, which is generally present in spectral data (Iordache *et al.*, 2011).

In the second objective, we target the development of a kernel NMF model to overcome the above limitations of the standard NMF. In addition to the kernel-induced nonlinear mappings applied to the endmembers (basis) matrix, this model should account for the spatial orthogonality imposed on the abundance (coefficient) matrix and the intrinsic geometrical structure of the data. To do this entails formulating the problem in a manner that allows obtaining endmembers representation in the original space (rather than the mapped Hilbert space (i.e., avoids the pre-image problem, see Fig. 2.4), and avoids calculating a huge Gram matrix for the sake of scalability.



(a) With pre-image issue.

(b) Without pre-image issue.

Figure 2.4 Illustration of the pre-image issue in kernels methods

Moreover, in order to preserve the intrinsic geometrical structure of document content that is generally lost by the vectorization of spectral images (see Fig. 2.5), we propose to incorporate a graph-based regularization term to provide graph structure information for document content.

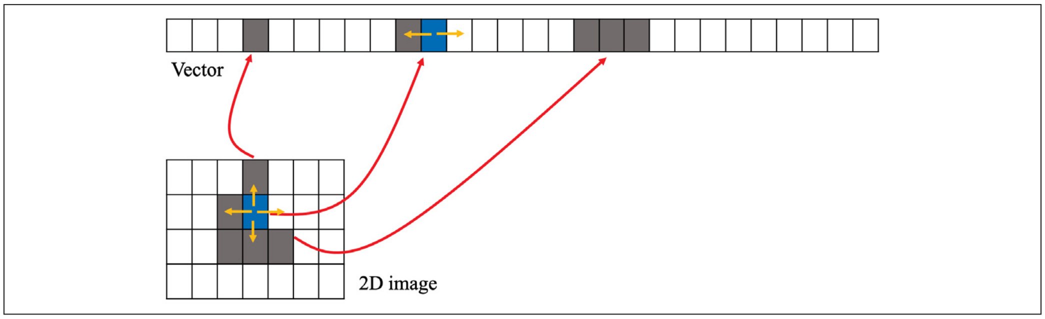


Figure 2.5 Local connectivity between pixels lost by the image vectorization operation

2.1.3 Objective 3: to develop a generative probabilistic NMF model for automatic estimation of the number of objects and model parameters estimation.

In the two first objectives, the emphasis is put on developing efficient orthogonal NMF models for multichannel sources separation in the linear and non-linear cases. In these models, the number of sources and parameters selection remain empirically determined, which is not practical.

Nevertheless, to build an NMF model that can estimate the number of sources (components) and accounts for parameters uncertainty, we target the development of a new probabilistic NMF model with such characteristics. Moreover, this probabilistic formulation of the NMF problem allows taking into account the noise inherent in the model. This point was also ignored in the first two objectives for simplicity. This probabilistic model should be data-driven. Based only on a few assumptions about the nature of sources, it should allow inferring the parameters of the different distributions involved. For the sake of scalability and efficiency, we avoid using sampling inference methods to build a model that can run in a reasonable time.

2.2 Research methodology

This thesis is devoted to developing new NMF models for blind decomposition of multichannel document images. In order to address the research questions and the sub-objectives defined previously, we adopted a similar methodology in each contribution of this research work. Therefore, for each research question:

1) We conducted a rigorous analysis of related work and existing methods in the related topic, synthesized their findings, identified their limitations, and defined possible improvements to address the shortcoming in the existing literature while advancing the state-of-the-art.
2) Based on a rational choice of the solution that will allow us to reach the stated objective, we formulated the problem mathematically and defined the proposed model. In practice, we investigated the orthogonality over the Stiefel manifold and how to incorporate it into the NMF formulation in a deterministic and a Bayesian fashion.
3) We devised efficient solutions to optimize the proposed NMF models. For the deterministic NMF models, our solutions were based on the Alternating Direction Method of Multipliers

(ADMM)[6] (Boyd *et al.*, 2011). As for the Bayesian NFM model, we adopted a Variational Bayesian inference scheme because of its low computational complexity compared to sampling methods.

4) Using different scenarios, we carried out extensive experiments to validate the proposed algorithms, where we used synthetic and real-world MS images of ancient documents as well. Our evaluation targeted several tasks such as complete MS document image decomposition, document image binarization, and inks differentiation. We compared our results with those of the related state-of-the-art techniques using standard quantitative metrics.
5) The last step consisted of disseminating our funding to the research community. The feedback we obtained during the peer review process helped shape our reports and improve our research quality. This step was also an opportunity for defending our choices and reflection moments about our solutions to figure out the limitations and possible improvement.

While iterating over the general steps mentioned above, a specific methodology was adopted to solve each one of our research questions. Hereafter we describe each methodology in more details.

2.2.1 Effective orthogonal NMF over the Stiefel manifold

Spectral document images data are high dimensional and sparse. Usually, document scenes consist of a few text entries among many background pixels. Analyzing such unbalanced data is a real challenge for supervised machine learning algorithms. ONMF variant has proven to be very effective in unsupervised data clustering and source separation tasks. Indeed, in our work (Rahiche & Cheriet, 2020) we have demonstrated that the orthogonality constraint yields sparse solutions and is useful to handle blind MS document image decomposition. However, most existing ONMF algorithms involve the nonnegativity as a penalty term added to the objective function. In such models, by minimizing the corresponding cost function, the orthogonality is achieved at its limit point of convergence.

[6] See the related description in Section 3.2.5 of Chapter 3.

Therefore, to achieve the first specific objective, in which we target performing good separation and achieving better decomposition results, we adopted a different strategy to incorporate the orthogonality into the NMF formulation. In contrast to existing approaches, our method works the opposite way by allowing the set of the orthogonal axis to exist in their natural topology, i.e., Stiefel manifold. The orthogonality constrain is transformed to an optimization problem over the Stiefel manifold. The orthogonality is strictly imposed at each iteration by performing steps on a set of orthogonal axis called Stiefel manifold[7]. Moreover, to investigate the issue on which factor the orthogonality is imposed, we developed three different models, a model for each case. The optimization problem related to each model is solved using the Alternating Direction Method of Multiplier (ADMM) scheme. ADMM is very effective in solving optimization problems with multiple variables and constraints. The problem formulation in each model involves a different objective function and optimization algorithm. The three models were compared against each other and several competitive approaches as well.

In **Chapter 3**, i.e. the first paper (Rahiche & Cheriet, 2021a), we developed three different orthogonal NMF models, called A-ONMF, M-ONMF, and MA-ONTF, respectively, according to on which factor we impose the orthogonality, i.e., over the matrix **A**, **M**, or both of them. The proposed algorithms demonstrated that incorporating the orthogonality over the Stiefel manifold provides better results than other strategies. The comparison made between the three models indicates that imposing the orthogonality on both factors provides better clustering results. The orthogonality yields a sparser solution, preventing adding other constraints to the objective function and keeping it simple. The proposed model was able to distinguish between materials that other conventional methods consider as one class. The obtained results outperform the results of several state-of-the-art algorithms, including deep learning-based models.

[7] To avoid redundancy, the reader could refer to Section 3.2.4, Chapter 3, for more in-depth details about optimization over the Stiefel manifold.

2.2.2 New nonlinear orthogonal NMF over the Stiefel manifold with graph-based total variation regularization

The model discussed in Objective 1 neither accounts for the non-linearity of data features nor preserves the local structure of data. To remedy these issues, we build upon our kernel orthogonal NMF model, called KONMF (Rahiche & Cheriet, 2021b), and our graph orthogonal NMF model, called GONMF (Rahiche & Cheriet, 2020), to design a new kernel NMF model. The core idea is to map the basis matrix into a higher-dimensional feature space (so-called Hilbert space $\mathcal{H}$) via a nonlinear transformation and include a graph regularization term to preserve the structure of data. Furthermore, we imposed the orthogonality over the Stiefel manifold on the coefficient matrix. In contrast to existing models, we reformulated the problem to prevent us from calculating a large size Gram matrix of the input data matrix (of size $n \times n$ samples).

In **Chapter 4**, we proposed a new nonlinear orthogonal NMF model with a graph-based total variation regularization, called GTV-ONNMF. In this model, our formulation promotes the preservation of the original feature space of endmembers and the geometrical structure of pixels. The ADMM technique allows simplifying solving our problem with all the considered constraints. We implemented the developed algorithms using a fast matrix operation toolbox, which helped avoid scalability and computational complexity issues. All the ingredients together allowed our new algorithm to boost the separation performance while being computationally efficient.

2.2.3 A Variational Bayesian Orthogonal NMF

As per Objectives 1 and 2, we targeted a non-probabilistic formulation of NMF, in which the noise is not considered for simplicity. This formulation also does not allow an easy investigation of the number of sources estimation. Fortunately, both challenges can be addressed efficiently through a probabilistic formulation that considers the decomposition factors as two independent random variables. It places priors distributions over them that act as regularization terms. As for the uniqueness of the solution, the orthogonality is incorporated via the von Mises-Fisher distribution, an orthogonal distribution over the Stiefel manifold.

Therefore, in **Chapter 5**, we developed a data-driven nonparametric Bayesian orthogonal NMF called Variational Bayesian Orthogonal NMF (VBONMF). In VBONMF, we incorporated the orthogonality using a family of orthogonal distributions called the Langevin (or von Mises-Fisher) distribution of random matrices on the Stiefel manifold. We also incorporated an automatic rank determination technique (ARD) prior to estimate the rank factorization (number of sources). We adopted a Variational Bayesian optimization scheme to solve the formulated problem. The results demonstrate that the proposed model outperforms state-of-the-art results and provides empirical evidence of its effectiveness.

2.3 Summary of contributions

Following our analysis of the state-of-the-art in this research area, the contributions we made were motivated by the limitations of existing approaches and the need to develop new techniques that can effectively handle the MS document image decomposition problem. In total, we contributed seven different NMF models. Five models among them are presented hereunder the chapters of this thesis. The remaining two models, called GONMF (Rahiche & Cheriet, 2020) and KONMF (Rahiche & Cheriet, 2021b), are not included in this dissertation.

The contributions made during this work provide practical tools for treating MS document images as a data-cube, processing them in an unsupervised way, and analyzing their content in a simple manner without targeting a specific task. These models should help simplify and improve the low-level processing of MS document images. This will also benefit many image analysts from the document image processing and signal processing community.

The proposed algorithms allow revealing latent information and targeting advanced applications that could not be addressed using traditional processing methods. The proposed algorithms will enable other potential applications with efficient analysis of MS document images (especially historical ones). As illustrated in Fig. 2.1, after the decomposition, subsequent MS document image analysis tasks are seen as instances and could be addressed easily. For example, we can cite three examples:

1) **Forgery detection:** after a successful decomposition, segments of text written by different materials will be automatically distinguished, and forgery will be spotted quickly.
2) **Material identification:** here one can simply measure the distance between the spectral signatures of the decomposed materials and some materials of references. Based on this measure, the materials found in the document scenes could be classified into the class with the shortest distance. This will open the door, for example, to materials-based object spotting in collections of MS document images.
3) **Text binarization:** this task could be done either by thresholding the corresponding text component or using another state-of-the-art binarization to convert the decomposed outcome to the requested binary form of the document images.

We also note that the proposed algorithms are not restricted to image data only. As demonstrated by our experiments, other types of data with the same properties can be handled with these approaches.

2.4 List of publications

A complete list of papers published or submitted during the preparation of this thesis is given as follows:

1) **Peer Reviewed Journals:**

1.1. **A. Rahiche**, M. Cheriet. "Nonlinear Orthogonal NMF on the Stiefel Manifold with Graph-based Total Variation Regularization." Submitted to **IEEE Signal Processing Letters**, 2022. (**IF: 3.109**)

1.2. **A. Rahiche** and M. Cheriet. "Variational Bayesian Orthogonal Nonnegative Matrix Factorization over the Stiefel Manifold." Submitted to **IEEE Transactions on Image Processing**, 2021. (**IF: 10.85**)

1.3. R. Hedjam, A. Abdesselam, **A. Rahiche**, and M. Cheriet. "Non-Negative Matrix Factorization with Scale Data Structure Preservation", Submitted to **Pattern Recognition Journal, Elsevier**, 2021. (**IF: 7.74**)

1.4. **A. Rahiche** and M. Cheriet. "Blind Decomposition of Multispectral Document Images using Orthogonal Nonnegative Matrix Factorization." In **IEEE Transactions on Image Processing**, Vol.30, p. 5997-6012, 2021. (**IF: 10.85**)

1.5. Y. E. Salehani, E. Arabnejad, **A. Rahiche**, A. Bakhta and M. Cheriet, "MSdB-NMF: MultiSpectral Document Image Binarization Framework via Non-Negative Matrix Factorization Approach." in **IEEE Transactions on Image Processing**, Vol. 29, p. 9099-9112, 2020. (**IF: 10.85**)

1.6. **A. Rahiche**, R. Hedjam, A. Al-maadeed, and M. Cheriet. "Historical documents dating using multispectral imaging and ordinal classification." in **Journal of Cultural Heritage**, Vol 45, p.71.80, 2020. (**IF: 2.95**)

2) **International Conferences:**

2.1. **A. Rahiche** and M. Cheriet. "Kernel Orthogonal Nonnegative Matrix Factorization: Application to Multispectral Document Image Decomposition." **ICASSP 2021, IEEE International Conference on Acoustics, Speech, and Signal Processing**, June 6-12, 2021, Toronto, Canada. p. 3275-3279. IEEE.

2.2. **A. Rahiche** and M. Cheriet. "Forgery Detection in Hyperspectral Document Images Using Graph Orthogonal Nonnegative Matrix Factorization." In Proceedings of the **IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR)**, p. 662-663, 2020.

2.3. **A. Rahiche** and M. Cheriet. "KFBin: Kalman filter-based approach for document image binarization." **International Conference on Image Analysis and Recognition**, 2019, August, p. 150-161. Springer.

CHAPTER 3

BLIND DECOMPOSITION OF MULTISPECTRAL DOCUMENT IMAGES USING ORTHOGONAL NONNEGATIVE MATRIX FACTORIZATION

Abderrahmane Rahiche[a] , Mohamed Cheriet[a]

[a] Department of Systems Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in IEEE Transaction on Image Processing, Vol. 30, p. 5997-6012, 2021.

Abstract

This paper addresses the challenge of Multispectral (MS) document image segmentation, which is an essential step for subsequent document image analysis. Most previous studies have focused only on binary (text/non-text) separation. They also rely on handcrafted features and techniques dedicated to conventional images that do not take advantage of MS images' spectral richness. In this work, we reformulate this task as a source separation problem, whereby we target the blind decomposition of entire MS document images via a new orthogonal nonnegative matrix factorization (ONMF). On the one hand, we incorporate orthogonality constraint as a Riemannian optimization on the Stiefel manifold. On the other hand, based on which factor we impose the orthogonality constraint, i.e., either on the endmember matrix, abundance matrix, or both, we propose three ONMF models to investigate this issue and determine which model is more suitable for this study. Minimizing the three models subject to nonnegativity and orthogonality constraints simultaneously is very challenging. Therefore, we extend the alternating direction method of multipliers scheme to solve them. We evaluated our models on synthetic Hyperspectral (HS) images and real-world MS document images. The experimental results confirm the effectiveness of the proposed models and demonstrate their generalization power compared with state-of-the-art techniques.

Keywords

Multispectral images, Orthogonal nonnegative matrix factorization, Stiefel manifold, Document image segmentation, Blind source separation.

3.1 Introduction

Multispectral (MS) imaging modalities provide higher spectral resolution than conventional color imaging systems. Traditional RGB (Red, Green, Blue) cameras operate in the visible spectrum, as they imitate the human visual system and capture images across three bands only. Conversely, MS imaging systems can capture high-resolution images at specific wavelengths from multiple electromagnetic spectrum bands, including the visible spectrum. Considering that objects reflect, absorb, or transmit different amounts of light energy at different wavelengths because of their physical properties, MS imaging measures specific spectral responses of scanned objects, which allows the characterization, identification, and discrimination of objects based on their unique spectral signatures.

Since the introduction of spectral imagery to the document image analysis field, it has emerged as a powerful tool for analyzing historical documents and paintings. This innovative technology has been demonstrated to be useful for the study of old documents (Fischer & Kakoulli, 2006), and it has become an attractive digitization technique in this field. Furthermore, it has enabled the derivation of new application perspectives such as pigment identification (Toque *et al.*, 2009), forgery detection, revealing hidden or erased writing (Tonazzini *et al.*, 2019), document image restoration (Hedjam & Cheriet, 2013b), and document age estimation (Rahiche *et al.*, 2020).

Document image segmentation is an important step in the process of document analysis. Nevertheless, this task is not trivial but rather highly challenging due to various issues related to this field, e.g., degradation of documents, especially historical ones, the complexity of scenes, the variability of contents, overlaps of objects, presence of stains, and noise, etc. Moreover, spectral data images have also raised additional challenges, such as the large volume of data that may contain redundant information and the variability of the number of images, which depends on the imaging system used and represents a big generalization issue. The research on MS document image processing is quite limited compared to other topics in this field, mainly due to the lack of datasets and the availability (cost) of MS systems. Thanks to the organizers of the first contest on multispectral text extraction (MS-TEx2015) (Hedjam *et al.*, 2015) for drawing the attention of the document image community to these kinds of problems.

Existing MS document image segmentation systems suffer from several significant drawbacks: i) most of them can perform only binary separation, i.e., text/non-text separation, as in (Tonazzini *et al.*, 2009b; Lettner & Sablatnig, 2010; Hedjam & Cheriet, 2013a); ii) usually they rely on traditional techniques developed for grey-level images, which requires manual selection of a single image from the whole stack of MS images, and do not benefit from their multidimensional nature (Hollaus *et al.*, 2015; Moghaddam & Cheriet, 2015); iii) many systems are conceived for a specific set of images and can not handle images of another spectral imager without heavy adaptations in this field, as in (Diem *et al.*, 2016; Hollaus, Diem & Sablatnig, 2018b); iv) moreover, the lack of labeled datasets thwarts the success of deep learning approaches, as in (Hollaus *et al.*, 2019). Therefore, all these challenges urge the development of new systems that exploit the spectral richness of MS images, don't target a specific application, and don't rely on labeled datasets.

In recent decades, Nonnegative Matrix Factorization (NMF) (Paatero & Tapper, 1994; Lee & Seung, 1999), which is a Blind Source Separation (BSS) approach (Cichocki *et al.*, 2009), has attracted significant attention from researchers in various areas of signal and image processing. NMF has been applied in many research fields, such as sound source separation (Virtanen, 2007), linear HS unmixing (Yang *et al.*, 2010; Bioucas-Dias *et al.*, 2012), and medical image analysis (Gobinet *et al.*, 2007). In particular, orthogonal NMF (ONMF), which is a variant of matrix factorizations with additional orthogonality constraints, has demonstrated successful clustering performance (Ding *et al.*, 2006) and is very effective for data mining. Surprisingly, despite the widespread use of NMF techniques in several fields, they are less attractive in the area of document image analysis.

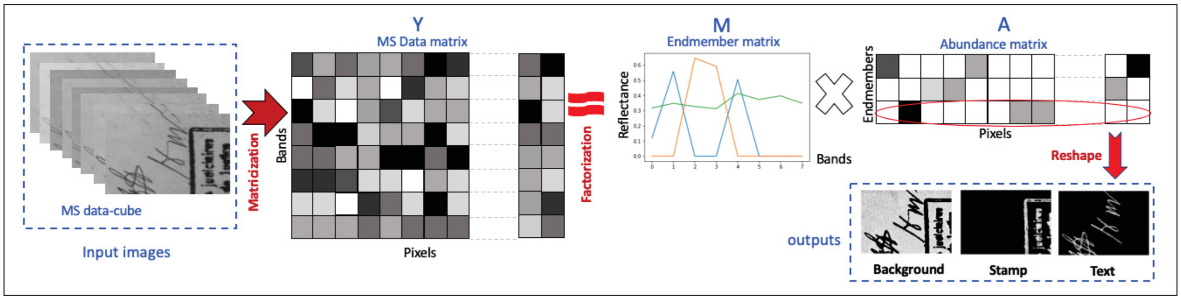


Figure 3.1 Flowchart of the proposed MS blind decomposition method

In this work, we address the problem of MS document image segmentation in the spirit of BSS[8], in which we consider document image data as a combination of superimposed layers of components. Thus, to separate between these components without prior knowledge about the mixture, we develop a blind decomposition approach based on a new orthogonal NMF formulation (see Fig. 3.1[9]). Moreover, to investigate which factor we should impose on the orthogonality constraint, we develop three orthogonal NMF models, two uni-orthogonal and one bi-orthogonal NMF model. Our approaches' key features are the incorporation of the orthogonality constraint as a Riemannian optimization over the Stiefel manifold (Absil, Mahony & Sepulchre, 2009), and the use of the Alternating Direction Method of Multipliers (ADMM) (Boyd *et al.*, 2011) to solve the formulated ONMF models. The advantages of the proposed models are: they neither rely on a specific number of bands, particular wavelengths, nor a given order of bands; they perform unsupervised full decomposition of MS document images; the factorization is more robust because the orthogonality constraint increases sparseness, reduces the number of local minima (Yang & Oja, 2010), and yields a unique solution; and due to the nature of ADMM, the solutions obtained are suitable for running in parallel.

[8] Thus, in some places, we use the two terms, i.e., decomposition and segmentation, interchangeably.

[9] The illustration corresponds to the uni-orthogonal NMF model described in Section 3.3.3. Outputs represent the fractional abundance maps of three endmembers (text, stamp, and paper). Each spatial map corresponds to one of the three rows of the abundance matrix reshaped back to the original input MS images' original size. The spatial maps are rescaled in the interval [0,1] and represented in gray-level, where black intensity means the full absence of the material, and white intensity indicates the material's full presence.

3.1.1 Related work

By considering the nonnegativity and orthogonality constraints simultaneously, the ONMF problem becomes more challenging than the standard NMF. Numerous ONMF models have been proposed in the literature. The main techniques adopted to handle the orthogonality can be broadly divided into the following strategies: most approaches incorporate orthogonality as an additive penalty term added to the cost function. For instance, in (Ding *et al.*, 2006), a hard orthogonality constraint is used, and a multiplicative update (MU) scheme has been adopted to solve the problem. The algorithm developed, called BiOR-NM3F, provided improved clustering performance over the K-means algorithm. The MU itself does not guarantee a global convergence and has been criticized for its slow convergence. To cope with the same problem, authors in (Li, Zhou & Cichocki, 2014) employed a soft orthogonality constraint as a penalty term. They adopted a Hierarchical Alternating Least Squares (HALS) and an Accelerated Proximate Gradient (APG) approach to optimize the two proposed models. A second strategy that does not rely on adding any penalty term to the cost function consists of projecting the gradient on a set of orthogonal matrices called the Stiefel manifold. In (Yoo & Choi, 2008), the natural (Euclidean) gradient on the Stiefel manifold is obtained by taking the objective function's derivative and then incorporating it into the tangent space equation. From which multiplicative updates are derived. The ONMF model obtained outperforms the standard NMF on document data clustering. In (Yoo & Choi, 2010), the authors used the same strategy proposed in (Choi, 2008) and (Yoo & Choi, 2008) to develop a bi-orthogonal NMF model called orthogonal nonnegative tri-factorization (ONMTF). Alternatively, Pompili *et al*. (Pompili, Gillis, Absil & Glineur, 2014) proposed a projected gradient approach to solve the sub-problem with orthogonality constraint. At each iteration, the algorithm minimizes the sub-problem by moving in the descent direction and then projecting the solution onto a feasible set of orthogonal matrices called Stiefel manifold; iii) another strategy that does not require the gradient information is based on a variables splitting technique called SOC (Splitting Orthogonality Constraints) (Lai & Osher, 2014), which is based on Bregman iterations. Based on SOC, the authors in (Rahiche & Cheriet, 2020) formulated a closed-form solution obtained via the SVD factorization, which is the optimal solution without nonnegativity. The proposed model performed well in the task of inks mismatch detection. A summary of the main ONMF approaches with their corresponding formulation, the optimization method used, and the application targeted is given in Table I-1, Appendix I.

In this work, we incorporate the orthogonality constraint in our ONMF models using a different strategy. We strictly enforce the orthogonality via Riemannian optimization over the Stiefel manifold, which has some advantages over other strategies. Riemannian optimization ensures the orthogonality of variables by projecting them onto an orthogonal manifold and keeps them orthogonal during their update. Besides, the corresponding optimization problem is solved by gradient-based approaches over a non-Euclidean space (which is more appropriate for several computer vision applications). The main steps involved are introduced hereafter in Section.3.2 and developed in detail in Section.3.3. As a use case, we validate our models on MS document image decomposition task, which is a very challenging problem, and less attention has been paid to it in the literature.

3.1.2 Contributions

We highlight the contributions of this work as follows:

1) We propose an unsupervised approach for blind decomposition of MS document images into their constitutive components via new orthogonal NMF models.
2) Instead of considering additive orthogonality constraints, we incorporate the orthogonality as an optimization over the Stiefel manifold. The nonnegativity is handled using an indicator function.
3) We propose three different ONMF models, two uni-orthogonal and one bin-orthogonal model, to investigate the issue of which factor we should be orthogonally constrained, i.e., the columns of the endmember matrix, the rows of the abundance matrix, or on both of them.
4) To cope with the simultaneous nonnegativity and orthogonality constraints, we propose a multi-block ADMM algorithm to solve the formulated problems.

5) We conduct extensive comparative analysis using synthetic HS images and several real-world MS document images, and we compare our results to several state-of-the-art techniques.

The remainder of the paper is structured as follows. Section 3.2 gives all the notations and the definitions of the main concepts used in this paper. In Section 3.3, we present the proposed models and the corresponding optimization algorithms. The results of our experiments are given in Section3.5. Finally, Section 3.6 concludes the paper and discusses future directions.

3.2 Notations and preliminaries

To make the paper self-contained, we describe in this section the notations used throughout this document, and we briefly introduce the main terminologies utilized in this paper, namely the Linear Mixture Model (LMM), NMF, Riemannian optimization over the Stiefel manifold, and ADMM.

3.2.1 Notations

To ease reading the paper, Table 3.1 gives a list of the main notations used hereafter.

Table 3.1 Notations used throughout this paper

Notations	Descriptions
$\mathbf{A}$, $\mathbf{a}$, a	Matrix, vector, and scalar of appropriate sizes
$\mathbf{A}^T$, $\mathbf{A}^{-1}$	Transpose and inverse of a matrix.
$\mathbb{R}_+^{b \times n}$	Set of b-by-n positive real-valued matrices.
b,n,k	Number of bands, pixels, and endmembers, respectively.
$\mathbf{I}_k$	k-by-k identity matrix.
$\max(.,0)$	Element-wise maximum of two matrices.
$\|.\|_F$	Frobenius norm.
$\nabla f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$	Gradient (first order derivative) of $f(\mathbf{X})$ w.r.t $\mathbf{X}$.
$\langle \mathbf{A}, \mathbf{B} \rangle = tr(\mathbf{A}^T \mathbf{B})$	Euclidean inner product.
$Tr(.)$	Matrix trace.

3.2.2 Linear Mixture Model (LMM)

The linear mixture model (LMM) (Settle & Drake, 1993) assumes that the reflectance of a given pixel in MS/HS images can be described by a linear combination of all the pure spectral signatures presented in that pixel. Under this assumption, MS/HS data can be modelled as follows:

$$\mathbf{Y} = \mathbf{M}\mathbf{A} + \mathbf{N}, \tag{3.1}$$

where $\mathbf{Y} \in \mathbb{R}_{+}^{b \times n}$ is a data matrix representing a sequence of b spectral images having n pixels each, where each row represents a vectorization of one image, $\mathbf{M} \in \mathbb{R}_{+}^{b \times k}$ is the endmember matrix, which is positive by nature, $\mathbf{A} \in \mathbb{R}^{k \times n}$ is the mixing (abundance) matrix that contains the coefficient of spectral signatures proportion, $\mathbf{N} \in \mathbb{R}^{b \times n}$ represents the matrix of additional noise, and k is the number of endmembers. Generally, $\mathbf{M}$, $\mathbf{A}$, $\mathbf{N}$, and the number k are all unknown.

3.2.3 Nonnegative Matrix Factorization (NMF)

The NMF aims to factorize an input positive data matrix (e.g. $\mathbf{Y}$) into two nonnegative latent matrices (e.g. $\mathbf{M}$, $\mathbf{A}$), that is, $\mathbf{Y} \approx \mathbf{M}\mathbf{A}$. For simplicity, we keep the same notations defined above in Eq.(3.1). The NMF problem can then be formulated as:

$$\min_{\mathbf{M},\mathbf{A}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_F^2, \quad \text{subject to} \quad \mathbf{M} \geqslant 0, \mathbf{A} \geqslant 0. \tag{3.2}$$

In the literature, beside the nonnegativity of the endmembers matrix, i.e., $\mathbf{M} \geq 0$, two other physical constraints are generally imposed to the abundance matrix (fractions of materials), namely, *the abundance nonnegativity constraint* (ANC), and *the abundance sum-to-one constraint* (ASC) (Settle & Drake, 1993; Heinz *et al.*, 2001), that is:

$$\mathbf{A} \geq 0, \qquad \text{(ANC)}, \tag{3.3}$$

$$\mathbf{1}_k^T \mathbf{A} = \mathbf{1}_n^T, \qquad \text{(ASC)}, \tag{3.4}$$

where, $\mathbf{1}_k^T$ and $\mathbf{1}_n^T$ are two vectors of size k and n, respectively, with all-one elements. The ASC condition requires that the sum of the abundance fractions in each column of the abundance matrix $\mathbf{A}$ must equal to one, i.e, $\sum_{i=1}^{k} \mathbf{A}_{ij} = 1$

3.2.4 Riemannian optimization on the Stiefel Manifold

Riemannian optimization over the Stiefel manifold is an elegant framework for solving problems under orthogonality assumptions. The Stiefel manifold is the set of all orthogonal matrices of a given size. Strictly enforcing the orthogonality constraints on matrices via Riemannian optimization over the Stiefel manifold has been shown to guarantee the orthogonality better than other techniques.

Riemannian optimization on the Stiefel manifold addresses problems of the form

$$\begin{aligned} &\min_{\mathbf{X}} f(\mathbf{X}), \\ &\text{subject to} \quad \mathbf{X} \in \mathcal{M} := St(k,n) = \{\mathbf{X} \in \mathbb{R}^{n\times k} | \mathbf{X}^T\mathbf{X} = \mathbf{I}_k\}, \end{aligned} \tag{3.5}$$

where, $f : \mathcal{M} \rightarrow \mathbb{R}$ is the objective function, the set $\mathcal{M}$ is called the Stiefel manifold of dimension $nk - \frac{1}{2}k(k+1)$, which consists of the set of rectangular matrices of fixed size with orthonormal columns (Absil *et al.*, 2009), $\mathbf{I}_k$ represents the k-by-k identity matrix ($k \leq n$).

The main idea of Riemannian optimization over the Stiefel manifold is to enforce the orthogonality by performing gradient descent iterations on the Stiefel manifold itself instead of the original Euclidean space. This can be achieved through the following steps. 1) The natural gradient of the objective function at a starting point is first calculated in the Euclidean space. 2) Using a proper projection operator, the Riemannian gradient is then obtained by projecting the natural (Euclidean) gradient onto the tangent space. 3) After calculating the new step, the update of the next iteration is mapped back to the original space (manifold) via an appropriate retraction operator. These steps are explained with more details hereafter. Fig. 3.2 illustrates a geometrical representation of the Stiefel manifold and the corresponding tangent space at a point X from the manifold.

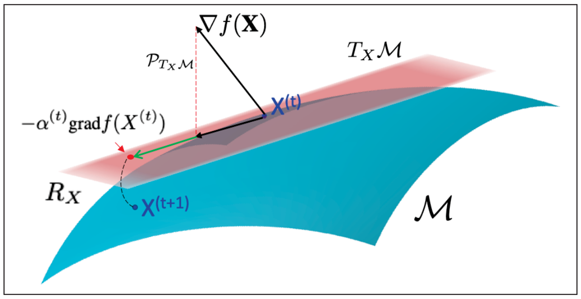


Figure 3.2 Graphical representation of the optimization over the Stiefel manifold

As shown in Fig. 3.2, the tangent space at point $\mathbf{X} \in \mathcal{M}$ is denoted $T_X\mathcal{M}$.

Lemma 1. *((Tagare, 2011)) If $\boldsymbol{U} \in T_X\mathcal{M}$, then $\boldsymbol{U}$ satisfies*

$$\boldsymbol{X}^T\boldsymbol{U} + \boldsymbol{U}^T\boldsymbol{X} = 0.$$

Proof. (adapted from (Tagare, 2011)) Let $C(t)$ be a curve defined in $\mathcal{M}$. The equation defining tangents to the Stiefel manifold is obtained by differentiating $C(t)^T C(t) = \mathbf{I}$ w.r.t t, yielding $C'(t)^T C(t) + C(t)^T C'(t) = 0$. By setting $C(0) = \mathbf{X}$ and $C'(0) = U$ at $t_0 = 0$, we obtain the corresponding tangent equation.

The Euclidian (natural) gradient of a smooth function $f(\mathbf{X})$ is denoted by ∇f, and the Riemannian gradient of $f(X)$ is denoted as grad$f$. The mathematical relationship between ∇f and grad$f$ is given by:

$$\mathrm{grad} f(\mathbf{X}) = \mathcal{P}_{T_X\mathcal{M}}(\nabla f(\mathbf{X})), \tag{3.6}$$

where $\mathcal{P}_{T_X\mathcal{M}}$ represents the orthogonal projection of $\nabla f(\mathbf{X})$ onto the tangent space $T_X\mathcal{M}$ and is defined as (Absil *et al.*, 2009, Example 3.6.2):

$$\mathcal{P}_{T_X\mathcal{M}}(\nabla f) = (\mathbf{I}_n - \mathbf{X}\mathbf{X}^T)\nabla f + \frac{1}{2}\mathbf{X}(\mathbf{X}^T\nabla f - \nabla f^T\mathbf{X}) \tag{3.7}$$

$$= \nabla f - \mathbf{X}\mathrm{Sym}(\mathbf{X}^T\nabla f), \tag{3.8}$$

where $\mathrm{Sym}(\mathbf{D}) := \frac{1}{2}(\mathbf{D} + \mathbf{D}^T)$ denotes the decomposition of $\mathbf{D}$ into a sum of two symmetric terms.

This orthogonal project $\mathcal{P}_{T_X\mathcal{M}}(\nabla f)$ is mapped back from the tangent plane $T_X\mathcal{M}$ to the manifold $\mathcal{M}$ via a retraction operator R_X:

$$\mathbf{X}^{(t+1)} = R_X(-\alpha^{(t)}\mathrm{grad}f(\mathbf{X}^{(t)})), \tag{3.9}$$

where $\alpha^{(t)}$ is the step size along the descent direction. The retraction on the Stiefel manifold based on polar decomposition has the following closed form expression (Absil *et al.*, 2009, Example 4.1.3):

$$R_X(\mathbf{Z}) = (\mathbf{X} + \mathbf{Z})(\mathbf{I}_P + \mathbf{Z}^T\mathbf{Z})^{-1/2}, \tag{3.10}$$

where $\mathbf{Z}$ satisfies $\mathbf{X}^T\mathbf{Z} + \mathbf{Z}^T\mathbf{X} = 0$.

A general framework for optimization over the Stiefel manifold is given in Algorithm 3.2. We refer the reader to (Absil *et al.*, 2009) for more details about the conceptual status of manifold optimization.

3.2.5 Alternating Direction Method of Multipliers (ADMM)

ADMM (Boyd *et al.*, 2011) is a powerful variable splitting technique in optimization designed to solve problems of the form

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{z}\in\mathcal{Z}} f(\mathbf{x}) + g(\mathbf{z}) \quad \text{s.t.} \quad \mathbf{B}\mathbf{x} + \mathbf{C}\mathbf{z} = \mathbf{d}, \tag{3.11}$$

where f and g are two convex functions, $\mathcal{X}$ and $\mathcal{Z}$ are two closed convex subsets, **B** and **C** are two matrices, and **d** is a vector of appropriate size. The augmented Lagrangian function of (3.11) is

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mu) = f(\mathbf{x}) + g(\mathbf{z}) + \langle \mu, \mathbf{Bx} + \mathbf{Cz} - \mathbf{d} \rangle + \frac{\rho}{2} \|\mathbf{Bx} + \mathbf{Cz} - \mathbf{d}\|_2^2, \tag{3.12}$$

where μ denotes a Lagrangian multiplier vector, and $\rho > 0$ is a penalty parameter. By combining the linear and quadratic terms in the augmented Lagrangian function and scaling the dual variable, Eq.(3.12) can be expressed as

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \lambda) = f(\mathbf{x}) + g(\mathbf{z}) + \frac{\rho}{2} \|\mathbf{r} + \lambda\|_2^2, \tag{3.13}$$

where $\lambda = \mu/\rho$ denotes the scaled dual variable, and $\mathbf{r} = \mathbf{Bx} + \mathbf{Cz} - \mathbf{d}$, is the residual.

ADMM (Boyd *et al.*, 2011) splits the problem in Eq.(3.13) into two sub-problems so that each sub-problem of them is easier to solve, and it performs alternating minimization with respect to each variable individually. it updates then the scaled dual variable λ. The iterative updating scheme is summarized as follows:

$$\mathbf{x}^{t+1} := \arg\min_{\mathbf{x}} (f(\mathbf{x}) + \frac{\rho}{2} \|\mathbf{Bx} + \mathbf{Cz}^t - \mathbf{d} + \lambda^t\|_2^2) \tag{3.14}$$

$$\mathbf{z}^{t+1} := \arg\min_{\mathbf{z}} (g(\mathbf{z}) + \frac{\rho}{2} \|\mathbf{Bx}^{t+1} + \mathbf{Cz} - \mathbf{d} + \lambda^t\|_2^2) \tag{3.15}$$

$$\lambda^{t+1} := \lambda^t + \mathbf{Bx}^{t+1} + \mathbf{Cz}^{t+1} - \mathbf{d}. \tag{3.16}$$

3.3 Problem Formulation

Considering the problem given in Eq.(3.2), the orthogonality constraint can be imposed either on the columns[10] of the endmembers matrix **M**, the rows of the abundance matrix **A**, or on both of them. Enforcing the orthogonality of any factor matrix, while strictly imposing the

[10] Similarly, the rows of the endmember matrix **M** could be also constrained to be orthogonal to investigate the band-to-band correlation issue. However, this issue is out the scope of this work.

nonnegativity on it, yields a sparser solution. Ideally, the optimal solution contains only one non-zero element in each column/row. This property is, in general, highly appreciated for many clustering applications. As a benefit, imposing the orthogonality obviates adding additional constraints to our NMF models, such as sparsity.

From a physical point of view, imposing the orthogonality on the endmembers matrix **M** is equivalent to assuming the non-correlation (statistical independence) (Karoui, Deville, Hosseini & Ouamri, 2012) of its columns' endmembers (materials). Indeed, this assumption is highly recommended in the literature (Van der Meer & De Jong, 2000; Van der Meer & Jia, 2012), as it reduces the error in the endmembers matrix inversion and improves the results of spectral unmixing. Imposing the orthogonality of abundance matrix A, i.e., orthogonal rows, alongside the nonnegativity yields a sparser solution, i.e., fewer non-zero elements in each column, which implies the disjointness of materials. Imposing the orthogonality on both factors at the same time yields materials indicators that do not overlap in the spatial domain and their sources do not correlate in the spectral domain.

Nevertheless, because of the ASC condition discussed above (3.4), the abundance matrix **A** can not be constrained to be orthogonal while obeying the ASC constraint at the same time. Consequently, to overcome this issue, we relax the problem with the following two assumptions: 1) The ASC condition has been criticized in the literature (Iordache *et al.*, 2011, pp.2018), owing to a strong signature variability (Bateson, Asner & Wessman, 2000) present in real-world spectral images. 2) According to the ANC condition (3.3), the coefficients are allowed to be ≥ 0 . Hence, if we assume that the corresponding fractions for some materials in a given pixel are zero, we can still model that pixel's reflected spectrum as a linear combination of those materials with some zero coefficients. According to these assumptions, the ASC is relaxed in this study, which fulfills the orthogonality condition as well.

Hence, following this discussion, and in order to evaluate the effect of imposing the orthogonality constraint on each matrix factor, we consider the three cases mentioned above, namely, orthogonal endmembers matrix, orthogonal abundance matrix, and finally bi-orthogonal factors. The detailed description of each model is given in the following subsections.

3.3.1 Uniqueness of Orthogonal NMF

NMF is an ill-posed problem. In general, the NMF of a given positive data matrix $\mathbf{Y} \in \mathbb{R}^{b \times n}$ has a large number of possible solutions. However, by considering additional constraints on the factor variables or incorporating proper regularization terms into the objective function, NMF can have a solution that is unique up to standard BSS indeterminacies, i.e., permutation and scaling (Hoyer, 2004; Ding *et al.*, 2006; Gillis, 2012).

For the same data matrix that we assume to be a full rank, we can find another set of solution $(\mathbf{M}', \mathbf{A}')$, such that

$$\mathbf{Y} = \mathbf{MA} = \mathbf{M}'\mathbf{A}', \text{ s.t } \mathbf{M}, \mathbf{A}, \mathbf{M}', \mathbf{A}' \geqslant 0. \tag{3.17}$$

It is easy to show that there exists a nonsingular matrix $\mathbf{G} \in \mathbb{R}^{r \times r}$ that satisfies

$$\mathbf{M}' = \mathbf{MG}, \quad \mathbf{A}' = \mathbf{G}^{-1}\mathbf{A}. \tag{3.18}$$

It follows that all other full rank factorizations are of the form $\mathbf{Y} = \mathbf{MGG}^{-1}\mathbf{A}$. In this case the only ambiguity lies in the matrix $\mathbf{G}$.

Lemma 2. *(Minc, 1988, Lemma 1.1). The inverse of a nonnegative matrix* $\mathbf{G}$ *is also nonnegative if and only if* $\mathbf{G}$ *is a generalized permutation matrix (monomial matrix) (see (Minc, 1988) for a proof).*

Definition 3.1. *The problem in Eq.(3.17) has unique solution if the permutation and the scaling indeterminacies of the columns in* $\boldsymbol{M}$ *and rows in* $\boldsymbol{A}$ *are the only sources of ambiguities.*

Based on Lemma.2, and Definition.3.1, we deduce the following result.

Proposition 1. *In the case of orthogonal NMF, with the condition $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$, the problem of Eq.(3.17) implies $\boldsymbol{G} = \boldsymbol{P}$, where $\boldsymbol{P}$ is a permutation matrix, which satisfies Eq.(3.17), Eq.(3.18), and the orthogonality condition $\boldsymbol{G}^{-1}\boldsymbol{A}(\boldsymbol{G}^{-1}\boldsymbol{A})^T = \boldsymbol{I}$.*

Proof. The orthogonality condition $(\mathbf{G}^{-1}\mathbf{A})(\mathbf{G}^{-1}\mathbf{A})^T = \mathbf{I}$ implies $\mathbf{G}^{-1}\mathbf{A}\mathbf{A}^T(\mathbf{G}^{-1})^T = \mathbf{I}$. We have $\mathbf{A}\mathbf{A}^T = \mathbf{I}$, which implies $(\mathbf{G}^T\mathbf{G})^{-1} = \mathbf{I}$. Multiplying $(\mathbf{G}^T\mathbf{G})$ in both sides we get $\mathbf{G}^T\mathbf{G} = \mathbf{I}$. A trivial choice of $\mathbf{G}$ would be $\mathbf{I}$, however, this is also satisfied by a permutation matrix. This removes the scaling ambiguity. Q.E.D.

3.3.2 M-orthogonal NMF

In the first model, which we call M-ONMF, we assume that all the endmembers are statistically independent. Accordingly, we force the columns of the matrix $\mathbf{M}$ to be orthogonal. Furthermore, the nonnegativity constraints on $\mathbf{M}$ and $\mathbf{A}$ are handled explicitly by incorporating an indicator functions $\iota_{\mathcal{S}}$ of the set $\mathcal{S} \subset \mathbb{R}_+$ that is defined as:

$$\iota_{\mathcal{S}}(\mathbf{C}) = \begin{cases} 0 & \text{if} \quad \mathbf{C} \in \mathcal{S}, \\ +\infty & \text{if not.} \end{cases} \tag{3.19}$$

Hence, we can write the corresponding mathematical model as follows:

$$\min_{\mathbf{M},\mathbf{A}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_F^2 + \iota_{\mathcal{S}}(\mathbf{M}) + \iota_{\mathcal{S}}(\mathbf{A}), \qquad \text{subject to:} \quad \mathbf{M}^T\mathbf{M} = \mathbf{I}_k. \tag{3.20}$$

We relax the problem by introducing two auxiliary variables $\mathbf{X} \in \mathbb{R}_+^{b\times k}$ and $\mathbf{Z} \in \mathbb{R}_+^{k\times n}$, and we formulate the following equivalent model

$$\min_{\mathbf{M},\mathbf{A},\mathbf{X},\mathbf{Z}} = \frac{1}{2}\|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_F^2 + \iota_{\mathcal{S}}(\mathbf{X}) + \iota_{\mathcal{S}}(\mathbf{Z}), \tag{3.21}$$

$$\text{subject to:} \quad \mathbf{M} = \mathbf{X}, \mathbf{A} = \mathbf{Z}, \mathbf{M} \in \mathcal{M}_1, \tag{3.22}$$

where $\mathcal{M}_1 := \{\mathbf{M} \in \mathbb{R}^{b\times k} | \mathbf{M}^T\mathbf{M} = \mathbf{I}_k\}$ is the Stiefel manifold for $\mathbf{M}$. The corresponding augmented Lagrangian function is

$$\mathcal{L}_1(\mathbf{M}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{\Lambda_1}, \mathbf{\Lambda_2}) = \frac{1}{2}||\mathbf{Y}-\mathbf{MA}||_F^2 + \iota_{\mathcal{S}}(\mathbf{X}) + \iota_{\mathcal{S}}(\mathbf{Z}) + \frac{\rho_x}{2}||\mathbf{M}-\mathbf{X}+\mathbf{\Lambda_1}||_2^2 + \frac{\rho_z}{2}||\mathbf{A}-\mathbf{Z}+\mathbf{\Lambda_2}||_2^2, \tag{3.23}$$

where $\mathbf{\Lambda_1} \in \mathbb{R}^{b\times k}$ and $\mathbf{\Lambda_2} \in \mathbb{R}^{k\times n}$ are the scaled Lagrangian multipliers (dual variables) associated with linear equality constraints in Eq.(3.22), and ρ_x and $\rho_z > 0$ are the corresponding penalty parameters. Applying the ADMM scheme to the problem of Eq.(3.23) yields the following updates:

$$\mathbf{M}^{t+1} := \arg\min_{\mathbf{M}\in\mathcal{M}_1} \frac{1}{2}||\mathbf{Y} - \mathbf{MA}||_F^2 + \frac{\rho_x}{2}||\mathbf{M} - \mathbf{X} + \mathbf{\Lambda_1}||_2^2, \tag{3.24}$$

$$\mathbf{A}^{t+1} := \arg\min_{\mathbf{A}} \frac{1}{2}||\mathbf{Y} - \mathbf{M}^{t+1}\mathbf{A}||_F^2 + \frac{\rho_z}{2}||\mathbf{A} - \mathbf{Z} + \mathbf{\Lambda_2}||_2^2, \tag{3.25}$$

$$\mathbf{X}^{t+1} := \arg\min_{\mathbf{X}\geqslant 0} \iota_{\mathcal{S}}(\mathbf{X}) + \frac{\rho_x}{2}||\mathbf{M}^{t+1} - \mathbf{X} + \mathbf{\Lambda}_1||_2^2, \tag{3.26}$$

$$\mathbf{Z}^{t+1} := \arg\min_{\mathbf{Z}\geqslant 0} \iota_{\mathcal{S}}(\mathbf{Z}) + \frac{\rho_z}{2}||\mathbf{A}^{t+1} - \mathbf{Z} + \mathbf{\Lambda}_2||_2^2, \tag{3.27}$$

$$\mathbf{\Lambda}_1^{t+1} := \mathbf{\Lambda}_1 + \mathbf{M}^{t+1} - \mathbf{X}^{t+1}, \tag{3.28}$$

$$\mathbf{\Lambda}_2^{t+1} := \mathbf{\Lambda}_2 + \mathbf{A}^{t+1} - \mathbf{Z}^{t+1}, \tag{3.29}$$

where t is the iterative variable.

Now, we can solve each subproblem separately, as an unconstrained problem, with an easier formulation, and in an iterative fashion. Due to the space restrictions, the optimization details are given in Section 2.1 of Appendix I. Algorithm 3.1 summaries the obtained iterative steps to solve the M-ONMF model. Note that the $prox_{\iota_{\mathcal{S}}}$ operator denotes the elements-wise projection on a positive set, i.e., $prox_{\iota_{\mathcal{S}}}(.) = max(., 0)$.

The $\mathbf{M}$ update step is solved using Algorithm 3.2, which summarizes the general framework of an optimization over the Stiefel manifold. In our experiment, this step is performed using the *Pymanopt* solver (Townsend, Koep & Weichwald, 2016). We note that the step size parameter α

Algorithm 3.1 M-ONMF model

```
**Input:** $\mathbf{Y}$, $k$, $\rho_x$, $\rho_z$, *Tol*, *niter*
**Initialization:** $\mathbf{M}^0$, $\mathbf{A}^0$, $\mathbf{X}^0 = \mathbf{A}^0$, $\mathbf{Z}^0 = \mathbf{M}^0$, $\mathbf{\Lambda_1} = 0$, $\mathbf{\Lambda_2} = 0$
**for** $i \leftarrow 1$ *niter* **do**
    Update $\mathbf{M}$ by solving Eq.(3.24) over $\mathcal{M}_1$ using Algorithm 3.2.
    $\mathbf{A} = (\mathbf{M^T M} + \rho_z I)^{-1}(\mathbf{M^T Y} + \rho_z(\mathbf{Z} - \mathbf{\Lambda_2}))$
    $\mathbf{X} = \mathrm{prox}_{\iota_S}(\mathbf{M} + \mathbf{\Lambda_1})$
    $\mathbf{Z} = \mathrm{prox}_{\iota_S}(\mathbf{A} + \mathbf{\Lambda_2})$
    $\mathbf{\Lambda_1} = \mathbf{\Lambda_1} + \mathbf{M} - \mathbf{X}$
    $\mathbf{\Lambda_2} = \mathbf{\Lambda_2} + \mathbf{A} - \mathbf{Z}$
    **if** $(\|r_1[i] - r_1[i-1]\|/\|r_1[i-1]\| < Tol)$ **then**
        break.
    **end if**
**end for**
**return** $\mathbf{M}$, $\mathbf{A}$
```

is selected using a line search strategy that satisfies the Armijo condition in each step, for more details see (Townsend *et al.*, 2016).

Algorithm 3.2 Framework of Riemannian optimization on the Stiefel manifold

```
**Input:** $\mathcal{L}$
**repeat**
    Compute the natural gradient $\nabla_{\mathbf{M}}\mathcal{L}$.
    Compute the Riemannian gradient using Eq.(3.6): $\mathrm{grad}\mathcal{L} = P_{T_{\mathbf{X}}\mathcal{M}}(\nabla_{\mathbf{M}}\mathcal{L})$.
    Update the step size $\alpha$
    Compute the retraction using Eq.(3.10): $\mathbf{M} \leftarrow R_X(-\alpha\mathrm{grad}\mathcal{L})$
    $i \leftarrow i + 1$
**until** *Convergence or i* $\geq$ *niter*;
**return X**
```

3.3.3 A-orthogonal NMF

For the second ONMF model, we instead constrain the rows of the abundance matrix **A** to be orthogonal, and we call it A-ONMF. We follow the same procedures, and we take the same steps as we did in the first model. Thus, we can formulate the A-ONMF model as follows:

$$\min_{\mathbf{M},\mathbf{A}} \quad \frac{1}{2}\|\mathbf{Y}-\mathbf{MA}\|_F^2 + \iota_{\mathcal{S}}(\mathbf{M}) + \iota_{\mathcal{S}}(\mathbf{A}), \qquad \text{subject to:} \quad \mathbf{AA}^T = \mathbf{I}_k. \tag{3.30}$$

As in the first model above, we relax the problem of Eq.(3.30) by introducing two auxiliary variables, $\mathbf{U} \in \mathbb{R}_+^{b\times k}$ and $\mathbf{V} \in \mathbb{R}_+^{k\times n}$, and we rewrite the model as follows:

$$\min_{\mathbf{M},\mathbf{A},\mathbf{U},\mathbf{V}} \quad \frac{1}{2}\|\mathbf{Y}-\mathbf{MA}\|_F^2 + \iota_{\mathcal{S}}(\mathbf{U}) + \iota_{\mathcal{S}}(\mathbf{V}), \tag{3.31}$$

$$\text{subject to:} \quad \mathbf{M} = \mathbf{U}, \mathbf{A} = \mathbf{V}, \mathbf{A} \in \mathcal{M}_2, \tag{3.32}$$

where $\mathcal{M}_2 := \{\mathbf{A}^T \in \mathbb{R}^{n\times k} | \mathbf{AA}^T = \mathbf{I}_k\}$ is the Stiefel manifold for $\mathbf{A}$. The corresponding augmented Lagrangian function of (3.31) is written as:

$$\begin{aligned}\mathcal{L}_2(\mathbf{M},\mathbf{A},\mathbf{U},\mathbf{V},\mathbf{\Pi_1},\mathbf{\Pi_2}) = {} & \frac{1}{2}\|\mathbf{Y}-\mathbf{MA}\|_F^2 + \iota_{\mathcal{S}}(\mathbf{U}) + \iota_{\mathcal{S}}(\mathbf{V}) \\ & + \frac{\rho_u}{2}\|\mathbf{M}-\mathbf{U}+\mathbf{\Pi}_1\|_2^2 + \frac{\rho_v}{2}\|\mathbf{A}-\mathbf{V}+\mathbf{\Pi}_2\|_2^2, \end{aligned} \tag{3.33}$$

where $\mathbf{\Pi_1} \in \mathbb{R}^{b\times k}$ and $\mathbf{\Pi_2} \in \mathbb{R}^{k\times n}$ are two Lagrangian dual variables associated with linear equality constraints in Eq.(3.32), and ρ_u and $\rho_v > 0$ are the penalty parameters.

Applying the ADMM paradigm to the problem of Eq.(3.33) yields the following updates

$$\mathbf{M}^{t+1} := \arg\min_{\mathbf{M}} \frac{1}{2}\|\mathbf{Y}-\mathbf{MA}\|_F^2 + \frac{\rho_u}{2}\|\mathbf{M}-\mathbf{U}+\mathbf{\Pi}_1\|_2^2, \tag{3.34}$$

$$\mathbf{A}^{t+1} := \arg\min_{\mathbf{A}\in\mathcal{M}_2} \frac{1}{2}\|\mathbf{Y}-\mathbf{M}^{t+1}\mathbf{A}\|_F^2 + \frac{\rho_u}{2}\|\mathbf{A}-\mathbf{V}+\mathbf{\Pi}_2\|_2^2, \tag{3.35}$$

$$\mathbf{U}^{t+1} := \arg\min_{\mathbf{U}\geqslant 0} \iota_{\mathcal{S}}(\mathbf{U}) + \frac{\rho_u}{2}\|\mathbf{M}^{t+1}-\mathbf{U}+\mathbf{\Pi}_1\|_2^2, \tag{3.36}$$

$$\mathbf{V}^{t+1} := \arg\min_{\mathbf{V}\geqslant 0} \iota_{\mathcal{S}}(\mathbf{V}) + \frac{\rho_v}{2}\|\mathbf{A}^{t+1}-\mathbf{V}+\mathbf{\Pi}_2\|_2^2, \tag{3.37}$$

$$\mathbf{\Pi}_1^{t+1} := \mathbf{\Pi}_1 + \mathbf{M}^{t+1} - \mathbf{U}^{t+1}, \tag{3.38}$$

$$\mathbf{\Pi}_2^{t+1} := \mathbf{\Pi}_2 + \mathbf{A}^{t+1} - \mathbf{V}^{t+1}. \tag{3.39}$$

Solving these subproblems yields the iterations summarized in Algorithm 3.3. The details of derivations are omitted here due to the space limit and are given in Section 2.2 of Appendix I.

Algorithm 3.3 A-ONMF model

```
Input: $\mathbf{Y}$, $k$, $\rho_u$, $\rho_v$, $Tol$, $niter$
Initialization: $\mathbf{M}^0$, $\mathbf{A}^0$, $\mathbf{U}^0 = \mathbf{A}^0$, $\mathbf{V}^0 = \mathbf{M}^0$, $\mathbf{\Pi_1} = 0$, $\mathbf{\Pi_2} = 0$
for $i \leftarrow 1$ $niter$ do
    $\mathbf{M} = (\mathbf{Y}\mathbf{A}^\mathbf{T} + \rho_u(\mathbf{U} - \mathbf{\Pi_1}))(\mathbf{A}\mathbf{A}^\mathbf{T} + \rho_u I)^{-1}$
    Update $\mathbf{A}$ by solving (3.35) using Algorithm 3.2.
    $\mathbf{U} = \mathrm{prox}_{\iota_S}(\mathbf{A} + \mathbf{\Pi_1})$
    $\mathbf{V} = \mathrm{prox}_{\iota_S}(\mathbf{M} + \mathbf{\Pi_2})$
    $\mathbf{\Pi_1} = \mathbf{\Pi_1} + \mathbf{M} - \mathbf{U}$
    $\mathbf{\Pi_2} = \mathbf{\Pi_2} + \mathbf{A} - \mathbf{V}$
    if $(\|r_2[i] - r_2[i-1]\|/\|r_2[i-1]\| < Tol)$ then
        break.
    end if
end for
return $\mathbf{M}$, $\mathbf{A}$
```

3.3.4 Bi-orthogonal NMF

As for the last model, the orthogonality constraint is imposed on both factor matrices at the same time. However, it turns out that this could be too strict. Thus, it is reasonable to relax the problem, as in (Ding *et al.*, 2006), by adding a third factor in the middle, that is not orthogonally constrained. This relaxation leads to the so-called nonnegative matrix tri-factorization (NTF), which is another variant of the standard NMF. The resulting bi-orthogonal NTF model is called hereafter MA-ONTF and is formulated as follows:

$$\begin{aligned} &\min_{\mathbf{M},\mathbf{Q},\mathbf{A}} \quad \frac{1}{2}\|\mathbf{Y}-\mathbf{MQA}\|_F^2, \qquad (3.40)\\ &\text{subject to:} \quad \mathbf{M} \geqslant 0, \mathbf{Q} \geqslant 0, \mathbf{A} \geqslant 0, \mathbf{M}^T\mathbf{M} = \mathbf{I}_k, \mathbf{A}\mathbf{A}^T = \mathbf{I}_k, \end{aligned}$$

where $\mathbf{Q} \in \mathbb{R}^{k\times k}$. Here, we introduce three auxiliary variables: $\mathbf{F} \in \mathbb{R}^{b\times k}$, $\mathbf{G} \in \mathbb{R}^{b\times k}$, and $\mathbf{H} \in \mathbb{R}^{k\times n}$ to relax this problem such that

$$\min_{\mathbf{M},\mathbf{Q},\mathbf{A},\mathbf{F},\mathbf{G},\mathbf{H}} \quad \frac{1}{2}\|\mathbf{Y}-\mathbf{FA}\|_F^2 + \iota_{\mathcal{S}}(\mathbf{G}) + \iota_{\mathcal{S}}(\mathbf{H}) \qquad (3.41)$$

$$\text{subject to:} \quad \mathbf{MQ} = \mathbf{F}, \mathbf{M} = \mathbf{G}, \mathbf{A} = \mathbf{H}, \qquad (3.42)$$

$$\mathbf{M} \in \mathcal{M}_1, \mathbf{A} \in \mathcal{M}_2, \qquad (3.43)$$

Note that here we do not enforce the positivity of $\mathbf{Q}$ explicitly because it is implicitly involved by $\mathbf{F} = \mathbf{MQ}$, $\mathbf{F} \geqslant 0$, and $\mathbf{M} \geqslant 0$ conditions. The corresponding augmented Lagrangian function is given as follows:

$$\begin{aligned} \mathcal{L}_3(\mathbf{M},\mathbf{Q},\mathbf{A},\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{\Gamma_1},\mathbf{\Gamma_2},\mathbf{\Gamma_3}) = &\frac{1}{2}\|\mathbf{Y}-\mathbf{FA}\|_F^2 + \frac{\rho_f}{2}\|\mathbf{MQ}-\mathbf{F}+\mathbf{\Gamma_1}\|_2^2 + \frac{\rho_g}{2}\|\mathbf{M}-\mathbf{G}+\mathbf{\Gamma_2}\|_2^2 \\ &+ \frac{\rho_h}{2}\|\mathbf{A}-\mathbf{H}+\mathbf{\Gamma_3}\|_2^2 + \iota_{\mathcal{S}}(\mathbf{G}) + \iota_{\mathcal{S}}(\mathbf{H}), \qquad (3.44) \end{aligned}$$

where $\mathbf{\Gamma_1} \in \mathbb{R}^{b\times k}$, $\mathbf{\Gamma_2} \in \mathbb{R}^{b\times k}$, $\mathbf{\Gamma_3} \in \mathbb{R}^{k\times n}$ are the Lagrangian dual variables associated with the linear equality constraints in Eq.(3.42), and ρ_f, ρ_g, and $\rho_h > 0$ are the corresponding penalty parameters. Finally, applying the ADMM scheme generates the following iterations:

$$\mathbf{M}^{t+1} := \arg\min_{\mathbf{M}\in\mathcal{M}_1} \frac{\rho_f}{2}\|\mathbf{MQ}-\mathbf{F}+\mathbf{\Gamma_1}\|_2^2 + \frac{\rho_g}{2}\|\mathbf{M}-\mathbf{G}+\mathbf{\Gamma}_2\|_2^2, \qquad (3.45)$$

$$\mathbf{A}^{t+1} := \arg\min_{\mathbf{A}\in\mathcal{M}_2} \frac{1}{2}\|\mathbf{Y}-\mathbf{FA}\|_F^2 + \frac{\rho_h}{2}\|\mathbf{A}-\mathbf{H}+\mathbf{\Gamma}_3\|_2^2, \qquad (3.46)$$

$$\mathbf{Q}^{t+1} := \arg\min_{\mathbf{Q}} \frac{\rho_f}{2}\|\mathbf{M}^{t+1}\mathbf{Q}-\mathbf{F}+\mathbf{\Gamma}_1\|_2^2, \qquad (3.47)$$

$$\mathbf{F}^{t+1} := \arg\min_{\mathbf{F}} \frac{1}{2}\|\mathbf{Y}-\mathbf{FA}^{t+1}\|_F^2 + \frac{\rho_f}{2}\|\mathbf{M}^{t+1}\mathbf{Q}^{t+1}+\mathbf{\Gamma}_1-\mathbf{F}\|_2^2, \qquad (3.48)$$

$$\mathbf{G}^{t+1} := \arg\min_{\mathbf{G}\geqslant 0} \iota_{\mathcal{S}}(\mathbf{G}) + \frac{\rho_g}{2}\|\mathbf{M}^{t+1} + \mathbf{\Gamma}_2 - \mathbf{G}\|_2^2, \tag{3.49}$$

$$\mathbf{H}^{t+1} := \arg\min_{\mathbf{H}\geqslant 0} \iota_{\mathcal{S}}(\mathbf{H}) + \frac{\rho_h}{2}\|\mathbf{A}^{t+1} + \mathbf{\Gamma}_3 - \mathbf{H}\|_2^2, \tag{3.50}$$

$$\mathbf{\Gamma}_1^{t+1} := \mathbf{\Gamma}_1 + \mathbf{M}^{t+1}\mathbf{Q}^{t+1} - \mathbf{F}^{t+1}, \tag{3.51}$$

$$\mathbf{\Gamma}_2^{t+1} := \mathbf{\Gamma}_2 + \mathbf{M}^{t+1} - \mathbf{G}^{t+1}, \tag{3.52}$$

$$\mathbf{\Gamma}_3^{t+1} := \mathbf{\Gamma}_3 + \mathbf{A}^{t+1} - \mathbf{H}^{t+1}. \tag{3.53}$$

The optimization of each subproblem is detailed in Section 2.3 of appendix I. The whole updating steps are summarized in Algorithm 3.4.

Algorithm 3.4 MA-ONTF model

```
**Input:** $\mathbf{Y}$, $k$, $\rho_f$, $\rho_g$, $\rho_h$, *Tol*, *niter*
**Initialization:** $\mathbf{M}^0$, $\mathbf{Q}^0$, $\mathbf{A}^0$, $\mathbf{F}^0 = \mathbf{M}^0\mathbf{Q}^0$, $\mathbf{G}^0 = \mathbf{M}^0$, $\mathbf{H}^0 = \mathbf{A}^0$, $\mathbf{\Gamma}_1 = 0$, $\mathbf{\Gamma}_2 = 0$, $\mathbf{\Gamma}_3 = 0$
**for** $i \leftarrow 1$ *niter* **do**
    Update $\mathbf{M}$ by solving (3.45) using Algorithm 3.2.
    Update $\mathbf{A}$ by solving (3.46) using Algorithm 3.2.
    $\mathbf{Q} \leftarrow (\mathbf{M}^T\mathbf{M})^{-1}(\mathbf{M}^T(\mathbf{F} - \mathbf{\Gamma}_1))$
    $\mathbf{F} \leftarrow (\mathbf{Y}\mathbf{A}^T + \rho_f(\mathbf{M}\mathbf{Q} + \mathbf{\Gamma}_1))(\mathbf{A}\mathbf{A}^T + \rho_f\mathbf{I})^{-1}$
    $\mathbf{G} \leftarrow \mathrm{prox}_{\iota_{\mathcal{S}}}(\mathbf{M} + \mathbf{\Gamma}_1)$
    $\mathbf{H} \leftarrow \mathrm{prox}_{\iota_{\mathcal{S}}}(\mathbf{A} + \mathbf{\Gamma}_2)$
    $\mathbf{\Gamma}_1 \leftarrow \mathbf{\Gamma}_1 + \mathbf{M}\mathbf{Q} - \mathbf{F}$
    $\mathbf{\Gamma}_2 \leftarrow \mathbf{\Gamma}_2 + \mathbf{M} - \mathbf{G}$
    $\mathbf{\Gamma}_3 \leftarrow \mathbf{\Gamma}_3 + \mathbf{A} - \mathbf{H}$
    **if** $(\|r_3[i] - r_3[i-1]\|/\|r_3[i-1]\| < Tol)$ **then**
        break.
    **end if**
**end for**
**return** $\mathbf{M}$, $\mathbf{Q}$, $\mathbf{A}$
```

3.4 Convergence analysis and computational complexity

In this section, we present the convergence analysis and the computational complexity of the proposed algorithms.

3.4.1 Computational complexity

We start by analyzing the complexity of Algorithm 3.2, which is involved in the other three algorithms and also is the most computationally expensive. We observe that Algorithm 3.2 has different complexities depending on the problem size. Thus, it has a complexity of $O(bkn)$ per iteration for the M-update step when called by Algorithm 3.1, $O(kn^2)$ per iteration for the **A**-update step when called by Algorithm 3.3, and $O(k^2b)$ and $O(bkn)$ per iteration for the **M**-update and **A**-update steps, respectively, when called by Algorithm 3.4. We suppose that Algorithm 3.2 stops after t_1 and t_2 iterations for the **M**-update and **A**-update steps, respectively. Therefore, the overall cost of Algorithm 3.2 in each case becomes $O(t_1bkn)$, $O(t_2kn^2)$, $O(t_2bk^2)$, and $O(t_1bkn)$, respectively.

The computational complexity of the proposed algorithms (M-ONMF, A-ONMF, and MA-ONTF) is summarized in Table 3.2. We supose that each algorithm converge after t iterations of the main loop in each algorithm.

Table 3.2 Computational time complexity of the proposed algorithms M-ONMF, A-ONMF, and MA-ONTF. Note that $k \ll b \ll n$

Step update	M-ONMF	A-ONMF	MA-ONTF
M update	$O(t_1bkn)$	$O(bkn)$	$O(t_1bk^2)$
A update	$O(bkn)$	$O(t_2kn^2)$	$O(t_2kn^2)$
Q update (Only for MAONTF)	-	-	$O(bk^2)$
1st Auxiliary update (**X**, **U**, and **F**)	$O(kn)$	$O(kn)$	$O(bkn)$
2nd Auxiliary update (**Z**, **V**, and **G**)	$O(bk)$	$O(bk)$	$O(bk)$
3rd Auxiliary update (**H**)	-	-	$O(kn)$
1st dual update ($\boldsymbol{\Lambda}_1$, $\boldsymbol{\Pi}_1$, and $\boldsymbol{\Gamma}_1$)	$O(bk)$	$O(bk)$	$O(bk^2)$
2nd dual update ($\boldsymbol{\Lambda}_2$, $\boldsymbol{\Pi}_2$, and $\boldsymbol{\Gamma}_2$)	$O(kn)$	$O(kn)$	$O(bk)$
3rd dual update ($\boldsymbol{\Gamma}_3$)	-	-	$O(kn)$
Overall	$O(tt_1bkn)$	$O(tt_2kn^2)$	$O(tt_2kn^2)$

From Table 3.2, we can observe that the M-ONMF is the less computationally expensive algorithm with $O(tt_1bkn)$. Updating the A variable is the most computationally expensive step in MA-ONTF and A-ONMF with $O(tt_2kn^2)$. The numerical results illustrated in Fig. 3.3 in the next section gives an idea about the time cost of each algorithm as a function of the number of pixels and the number of bands as well.

3.4.2 Convergence analysis

Due to the space limit, only the convergence of Algorithm 3.1 is detailed here. The same steps are used to analyze the convergence of Algorithm 3.3 and Algorithm 3.4. To facilitate our analysis, we start the analysis by making the following common assumptions.

Assumption 1. *The gradient $\nabla\mathcal{L}_1$ of the Lagrangian function $\mathcal{L}_1$ w.r.t $\boldsymbol{M}$ and $\boldsymbol{A}$, respectively, is Lipschitz continuous, which means*

$$||\nabla\mathcal{L}_1(\boldsymbol{M}_1) - \nabla\mathcal{L}_1(\boldsymbol{M}_2)|| \leq L_m||\boldsymbol{M}_1 - \boldsymbol{M}_2||, \quad \forall\, \boldsymbol{M}_1, \boldsymbol{M}_2, \tag{3.54}$$

and

$$||\nabla\mathcal{L}_1(\boldsymbol{A}_1) - \nabla\mathcal{L}_1(\boldsymbol{A}_2)|| \leq L_a||\boldsymbol{A}_1 - \boldsymbol{A}_2||, \quad \forall\, \boldsymbol{A}_1, \boldsymbol{A}_2, \tag{3.55}$$

where L_m, $L_a \geq 0$ are two Lipschitz constants.

Assumption 2. *The proximal mapping function $prox_i$ is non-expansive, which means*

$$||prox_{\iota_S}(\boldsymbol{X}_1) - prox_{\iota_S}(\boldsymbol{X}_2)|| \leq ||\boldsymbol{X}_1 - \boldsymbol{X}_2||, \quad \forall\, \boldsymbol{X}_1, \boldsymbol{X}_2, \tag{3.56}$$

The Karush-Kuhn-Tucker (KKT) (Bertsekas, 1997; Boyd *et al.*, 2011) conditions associated with problem (3.21) are:

$$-\mathbf{M}^T(\mathbf{Y} - \mathbf{MA}) + \mathbf{\Lambda}_2 = 0, \tag{3.57}$$

$$\mathbf{M} - \mathbf{X} = 0, \tag{3.58}$$

$$\mathbf{A} - \mathbf{Z} = 0, \tag{3.59}$$

$$\mathbf{\Lambda}_1 + \mathbf{M} \in N_{\iota_{\mathcal{S}}}(\mathbf{X}) \tag{3.60}$$

$$\mathbf{\Lambda}_2 + \mathbf{A} \in N_{\iota_{\mathcal{S}}}(\mathbf{Z}) \tag{3.61}$$

where $N_{\iota_{\mathcal{S}}}(\mathbf{X})$, called the normal cone, is the sub-differential of the indicator function, which is defined as $N_{\iota_{\mathcal{S}}}(\mathbf{X}) = \partial \iota_{\mathcal{S}}(\mathbf{X})$ at a point $\mathbf{X} \in C$.

To simplify the notation, let us define a point $\mathbf{W} \triangleq (\mathbf{A}, \mathbf{M}, \mathbf{X}, \mathbf{Z}, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$. Let $\{\mathbf{W}^t\}$ be a sequence generated by the update formulas (15a-15f). A point $\mathbf{W}$ is called a KKT point if it satisfies the above KKT conditions.

Theorem 1. *The sequence $\{\boldsymbol{W}^t\}$ is bounded, and every limit point of it is a KKT point of problem (3.21).*

Proof. The proof consists of the following.

Proposition 2. *Let $\{M^{t+1}\} \subset R^{m \times k}$ be a sequence such that*

$$||\boldsymbol{M}^{t+1} - \boldsymbol{M}^*|| \leq ||\boldsymbol{M}^t - \boldsymbol{M}^*|| + \epsilon_t \tag{3.62}$$

Then $\{M^t\}$ is bounded.

Proof. We have

$$\begin{aligned} ||\mathbf{M}^{t+1} - \mathbf{M}^*|| &= ||R_X(\mathbf{M}) - \mathbf{M}^*|| \\ &= ||R_X(\mathbf{M}) - \mathbf{M}^* + \mathbf{M}^t - \mathbf{M}^t|| \\ &= ||(R_X(\mathbf{M}) - \mathbf{M}^t) + (\mathbf{M}^t - \mathbf{M}^*)|| \\ &\leq ||(\mathbf{M}^t - \mathbf{M}^*)|| + ||(R_X(\mathbf{M}) - \mathbf{M}^t)|| \end{aligned} \tag{3.63}$$

Moreover, considering the Lipschitz Assumption 1 in (3.54) and the boundedness of the Stiefel manifold $\mathcal{M}_1$, it follows that both the sub-gradient $\nabla \mathcal{L}_1(\mathbf{M})$ and its Stiefel manifold gradient grad$\mathcal{L}_1$ are bounded, which can be written as:

$$\|\text{grad}\mathcal{L}_1\| \leq \|\nabla\mathcal{L}_1(\mathbf{M})\| \leq c, \quad \forall\, \mathbf{M} \in \mathcal{L}_1, \tag{3.64}$$

where $c > 0$ is a constant. Therefore, we can conclude that the sequence $\{\mathbf{M}^t\}$ is bounded, i.e., $\lim\limits_{t\to+\infty} \|\mathbf{M}^{t+1} - \mathbf{M}^t\|_F^2 = 0$. Q.E.D.

For the subproblem (3.26), let $\mathbf{X}^*$ be its optimal point, which implies that (see (Rockafellar, 1970, Theorem 23.8) and the proof therein)

$$0 \in \partial\iota_{\mathcal{S}}(\mathbf{X}^*) + \partial(\frac{\rho_x}{2}\|\mathbf{X}^* - (\mathbf{M} + \mathbf{\Lambda}_1)\|_2^2). \tag{3.65}$$

We can demonstrate the boundedness of the Lagrangian variable $\mathbf{\Lambda}_1$ as follows. According to (Eckstein & Yao, 2015, Lamma 4), there exist $\mathbf{B} \in \partial\iota_{\mathcal{S}}(\mathbf{X}^*)$ such that

$$0 = \mathbf{B} + \rho_x(\mathbf{X}^* - (\mathbf{M} + \mathbf{\Lambda}_1). \tag{3.66}$$

Which implies that

$$\mathbf{\Lambda}_1 = \mathbf{X}^* + \frac{1}{\rho_x}\mathbf{B} - \mathbf{M} \tag{3.67}$$

The difference of two successive updates of $\mathbf{\Lambda}_1$ can be written as

$$\begin{aligned} \|\mathbf{\Lambda}_1^{t+1} - \mathbf{\Lambda}_1^t\| &= \|(\mathbf{X}^{t+1} + \frac{\mathbf{B}^{t+1}}{\rho_x} - \mathbf{M}^{t+1}) - (\mathbf{X}^t + \frac{\mathbf{B}^t}{\rho_x} - \mathbf{M}^t)\| \\ &\leq \|\mathbf{X}^{t+1} - \mathbf{X}^t\| + \frac{1}{\rho_x}\|\mathbf{B}^{t+1} - \mathbf{B}^t\| - \|\mathbf{M}^{t+1} - \mathbf{M}^t\| \end{aligned} \tag{3.68}$$

Moreover, Assumption 2 invokes that

$$\begin{aligned} &\|prox_{\iota_{\mathcal{S}}}(\mathbf{\Lambda}_1^{t+1} + \mathbf{M}^{t+1}) - prox_{\iota_{\mathcal{S}}}(\mathbf{\Lambda}_1^t + \mathbf{M}^t)\| \\ &\leq \|\mathbf{\Lambda}_1^{t+1} + \mathbf{M}^{t+1} - \mathbf{\Lambda}_1^t + \mathbf{M}^t\| \\ &\leq \|\mathbf{\Lambda}_1^{t+1} - \mathbf{\Lambda}_1^t\| + \|\mathbf{M}^{t+1} - \mathbf{M}^t\| \end{aligned} \tag{3.69}$$

Considering the boundedness of $\{\mathbf{M}^t\}$, we can conclude that $\mathbf{\Lambda}_1^t$ is bounded, which implies $\lim_{t\to+\infty} \|\mathbf{\Lambda}_1^{t+1} - \mathbf{\Lambda}_1^t\| = 0$, and implies also

$$\lim_{t\to+\infty} \|\mathbf{X}^{t+1} - \mathbf{X}^t\| = 0,$$

Using the same steps above, we can show that $\mathbf{\Lambda}_2^t$ is bounded (the details are omitted due to the space limit), which means

$$\lim_{t\to+\infty} \|\mathbf{\Lambda}_2^{t+1} - \mathbf{\Lambda}_2^t\| = 0.$$

Finally, by rearranging the formulas of the optimality conditions in (3.57-3.61) we get

$$\begin{aligned}(\mathbf{M}^T\mathbf{M} + \rho_Z\mathbf{I})(\mathbf{A}^{t+1} - \mathbf{A}^t) &= \mathbf{M}^T(\mathbf{Y} - \mathbf{M}\mathbf{A}^t) \\ &\quad - \mathbf{\Lambda}_2^t + \rho_Z(\mathbf{A}^t - \mathbf{Z}^t), & (3.70) \\ \mathbf{X}^{t+1} - \mathbf{X}^t &= \mathrm{prox}_{\iota_S}(\mathbf{M}^{t+1} + \mathbf{\Lambda}_1^t) - \mathbf{X}^t, & (3.71) \\ \mathbf{Z}^{t+1} - \mathbf{Z}^t &= \mathrm{prox}_{\iota_S}(\mathbf{A}^{t+1} + \mathbf{\Lambda}_2^t) - \mathbf{Z}^t, & (3.72) \\ \mathbf{\Lambda}_1^{t+1} - \mathbf{\Lambda}_1^t &= \mathbf{M}^{t+1} - \mathbf{X}^{t+1}, & (3.73) \\ \mathbf{\Lambda}_2^{t+1} - \mathbf{\Lambda}_2^t &= \mathbf{A}^{t+1} - \mathbf{Z}^{t+1}. & (3.74)\end{aligned}$$

The assumption $\lim_{t\to+\infty} \|\mathbf{W}^{t+1} - \mathbf{W}^t\| \to 0$ implies that both sides of the above formulas go to zero.

The boundedness of $\mathbf{\Lambda}_1^t$ and $\mathbf{\Lambda}_2^t$, respectively, implies

$$\mathbf{M}^{t+1} - \mathbf{X}^{t+1} = 0, \quad (3.75)$$

$$\mathbf{A}^{t+1} - \mathbf{Z}^{t+1} = 0, \quad (3.76)$$

which represent the 2^{nd} and the 3^{rd} KKT conditions, respectively.

According to Assumption 1 and due to the assumption $\lim_{t \to +\infty} \|\mathbf{A}^{t+1} - \mathbf{A}^t\| \to 0$, we must have

$$(\mathbf{M}^{t+1})^T(\mathbf{Y} - \mathbf{M}^{t+1}\mathbf{A}^{t+1}) - \mathbf{\Lambda}_2^{t+1} = 0, \tag{3.77}$$

which is the 1^{st} KKT condition.

For (3.71) and (3.72), it easy to show that invoking the boundedness of $\mathbf{X}^t$ and $\mathbf{Z}^t$, respectively, yields

$$\mathrm{prox}_{\iota_S}(\mathbf{M}^{t+1} + \mathbf{\Lambda}_1^t) - \mathbf{X}^{t+1} = 0, \tag{3.78}$$

$$\mathrm{prox}_{\iota_S}(\mathbf{A}^{t+1} + \mathbf{\Lambda}_2^t) - \mathbf{Z}^{t+1} = 0. \tag{3.79}$$

These results are the KKT conditions of problem (3.21). This completes the proof. Q.E.D.

Similarly, we can show that both A-ONMF and MA-ONTF convergence to their corresponding KKT points. In practice, the graphs of Fig.3.4 shows that the proposed algorithms empirically converge after a few tens of iterations.

3.5 Experimental results and analysis

In this section, we present the results of the different experiments we conducted to evaluate the performance of the proposed models. Several quantitative and qualitative tests were performed on different collections of MS and HS images. Furthermore, two scenarios of experiments were considered in this study to show the possible use cases of our approaches, namely, blind source separation and text extraction (binarization).

3.5.1 Datasets

To assess the effectiveness of our models, we considered three different datasets. While the focus of this study was on the decomposition of MS document images, mainly due to the availability of a well know MS dataset, we validated the proposed approaches on the HS dataset as well.

Thus, in this study, we consider one synthetic HS dataset and two collections of MS images of real old documents. The description of each dataset is given as follows:

3.5.1.1 HS Hubble Telescope dataset

A set of 100 simulated spectral images of the Hubble telescope itself, 128 x 128 pixels each. The Hubble telescope is composed of eight materials, namely (Zhang, Wang, Plemmons & Pauca, 2008): 1) Bolts (BO), 2) Copper Stripping (CS), 3) HPCA doesn't seek to separate the sources into their original format, rather it presents them as the difference between two extremes. Blind source separation seeks to go further than PCA and return factors that are more comparable to the the original signals by exploiting properties of the original signal type, e.g. having no negatives weights in the factors.oneycomb Side (HS), 4) Honeycomb Top (HT), 5) Aluminum (AL), 6) Green Glue (GG), 7) Solar Cell (SC), and 8) Black Rubber Edge (RE). The dataset was made by mixing (varying the compositions) the true spectral signatures ($4\mu m$ to $2.5\mu m$) of the eight materials provided by the NASA Johnson Space Center (Pauca, Piper & Plemmons, 2006).

3.5.1.2 MSTEx-1

A collection of 21 MS data-cubes of real ancient documents, provided by the BAnQ (Bibliothèque et Archives nationales du Québec), digitized and published by Synchromedia research group[11] (Hedjam *et al.*, 2015). The MS images were captured using 8 different spectral bands, one in the ultra-violet (UV) spectrum ($340nm$), three bands in the visible spectrum (Blue ($500nm$), Green ($600nm$), and Red ($700nm$)), and four bands in the infrared (IR) spectrum ($800nm$, $900nm$, $1000nm$, $1100nm$). The MS camera's sensor used is a Chroma X3 KAF 6303E (Kodak), with a resolution of 6 megapixels ($9 \times 9\mu m$). More technical details about the digitization and calibration protocols can be found in (Hedjam, 2013).

[11] Available at http://www.synchromedia.ca/databases/MSI-HISTODOC

3.5.1.3 MSTEx-2

A second collection of 10 MS data-cubes of real ancient documents, collected by the same research group[0] and digitized using the same MS imaging system.

We notice that the MSTEx-1 and MSTEx-2 datasets are partially labeled. Each ground-truth (GT) image is a binary image, with each pixel is assigned to one of the two classes, text or non-text. In these GT images, black pixels represent the foreground (text), and white pixels correspond to the background (which covers annotations, degradation, background, and stamps). The GT (binary) images were generated by a two-steps protocol. First, for each MS cube, images from different bands are processed to separate the foreground from the background using a phase-based method developed in (Nafchi *et al.*, 2014) and combined to produce a rough binary image. The combined image is then manually post-processed to obtain the final binary output. The process of GT generation was the same for all MS cubes.

3.5.2 Evaluation metrics

For the blind decomposition task, we adopt Pearson's correlation coefficient (PCC) as a metric (Boslaugh & Watters, 2008) to evaluate the performance of the proposed methods. The value of PCC ranges between -1 to +1, where a perfect match between two images (components) will be indicated by +1 and a perfect mismatch by -1. As for the text extraction task, we evaluate the efficiency of our algorithms compared to other benchmarks in terms of four widely used quantitative metrics: the F1-score (Sokolova & Lapalme, 2009), also called F-measure (FM), the Distance Reciprocal Distortion (DRD) (Lu, Kot & Shi, 2004), the Negative Rate Metric (NRM) (Young & Ferryman, 2005), and the accuracy metric (ACC)[12]. The definition of each metric is given in Section 3 of Appendix I. The F-Measure (FM) is expressed as a percentage, whereas the values of DRD and the NRM range from 0 to 1. A better binarization quality has

[12] DRD and NRM are typically used to assess text binarization results, while FM and ACC are used to evaluate general classification and clustering tasks.

higher F-Measure and ACC values. In contrast, the lowest is the value of NRM or DRD the best is the quality of the binarization.

3.5.3 Parameters setup

For the sake of reproducibility, we provide here the different settings we used in our experiments. The initialization of the matrix factors of our models, i.e., **M**, **A**, and **Q**, is done using a deterministic initialization based on SVD-decomposition. In addition, for each model, every decomposition experiment was run only one single time, because the orthogonality constraint leads to a unique decomposition. The penalty parameters of our models are tuned using a grid search strategy, with the range $\{10^{-5}, \ldots, 10^{5}\}$ for each parameter. Due to the variability of the spectral signature of each sample, we tuned the corresponding hyperparameters for each model. The $thresh$ value used in in the text separation (binarization) task was set manually for all samples (see Section. 3.5.7). The best settings of the different hyperparameters used in our experiments are provided in Section 5 of Appendix I. As a stopping criteria, we set the maximum number of iterations to $niter = 100$, and the tolerance $tol = 10^{-5}$ for the three algorithms. For the number of sources (rank) selection, in this study, we did not investigate the automatic rank selection issue; therefore, we set it manually.

3.5.4 Computational analysis

Fig. 3.3 shows the average processing time required by each algorithm as a function of (a) the size of the image (in terms of pixels number) and (b) the number of bands. Experiments were done on a machine with Ubuntu 16.04 OS, CPU i7 with 3.6GHz, and 16GB of RAM.

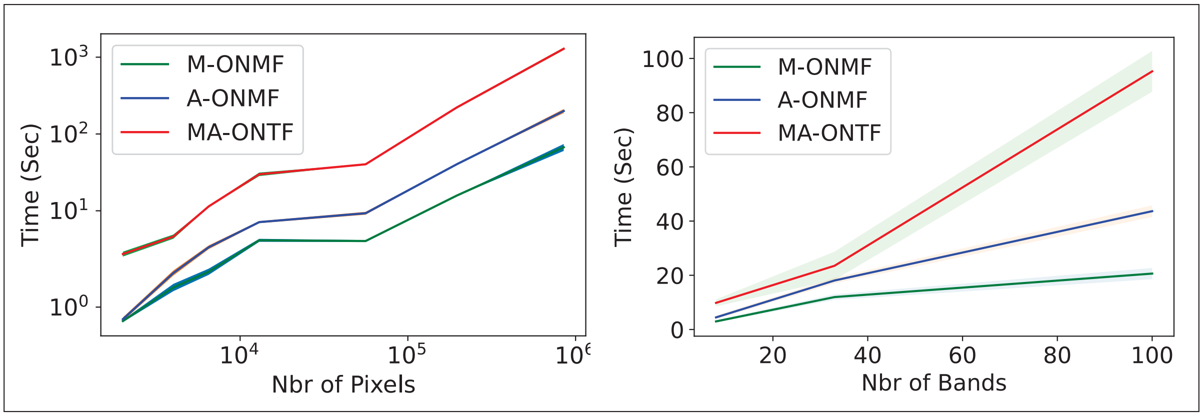
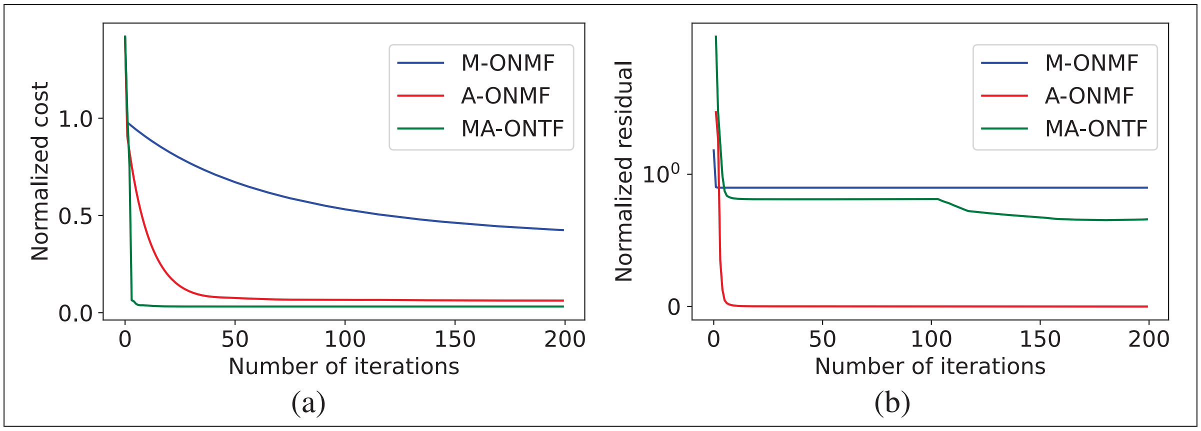


Figure 3.3 Average running time of the proposed algorithms. In the left, the running time as a function of the number of pixels with a constant number of bands (8), and in the right, as a function of the number of bands with a constant number of pixels (128 × 128)

As it can be observed in Fig. 3.3, it is not surprising to see that the MA-ONTF method is the most computationally expensive among the three models due to the double optimization over the Stiefel manifold. M-ONMF is the less expensive model because the optimization over the Stiefel manifold is applied on the **M** matrix, which is generally of a small size.

Next, Fig. 3.4 illustrates the convergence performance of the proposed models in terms of the normalized cost and the normalized residual given by $\|\mathbf{Y}-\mathbf{MA}\|_F/\|\mathbf{Y}\|_F$ and $\|r_i^t-r_i^{t-1}\|_F/\|r_i^{t-1}\|_F$ respectively.

In Fig. 3.4, we set the constant $Tol = 10^{-5}$, which is a very small precision number, as the smallest objective value / residual error. We also set $niter = 200$ in order to show that the three algorithms are monotonously decreasing. In general the three algorithms converge in few tens of iterations. In Fig. 3.4-a, the MA-ONTF model achieves the lowest minimum of the cost function, i.e., a lowest reconstruction error in less than 50 iterations. A-ONMF stabilizes between 50 and 100 iterations and performs better than M-ONMF in terms of both measures. It also achieves the lowest precision in terms of the residual error. Both, A-ONMF and MA-ONTF models converge slightly faster than M-ONMF.



(a) (b)

Figure 3.4 Convergence properties of the proposed models. In (a), the convergence rate in terms of the normalized reconstruction error, and in (b) in terms of the normalized residual error

3.5.5 Blind decomposition of the Hubble Telescope dataset

The first experiment was carried out on the Hubble Telescope dataset in order to show how our models perform in HS image decomposition tasks. We compare the performance of our models against eight methods: 4 existing orthogonal NMF methods (whose code is available), namely ONMTF[13] (Orthogonal Nonnegative Matrix Tri-Factorization) (Yoo & Choi, 2010), OPNMF[14] (Orthogonal Projective NMF) (Yang & Oja, 2010), EM-ONMF[15] (Expectation-Maximization Orthogonal NMF) (Pompili *et al.*, 2014), ON-PMF[0] (Orthogonal Nonnegatively Penalized Matrix Factorization) (Pompili *et al.*, 2014), the standard NMF, two clustering techniques, k-means and SKKHM[16] (Spatial Kernel K-Harmonic Means) (Li *et al.*, 2007), and a BSS approach called the Reconstruction Independent Components Analysis (RICA)[17] (Le, Karpenko, Ngiam & Ng, 2011). For all methods, the number of materials was set manually to $k = 8$. It should be noticed that we ran RICA, k-means, SKKHM, NMF, ONMTF, OPNMF, and EM-ONMF several times and we selected the best results obtained. For RICA, another clustering

[13] In NMFlib https://www.ee.columbia.edu/~grindlay/code.html

[14] An implementation is provided at https://github.com/asotiras/brainparts

[15] Code and Hubble dataset are available at https://github.com/filippo-p/onmf

[16] Code is available at http://utopia.duth.gr/~nmitiano/

[17] We used the build-in function of Matlab.

step was required to perform the separation, for which we used k-means (it also works with any other clustering approach). For the remaining methods, as a positive effect of the deterministic initialization and the orthogonality constraint, only one run is needed.

Fig.3.5 shows the outcomes of each method. In the case of NMF decomposition methods, the rows of the abundance (coefficient) matrix were reshaped to the size of input images after normalization. It is obvious to see that among all methods, only the three models A-ONMF, MA-ONTF, and ONP-MF, were able to successfully isolate and cluster the pixels of the eight materials without any overlapping. The M-ONMF model was able to separate six elements correctly but failed to cluster two of them. EM-ONMF performs slightly worse than M-ONMF, with only four materials correctly identified. It is also surprising to observe that the standard NMF outperforms OPNMF and ONMTF in this task. Both methods have the lowest performance and give unexpected results. RICA, SKKHM and k-means performed better than OPNMF and ONMTF also. The OPNMF and ONMTF do not seem to achieve good orthogonality. For that reason, their results are not good enough compared to the standard NMF. Indeed, that what motivated us to consider the optimization over the Stiefel manifold instead of adding orthogonality as a regularization term to the objective function. The qualitative illustration confirms, with the exception of certain models (OPNMF and ONMTF), orthogonal methods (5 over 7 methods), especially those based on the Stiefel manifold, tend to be more the effective over the standard NMF model and other clustering approaches in the task of HS image decomposition.

We also notice that even if the number of pixels of each material is highly imbalanced, the A-ONMF, MA-ONTF, and ONP-MF methods clustered the pixels of small parts properly. For instance, the Bolts elements (Fig.3.5-h) have the smallest number of pixels compared to other elements. Such an issue may cause a real challenge for supervised classification methods that rely on training samples.

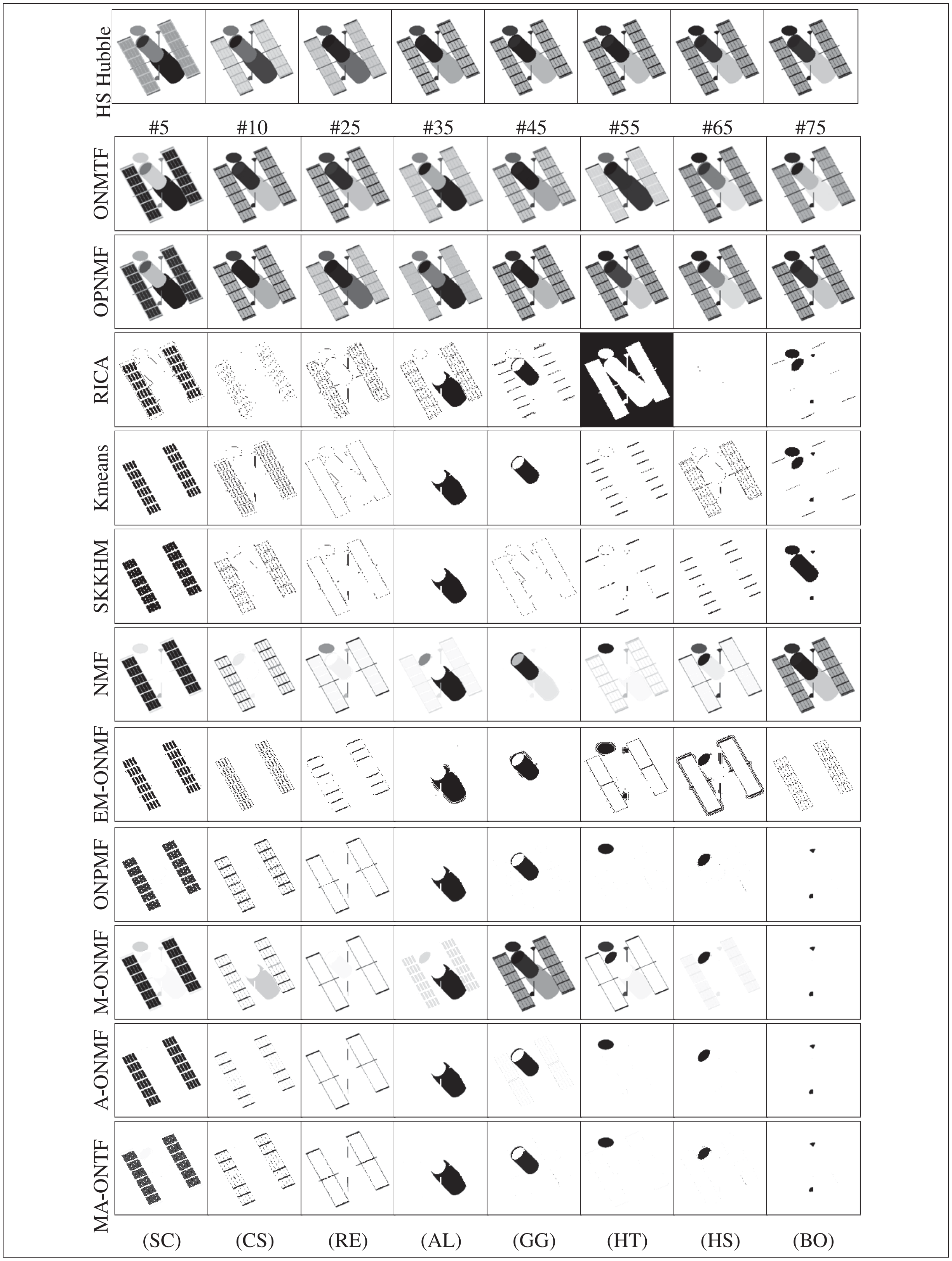
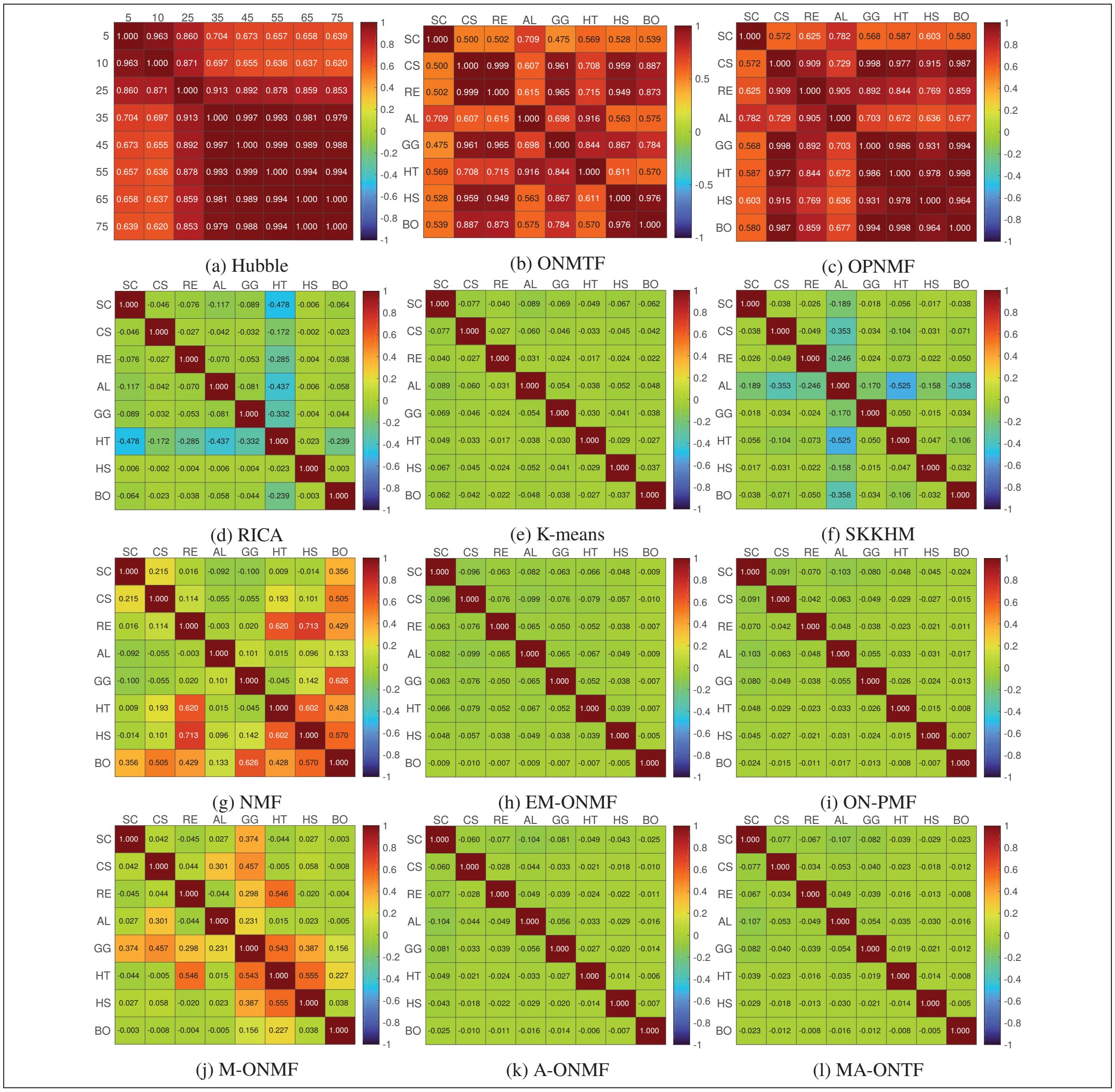


Figure 3.5 Results of the several approaches in the decomposition of the Hubble dataset into its 8 components. From top to bottom, the 1^{st} row shows raw HS images with their corresponding bands, the remaining rows show the result of each method

Likewise, although it is not hard to visually assess the quality of the decomposition, we calculated the correlation (PCC) between all pair of images (columns a-h) for each method to quantify decomposition quality numerically. The resulting correlation matrices are shown in Fig.3.6.



	5	10	25	35	45	55	65	75
5	1.000	0.963	0.860	0.704	0.673	0.657	0.658	0.639
10	0.963	1.000	0.871	0.697	0.655	0.636	0.637	0.620
25	0.860	0.871	1.000	0.913	0.892	0.878	0.859	0.853
35	0.704	0.697	0.913	1.000	0.997	0.993	0.981	0.979
45	0.673	0.655	0.892	0.997	1.000	0.999	0.989	0.988
55	0.657	0.636	0.878	0.993	0.999	1.000	0.994	0.994
65	0.658	0.637	0.859	0.981	0.989	0.994	1.000	1.000
75	0.639	0.620	0.853	0.979	0.988	0.994	1.000	1.000

(a) Hubble

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	0.500	0.502	0.709	0.475	0.569	0.528	0.539
CS	0.500	1.000	0.999	0.607	0.961	0.708	0.959	0.887
RE	0.502	0.999	1.000	0.615	0.965	0.715	0.949	0.873
AL	0.709	0.607	0.615	1.000	0.698	0.916	0.563	0.575
GG	0.475	0.961	0.965	0.698	1.000	0.844	0.867	0.784
HT	0.569	0.708	0.715	0.916	0.844	1.000	0.611	0.570
HS	0.528	0.959	0.949	0.563	0.867	0.611	1.000	0.976
BO	0.539	0.887	0.873	0.575	0.784	0.570	0.976	1.000

(b) ONMTF

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	0.572	0.625	0.782	0.568	0.587	0.603	0.580
CS	0.572	1.000	0.909	0.729	0.998	0.977	0.915	0.987
RE	0.625	0.909	1.000	0.905	0.892	0.844	0.769	0.859
AL	0.782	0.729	0.905	1.000	0.703	0.672	0.636	0.677
GG	0.568	0.998	0.892	0.703	1.000	0.986	0.931	0.994
HT	0.587	0.977	0.844	0.672	0.986	1.000	0.978	0.998
HS	0.603	0.915	0.769	0.636	0.931	0.978	1.000	0.964
BO	0.580	0.987	0.859	0.677	0.994	0.998	0.964	1.000

(c) OPNMF

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	-0.046	-0.076	-0.117	-0.089	-0.478	-0.006	-0.064
CS	-0.046	1.000	-0.027	-0.042	-0.032	-0.172	-0.002	-0.023
RE	-0.076	-0.027	1.000	-0.070	-0.053	-0.285	-0.004	-0.038
AL	-0.117	-0.042	-0.070	1.000	-0.081	-0.437	-0.006	-0.058
GG	-0.089	-0.032	-0.053	-0.081	1.000	-0.332	-0.004	-0.044
HT	-0.478	-0.172	-0.285	-0.437	-0.332	1.000	-0.023	-0.239
HS	-0.006	-0.002	-0.004	-0.006	-0.004	-0.023	1.000	-0.003
BO	-0.064	-0.023	-0.038	-0.058	-0.044	-0.239	-0.003	1.000

(d) RICA

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	-0.077	-0.040	-0.089	-0.069	-0.049	-0.067	-0.062
CS	-0.077	1.000	-0.027	-0.060	-0.046	-0.033	-0.045	-0.042
RE	-0.040	-0.027	1.000	-0.031	-0.024	-0.017	-0.024	-0.022
AL	-0.089	-0.060	-0.031	1.000	-0.054	-0.038	-0.052	-0.048
GG	-0.069	-0.046	-0.024	-0.054	1.000	-0.030	-0.041	-0.038
HT	-0.049	-0.033	-0.017	-0.038	-0.030	1.000	-0.029	-0.027
HS	-0.067	-0.045	-0.024	-0.052	-0.041	-0.029	1.000	-0.037
BO	-0.062	-0.042	-0.022	-0.048	-0.038	-0.027	-0.037	1.000

(e) K-means

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	-0.038	-0.026	-0.189	-0.018	-0.056	-0.017	-0.038
CS	-0.038	1.000	-0.049	-0.353	-0.034	-0.104	-0.031	-0.071
RE	-0.026	-0.049	1.000	-0.246	-0.024	-0.073	-0.022	-0.050
AL	-0.189	-0.353	-0.246	1.000	-0.170	-0.525	-0.158	-0.358
GG	-0.018	-0.034	-0.024	-0.170	1.000	-0.050	-0.015	-0.034
HT	-0.056	-0.104	-0.073	-0.525	-0.050	1.000	-0.047	-0.106
HS	-0.017	-0.031	-0.022	-0.158	-0.015	-0.047	1.000	-0.032
BO	-0.038	-0.071	-0.050	-0.358	-0.034	-0.106	-0.032	1.000

(f) SKKHM

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	0.215	0.016	-0.092	-0.100	0.009	-0.014	0.356
CS	0.215	1.000	0.114	-0.055	-0.055	0.193	0.101	0.505
RE	0.016	0.114	1.000	-0.003	0.020	0.620	0.713	0.429
AL	-0.092	-0.055	-0.003	1.000	0.101	0.015	0.096	0.133
GG	-0.100	-0.055	0.020	0.101	1.000	-0.045	0.142	0.626
HT	0.009	0.193	0.620	0.015	-0.045	1.000	0.602	0.428
HS	-0.014	0.101	0.713	0.096	0.142	0.602	1.000	0.570
BO	0.356	0.505	0.429	0.133	0.626	0.428	0.570	1.000

(g) NMF

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	-0.096	-0.063	-0.082	-0.063	-0.066	-0.048	-0.009
CS	-0.096	1.000	-0.076	-0.099	-0.076	-0.079	-0.057	-0.010
RE	-0.063	-0.076	1.000	-0.065	-0.050	-0.052	-0.038	-0.007
AL	-0.082	-0.099	-0.065	1.000	-0.065	-0.067	-0.049	-0.009
GG	-0.063	-0.076	-0.050	-0.065	1.000	-0.052	-0.038	-0.007
HT	-0.066	-0.079	-0.052	-0.067	-0.052	1.000	-0.039	-0.007
HS	-0.048	-0.057	-0.038	-0.049	-0.038	-0.039	1.000	-0.005
BO	-0.009	-0.010	-0.007	-0.009	-0.007	-0.007	-0.005	1.000

(h) EM-ONMF

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	-0.091	-0.070	-0.103	-0.080	-0.048	-0.045	-0.024
CS	-0.091	1.000	-0.042	-0.063	-0.049	-0.029	-0.027	-0.015
RE	-0.070	-0.042	1.000	-0.048	-0.038	-0.023	-0.021	-0.011
AL	-0.103	-0.063	-0.048	1.000	-0.055	-0.033	-0.031	-0.017
GG	-0.080	-0.049	-0.038	-0.055	1.000	-0.026	-0.024	-0.013
HT	-0.048	-0.029	-0.023	-0.033	-0.026	1.000	-0.015	-0.008
HS	-0.045	-0.027	-0.021	-0.031	-0.024	-0.015	1.000	-0.007
BO	-0.024	-0.015	-0.011	-0.017	-0.013	-0.008	-0.007	1.000

(i) ON-PMF

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	0.042	-0.045	0.027	0.374	-0.044	0.027	-0.003
CS	0.042	1.000	0.044	0.301	0.457	-0.005	0.058	-0.008
RE	-0.045	0.044	1.000	-0.044	0.298	0.546	-0.020	-0.004
AL	0.027	0.301	-0.044	1.000	0.231	0.015	0.023	-0.005
GG	0.374	0.457	0.298	0.231	1.000	0.543	0.387	0.156
HT	-0.044	-0.005	0.546	0.015	0.543	1.000	0.555	0.227
HS	0.027	0.058	-0.020	0.023	0.387	0.555	1.000	0.038
BO	-0.003	-0.008	-0.004	-0.005	0.156	0.227	0.038	1.000

(j) M-ONMF

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	-0.060	-0.077	-0.104	-0.081	-0.049	-0.043	-0.025
CS	-0.060	1.000	-0.028	-0.044	-0.033	-0.021	-0.018	-0.010
RE	-0.077	-0.028	1.000	-0.049	-0.039	-0.024	-0.022	-0.011
AL	-0.104	-0.044	-0.049	1.000	-0.056	-0.033	-0.029	-0.016
GG	-0.081	-0.033	-0.039	-0.056	1.000	-0.027	-0.020	-0.014
HT	-0.049	-0.021	-0.024	-0.033	-0.027	1.000	-0.014	-0.006
HS	-0.043	-0.018	-0.022	-0.029	-0.020	-0.014	1.000	-0.007
BO	-0.025	-0.010	-0.011	-0.016	-0.014	-0.006	-0.007	1.000

(k) A-ONMF

	SC	CS	RE	AL	GG	HT	HS	BO
SC	1.000	-0.077	-0.067	-0.107	-0.082	-0.039	-0.029	-0.023
CS	-0.077	1.000	-0.034	-0.053	-0.040	-0.023	-0.018	-0.012
RE	-0.067	-0.034	1.000	-0.049	-0.039	-0.016	-0.013	-0.008
AL	-0.107	-0.053	-0.049	1.000	-0.054	-0.035	-0.030	-0.016
GG	-0.082	-0.040	-0.039	-0.054	1.000	-0.019	-0.021	-0.012
HT	-0.039	-0.023	-0.016	-0.035	-0.019	1.000	-0.014	-0.008
HS	-0.029	-0.018	-0.013	-0.030	-0.021	-0.014	1.000	-0.005
BO	-0.023	-0.012	-0.008	-0.016	-0.012	-0.008	-0.005	1.000

(l) MA-ONTF

Figure 3.6 Heat maps representation of correlation matrices. In (a) the PCC is calculated between the bands 5, 15, 25, 35, 45, 55, 65, et 75. In (b-l), for each method the PCC is calculated between pair of output images obtained by the decomposition of the Hubble dataset (results shown in Fig.3.5)

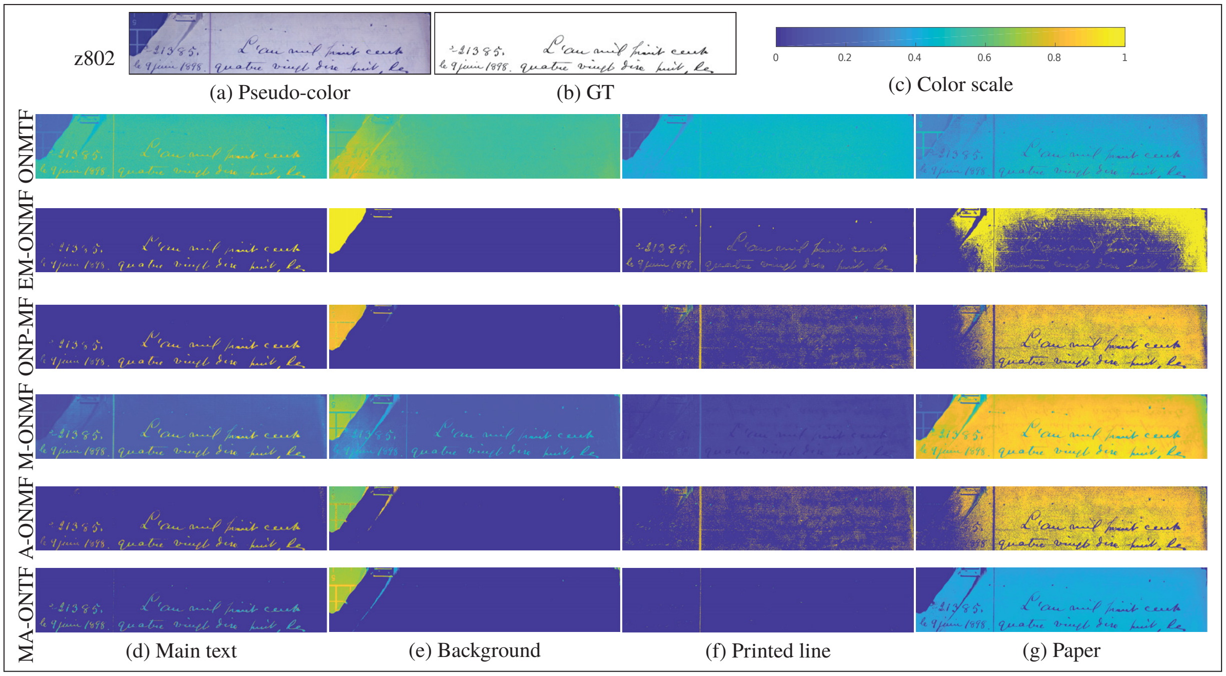
As expected, perfect diagonal matrices are obtained for ONP-MF, A-ONMF, and MA-ONTF. The first two matrices (a) and (b) reflect the failure of the ONMTF and OPNMF. However, for the hard clustering methods (K-means, SKKHM, and EM-ONMF) and RICA, we also obtained diagonal matrices, even the materials are not well separated from each other. This is due to the fact that the corresponding images contain non-overlapping pixels, which does not imply that pixels were correctly clustered. Based on this observation, we can say that perfect decomposition yields a diagonal correlation matrix, but the inverse is not always true. This is an important point to be considered when using the PCC metric.

3.5.6 Blind decomposition of MS document images

The second set of experiments was conducted on MSTEx-1 and MSTEx-2 datasets, two collections of real ancient documents. For the MSTEx-2 dataset, we used the whole set of MS cubes, and we randomly selected a set of 11 out of 21 MS cubes from the MSTEx-1 dataset. The decomposition of MS document images is conducted using the following algorithms: ONMTF, EM-ONMF, ONP-MF, M-ONMF, A-ONMF, MA-ONTF.

For visual comparison purpose, the abundance maps corresponding to different materials (endmembers) obtained by different methods are shown in Fig.3.7 and Fig. I-1, where Fig.3.7 shows the results of a sample (z802) taken from MSTEx-1, and Fig. I-1 shows another sample (z35) taken from the MSTEx-2 collection. In these two illustrations, the images in each column are reshaped form the corresponding abundance matrix and then normalized. As indicated by the color scale, a high intensity indicates the presence of materials and a low intensity indicates the opposite.

By taking a look at Fig.3.7, it can be clearly seen that among all the considered methods, only the MA-ONTF model is able to achieve the best decomposition. All the materials are well isolated from each other, and their pixels do not overlap. Furthermore, the details of the support on which the document was put (background Fig.3.7-b) are preserved. The ink pixels are also well separated with the A-ONMF, EM-ONMF, and ONP-MF methods. The EM-ONMF does

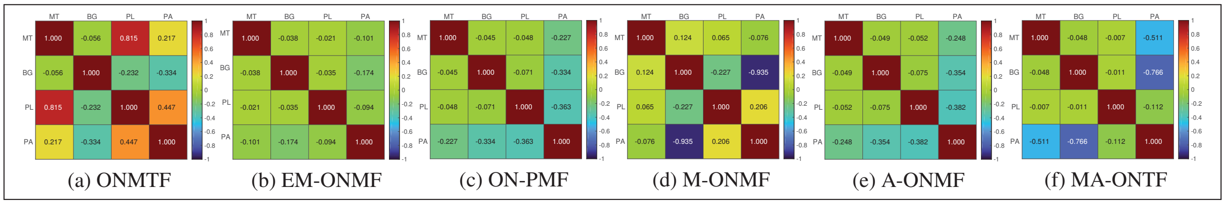


(a) Pseudo-color (b) GT (c) Color scale

(d) Main text (e) Background (f) Printed line (g) Paper

Figure 3.7 Abundance maps of different materials decomposed by different ONMF methods applied on a sample from MSTEx-1 dataset. (a) pseudo-colour image of the sample $z802$ generated from 3 channels of the visible band, (b) corresponding GT image, (c) color scale, (d) text material, (e) background (table support, fold, and staple), (f) printed line, and (g) paper

not preserve the details of the materials as it performs hard clustering. The same behavior can be observed in Fig. I-1, which represents a very challenging scenario due to the presence of several overlapped materials (handwriting text, machine writing, and stamps).



(a) ONMTF (b) EM-ONMF (c) ON-PMF (d) M-ONMF (e) A-ONMF (f) MA-ONTF

Figure 3.8 Heat maps representation of Pearson's correlation coefficient matrices. Pairwise correlation is calculated for the four abundance maps corresponding to MT (main text), BG (background), PL (printed line), PA (paper)

Besides, for objective assessment purposes, we have also computed the correlation between images of different materials obtained in Fig.3.7. The correlation matrices shown in Fig.3.8

demonstrate the effectiveness of proposed models. We also can observe that the correlation matrix corresponding to the ONMTF method is not diagonal. Indeed, the corresponding images in the first row in Fig.3.7 shows that the pixels of different materials overlap spatially over all images.

3.5.7 Binarization enhancements of MS document images

In this set of experiments, we evaluate the proposed models' performance on the binarization task, which is considered as a primary stage of the document image analysis and understanding pipeline. This step produces valuable outputs for the subsequent processing tasks, such as text recognition, writers identification, etc. Therefore, it is evident to deduce that a good decomposition yields automatically to good text extraction and binarization.

However, the issue that arises here is that the outputs (abundance matrices) of all orthogonal NMF models (except for the EM-ONMF) are not binary. Therefore, it is not possible to perform a qualitative comparison with binary GT images without converting these outputs to binary forms. To this end, we adopted two techniques. The first one consists of thresholding (globally) the image intensity. In the second, we used Howe's (Howe, 2013) methods[18], one of the best-known techniques for text images binarization that won the HDIBCO 2012 Competition (Pratikakis, Gatos & Ntirogiannis, 2012). Howe's method is an energy-based binarization technique that takes one single band as input[19] and produces one binary image. Moreover, we compare the results of two other approaches that have been proposed to address the same challenge, i.e., text extraction on MSTEx-2 document images, namely SKKHM (Li *et al.*, 2007) and GMMs *et al*. (Hollaus *et al.*, 2018b)(Gaussian Mixture Model). In this scenario, the binarization performance is measured through four metrics: F1-score (FM), DRD, NRM, and ACC. The results of different methods are reported in Table 3.3, where best scores are set in bold and underlined, and the second-best scores are in bold.

[18] Code is available at http://www.science.smith.edu/~nhowe/research/code/

[19] In our experiment, we used Algorithm 3 that offers automatic tuning of parameters. We applied it to pseudo color images generated by combining the three-channel images of the visible band.

Table 3.3 Binarization results of different methods on MSTEx-2 dataset. In bold and underlined, best performance; bold only, second-best scores. FM & ACC are expressed in %, DRD $\times 10^{-3}$, and NRM $\times 10^{-2}$. Size of MS cubes reads bands×width× height

MS-cube (size)	Metric	RICA	Howe	SKKHM	GMM	EM-ONMF	Factorization + Threshold				Factorization + Howe				
							ONP-MF	M-ONMF	A-ONMF	MA-ONTF	ONMTF	ONP-MF	M-ONMF	A-ONMF	MA-ONTF
z27 (8x330x1001)	FM	53.92	78.57	70.59	81.67	79.76	79.94	81.49	81.56	81.27	82.04	78.33	**84.36**	80.62	**85.08**
	DRD	15.31	6.96	10.35	4.86	5.28	5.25	5.77	5.17	5.43	4.88	5.45	**4.57**	5.18	**4.26**
	NRM	26.19	10.20	14.50	11.67	13.17	12.98	8.65	10.64	10.00	9.70	14.73	**6.84**	11.85	**6.76**
	ACC	88.44	94.31	92.03	95.60	95.22	95.25	95.12	95.41	95.24	95.45	95.02	**95.87**	95.29	**96.09**
z31 (8x628x1228)	FM	43.51	66.22	59.34	**86.25**	65.93	67.81	69.93	69.68	70.86	67.18	67.48	68.05	68.28	**80.84**
	DRD	13.20	12.48	13.01	**2.77**	10.23	10.15	10.77	12.15	9.98	12.90	10.22	11.00	10.93	**4.70**
	NRM	31.06	9.29	14.23	**5.15**	15.07	13.16	8.48	**5.17**	9.29	7.03	13.42	10.46	10.36	8.43
	ACC	95.71	96.41	96.13	**98.80**	96.91	97.01	96.94	96.64	97.14	96.37	96.98	96.81	96.84	**98.34**
z43 (8x625x1933)	FM	51.56	67.89	50.01	**81.00**	55.30	78.57	63.90	70.28	77.95	75.21	78.44	67.59	68.64	**80.81**
	DRD	14.89	10.78	13.56	**4.56**	11.68	5.07	11.12	7.89	6.94	8.91	5.09	11.22	8.32	**3.25**
	NRM	25.01	13.33	27.58	9.14	25.22	12.74	17.90	16.20	**8.31**	**6.50**	12.87	12.99	17.09	14.71
	ACC	93.64	95.43	93.97	**97.54**	94.70	97.40	95.22	96.29	96.94	96.24	97.38	95.31	96.09	**98.80**
z582 (8x225x1333)	FM	19.21	19.94	18.13	**79.09**	76.81	71.89	73.01	71.11	67.33	**79.87**	74.80	77.61	78.22	76.86
	DRD	28.06	11.51	71.34	**4.71**	7.22	9.60	6.66	7.09	7.83	5.30	8.30	6.05	**5.12**	7.03
	NRM	41.94	44.37	36.54	11.21	**8.77**	10.17	15.94	17.25	19.01	9.46	12.16	11.12	13.06	**9.13**
	ACC	91.04	95.45	79.55	**97.94**	97.47	96.81	97.46	97.31	96.95	**97.94**	97.12	97.73	97.93	97.50
z592 (8x366x1380)	FM	63.70	81.53	72.92	**86.46**	80.11	74.27	81.10	83.07	82.71	81.77	74.81	**86.65**	85.75	83.70
	DRD	8.89	5.51	10.08	**3.19**	6.01	9.64	4.62	4.58	3.91	5.54	9.25	3.49	**3.29**	3.75
	NRM	21.19	6.42	7.09	**4.67**	7.61	5.85	10.60	7.79	10.94	6.01	6.01	**4.67**	6.97	9.43
	ACC	96.62	97.98	96.69	**98.56**	97.85	96.82	98.13	98.25	98.34	97.99	96.93	**98.58**	**98.56**	98.40
z65 (8x299x1490)	FM	67.27	84.35	74.05	82.27	72.98	68.65	75.66	78.97	76.46	84.99	70.12	**85.49**	78.71	**87.97**
	DRD	9.01	4.21	6.55	4.45	5.10	12.62	6.50	4.15	5.78	4.32	11.40	**3.91**	5.85	**2.45**
	NRM	17.46	6.68	14.61	9.28	19.62	8.56	12.89	15.29	13.75	**4.83**	8.97	**5.58**	10.37	7.20
	ACC	94.89	97.46	96.05	97.23	96.49	93.73	96.20	97.11	96.47	97.48	94.25	**97.62**	96.59	**98.17**
z802 (8x339x1680)	FM	40.93	81.28	81.52	**92.73**	73.82	88.67	90.61	89.25	**91.53**	82.82	88.69	90.15	87.27	89.28
	DRD	37.86	7.31	5.25	**1.53**	4.79	2.36	2.62	2.46	**1.94**	5.09	2.35	2.66	2.93	2.39
	NRM	18.52	**2.35**	10.28	4.12	20.70	8.47	**3.97**	6.87	4.96	5.43	8.50	4.43	8.11	5.79
	ACC	92.66	98.37	98.67	**99.47**	98.48	99.22	99.30	99.24	**99.39**	98.61	99.22	99.27	99.10	99.22
z822 (8x286x1934)	FM	67.85	84.64	79.09	78.26	15.38	85.09	85.18	84.31	80.15	84.31	84.83	86.49	**86.80**	**87.48**
	DRD	6.79	2.64	3.93	4.51	144.93	2.65	2.80	2.69	3.30	2.79	2.66	2.29	**2.32**	**2.21**
	NRM	20.07	10.41	14.13	13.24	30.30	10.26	9.37	11.55	15.13	10.20	10.73	9.24	**8.27**	**8.86**
	ACC	97.59	98.75	98.40	98.27	66.04	98.83	98.81	98.80	98.55	98.75	98.82	**98.93**	**98.93**	**98.99**
z90 (8x489x1359)	FM	43.64	67.56	51.31	**83.77**	56.63	55.90	58.27	64.34	57.37	76.07	56.57	78.89	67.41	**86.22**
	DRD	35.97	15.73	26.81	**4.68**	19.52	25.58	15.46	15.69	18.79	9.23	24.45	7.47	13.76	**3.36**
	NRM	17.77	**7.12**	15.43	7.75	16.77	10.57	19.81	15.74	16.79	6.93	10.98	7.16	11.10	**6.67**
	ACC	91.06	95.89	**98.20**	98.42	94.85	93.61	95.69	85.65	95.02	97.35	93.87	97.77	96.27	**98.68**
z92 (8x262x485)	FM	58.35	72.11	69.88	69.01	78.42	71.79	80.74	**81.89**	75.52	77.33	77.87	**82.72**	81.09	80.32
	DRD	11.35	8.68	10.74	10.09	6.14	9.78	4.04	4.42	6.81	6.57	6.30	**3.56**	4.64	**3.76**
	NRM	22.65	10.46	8.93	10.47	7.57	7.52	11.49	9.89	10.8	**7.18**	7.37	9.55	**6.26**	12.26
	ACC	92.35	94.10	93.06	93.05	95.57	95.39	93.49	**96.76**	95.18	95.22	95.39	**96.87**	96.37	96.58
Average	FM	50.67	70.41	62.68	**82.05**	65.51	74.24	75.99	77.45	76.12	79.16	75.19	80.64	78.44	**83.86**
	DRD	21.99	8.58	17.16	**4.54**	22.09	9.27	7.04	6.63	7.07	6.55	8.55	5.73	6.13	**3.72**
	NRM	23.10	12.06	16.33	8.67	16.48	10.03	11.91	11.64	11.90	**7.33**	10.57	**7.88**	10.67	8.92
	ACC	87.18	96.42	93.78	**97.49**	93.36	96.22	96.95	96.15	96.92	97.14	96.50	97.43	97.25	**98.08**

By analyzing the results reported in Table 3.3, we can make the following observations: i) factorization based approaches outperform the state-of-the-art binarization approaches Howe in all experiments in terms of all metrics; ii) our proposed models outperform all other orthogonal NMF methods, and the RICA method ranked the last; iii) the factorization improved the performance of Howe's method dramatically. For instance, combining MA-ONTF with Howe's method achieved 13.45% more scores in terms of the FM metric than Howe's method alone. iv) the second-best performance is achieved by the GMMs method. However, this method is designed especially for the MSTEx dataset, which makes it not useful to handle other datasets as it without modification.

Since the MSTEx-2 collection was officially adopted for the MSTEx 2015 Competition (Hedjam *et al.*, 2015), we though it is judicious to compare the performances of our approaches to the methods that were submitted to that contest. Thus, besides the algorithms we considered in Table 3.3, the top two methods that won the MSTEx context, namely MSIO (Diem *et al.*, 2016) and Hollas *et al*. (Hedjam *et al.*, 2015), have been also compared as they were developed especially for the same purpose. In addition, we also include the results of a recent Convolutional Neural Network (CNN) based method developed by Hollaus *et al*. (Hollaus *et al.*, 2019). In Table 3.4, we report the average performances obtained by all methods.

Taking into account the results illustrated in Figs.3.5-3.8, Figs. I-1-I-2, Table 3.3, and Table I-2, it is not surprising to see in Table 3.4 that the MA-ONTF method achieves the top performance among all methods, including the two winners of the MSTEx 2015 contest. Two of our models, M-ONMF and MA-ONTF, outperform all other orthogonal NMF methods. The ONP-MF method, in turn, performs better than the EM-ONMF, which confirms the observation made in (Pompili *et al.*, 2014). Moreover, due to the lake of labeled MS document image datasets, it is not surprising to see the low performance achieved by the CNN-based technique (Hollaus *et al.*, 2019).

Table 3.4 Average scores in the binarization of the MSTEx-2 dataset and comparison with state-of-the-art methods

Method	FM (%)	DRD ($\times 10^{-3}$)	NRM ($\times 10^{-2}$)
MA-ONTF (Ours)	**83.86**	**3.72**	8.92
MSIO (Diem *et al.*, 2016)	83.33	4.24	9.25
GMM (Hollaus *et al.*)(Hollaus *et al.*, 2018b)	82.95	4.16	8.47
Hollaus *et al.* (Hedjam *et al.*, 2015)	81.90	4.74	10.10
M-ONMF (Ours)	80.64	5.73	7.88
Hollaus *et al.* (CNN)(Hollaus *et al.*, 2019)	79.90	4.71	9.44
ONMTF (Yoo & Choi, 2010)	79.16	6.55	**7.33**
A-ONMF (Ours)	78.44	6.13	10.67
ONP-MF (Pompili *et al.*, 2014)	75.19	8.55	10.57
Howe (Howe, 2013)	70.41	8.58	12.06
SKKHM (Li *et al.*, 2007)	62.68	17.16	16.33
EM-ONMF (Pompili *et al.*, 2014)	65.51	22.09	16.48
RICA (Le *et al.*, 2011)	50.67	21.99	23.10

In order to give a better idea about the content complexity of the MSTEx-2 collection, it is helpful to examine some samples visually. Thus, another qualitative evaluation is provided in Fig.3.9 to assess the effectiveness of the proposed methods more deeply.

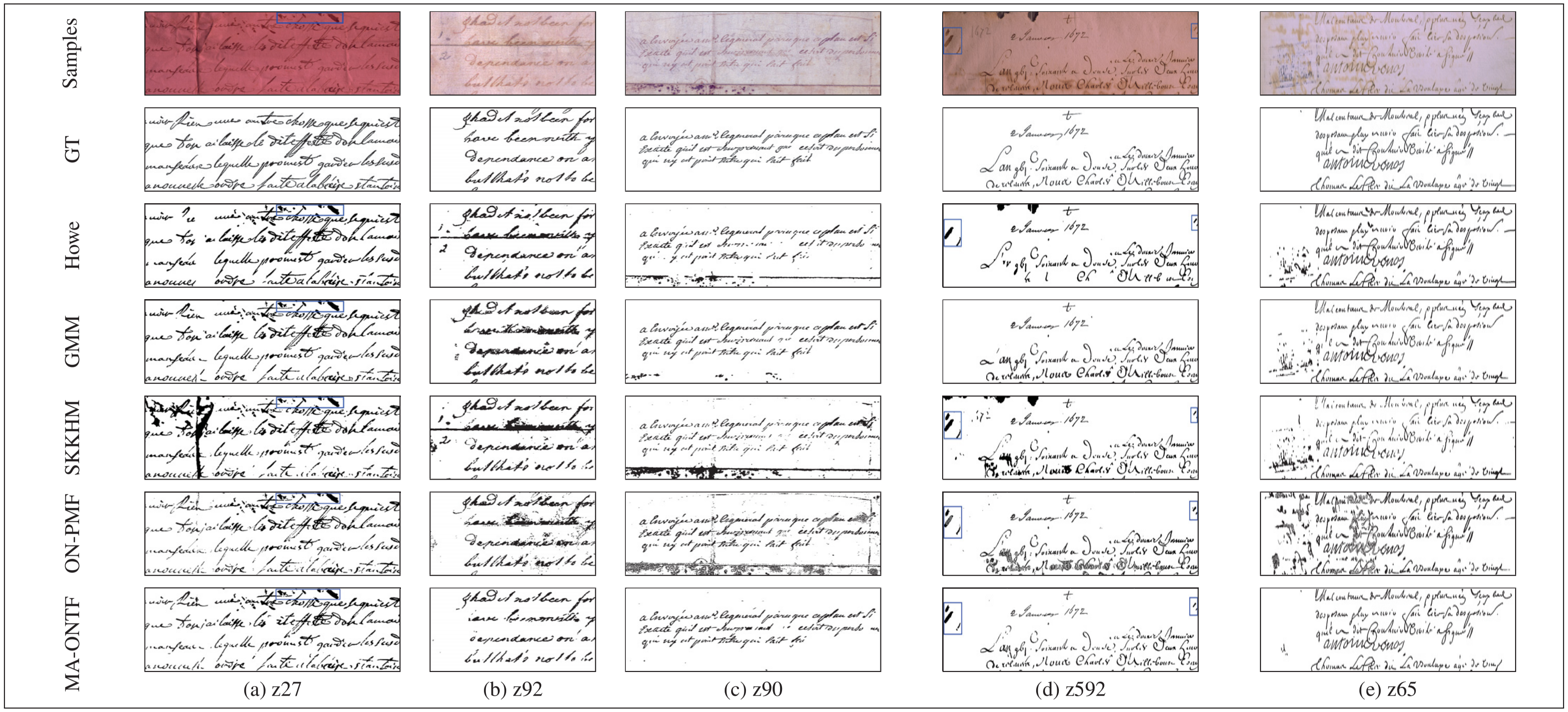
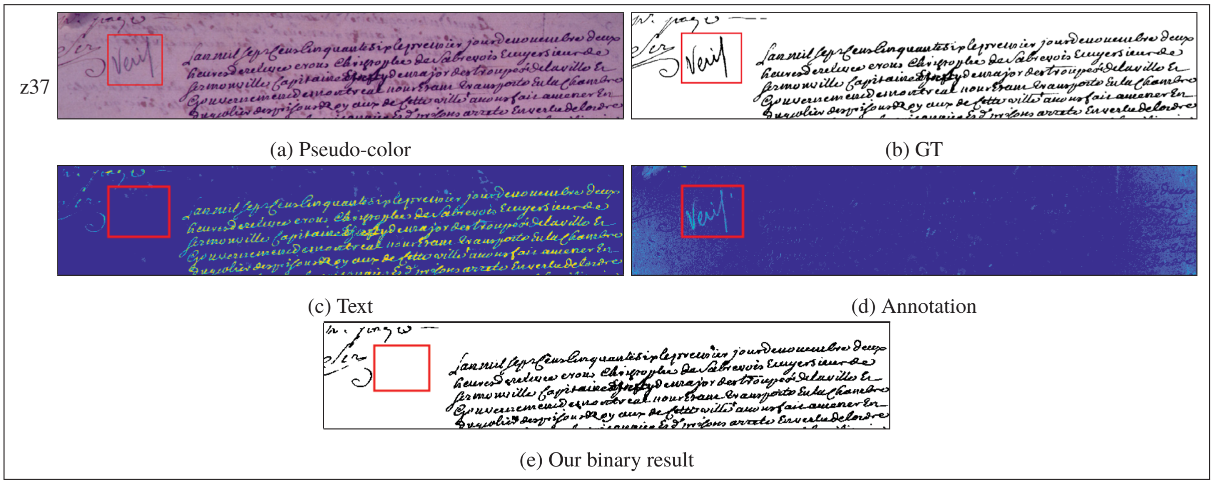


Figure 3.9 Subjective assessment of the binarization results of some complex scenes from the MSTEx-2 dataset. Blue boxes surround inkblots that are present in input images, the output of ONP-MF and MA-ONTF, but removed from GT images

Fig.3.9 compares the binarization results obtained by our MA-ONTF method, ONP-MF, SKKHM, GMM, and Howe's method to the GT images provided. The results obtained by our MA-ONTF approach are conclusive. In most cases, the proposed method separates the text from the degradation and generates images comparable to the GT images. Both ONP-MF and GMMs methods perform well also. However, Howe's approach and SKKHM fail in handling the degradations in complex scenes. Indeed, this is well reflected by the scores obtained in the Tables shown previously. Moreover, we observed that most approaches were unable to remove inkblots present on image scenes (areas delimited by blue rectangles in the corresponding figures). These inkblots are not present in the GT because they were manually removed. However, they are present in the outputs of many approaches, including our models. This particularly explains why it is very challenging to obtain highest scores with some images.

Distinguishing between different types of inks present in document scenes is another interesting advantage of our approaches, which represents valuable complementary information for further document analysis. Fig.3.10 shows an example where our approach was able to distinguish the annotation from the main text written by two different inks. We notice that the corresponding GT provided did not consider this issue. Thus, the two objects belong to the same class (text).



(a) Pseudo-color

(b) GT

(c) Text

(d) Annotation

(e) Our binary result

Figure 3.10 Inks differentiation in MS document images. Red box surrounds an annotation written by a different ink, (a) pseudo-colour image of z37 sample, (b) GT image of z37 sample, (c) and (d) abundance maps of the text and annotation obtained via MA-ONTF, (e) binary output image

3.5.8 Further Discussion

Once again, optimization over the Stiefel manifold has demonstrated to be useful to handle problems with orthogonality constraints. From the conceptual point of view, the common factor among the A-ONMF, ONP-MF, and EM-ONMF models is constraining the abundance matrix rows to be orthogonal. In our experiments, we observed that A-ONMF outperforms by far the EM-ONMF model. The M-ONMF model outperforms both EM-ONMF and ONP-MF. The MA-ONTF model is compared to ONMTF as it imposes the orthogonality constraints on both factors. Nevertheless, our MA-ONTF exhibits superior segmentation performance compared with ONMTF and all other orthogonal NMF variants in all experiments.

It is worthy of mentioning that the three models proposed here rely neither on a specific number of bands, specific wavelengths, nor a given order of bands in the input matrix, which expand their application scope, unlike other methods, such as GMMs, that is designed to work exactly on the same set of bands and does not allow permutation between bands. Indeed, it is not possible to apply the GMM-based method in its current form on the Hubble dataset or any other dataset. In our experiments, since each document image is unique, tuning the parameters of the proposed models is needed to achieve the best binarization performances. However, based on our empirical validation and analysis, tuning the parameters for each MS image will not be required if the images are scanned from the same document (book), which means that similar spectral signatures will be extracted over the range of the background and inks for all images.

Finally, we conclude this section with a number of observations worth making. Selecting the suitable model to use would depend upon the problem at hand and the nature of the data. Clearly, nothing in the three objective functions of the proposed models constrain them to be used only for MS document images. Thus, they could be used for other applications. Based on our empirical experiments, the bi-orthogonal MA-ONTF model and the M-ONMF model could be the first choices if the problem at hand allows constraining the basis matrix to be orthogonal. The bi-orthogonal MA-ONTF is an effective decomposition model that shows the highest efficiency, but its running time has to be taken into account. The results of the ONMTF, which is a bi-orthogonal NMF model also, confirm this observation. Outside our application, the MA-ONTF is more suitable for co-clustering applications such as community detection, recommendation system, and document clustering. In such applications, each of the two-factor matrices contains clues about the data from a different perspective that needs to be clustered separately. The M-ONMF model is a best-tradeoff between time complexity and performance. It would be preferred where the clustering is performed on the basis matrix M. The last model, i.e., A-ONMF, can be used for a wide range of applications where matrix A contains the information to be clustered.

In general, ADMM based algorithms belong to the category of algorithms that have a high cost per iteration and fast convergence (small number of iterations). In contrast, there is another category of algorithms, such as the multiplicative update, that have a low cost per iteration but they are slow in convergence. Thus, a GPU implementation of the proposed algorithms will decrease the cost per iteration and accelerate the update steps.

3.6 Conclusion

This paper addresses the problem of MS document image decomposition using three new orthogonal NMF methods. The core idea was to incorporate the orthogonality as a Riemannian optimization on the Stiefel manifold. Based on that, we developed three new ONMF models. We validated our methods on various challenging datasets. We experimentally demonstrate the effectiveness of the proposed models and their capability to handle blind decomposition of MS document images efficiently. We confirmed the power of generalization of the proposed approaches on the HS dataset. We conducted qualitative and quantitative evaluations of the proposed models through several experiments and compared them to state-of-the-art techniques. The experiments we conducted demonstrate that our methods achieved the top performance among all the benchmarks used. However, some relevant issues remain to be investigated in future works. The most critical one concerns the rank selection step in order to make it automatic and less problematic. An adaptive selection of different hyperparameters will also be investigated. A GPU implementation of the proposed methods is also planned for the near future.

CHAPTER 4

NONLINEAR ORTHOGONAL NMF ON THE STIEFEL MANIFOLD WITH GRAPH-BASED TOTAL VARIATION REGULARIZATION

Abderrahmane Rahiche[a] , Mohamed Cheriet[a]

[a] Department of Systems Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article submitted to IEEE Signal Processing Letters, Jan. 2022.

Abstract

This paper proposes a novel Nonlinear Orthogonal NMF model with Graph-based Total Variation Regularization (GTV) for Multispectral document images decomposition. In this model, a GTV regularization is incorporated to preserve the intrinsic geometrical structure of document content that is generally lost by the vectorization of spectral images. A spatial orthogonality constraint over the Stiefel manifold is included to improve the sparsity of the solution and ensure its uniqueness. The kernel trick is involved to account for the non-linear correlation inherent to spectral data. We devised an efficient algorithm to solve the resulting problem and implemented it using recent fast kernel operations toolboxes that allowed us to process larges data images without memory overflows. The experimental results on real data shows that the proposed model achieves better decomposition performance compared to recent competitive methods and outperforms some existing state-of-the-art methods.

Keywords

Non-linear nonnegative matrix factorization, Graph total variation, Orthogonality, Stiefel Manifold, Image decomposition, Multispectral document image.

4.1 Introduction

Multispectral (MS) imagery is an effective non-destructive imaging technology that continues to attract increasing attention from researchers in the field of historical document analysis. However, most related MS document image processing practices still rely on traditional methods developed for conventional RGB and gray-scale images.

Document image decomposition is a challenging problem aiming to split the image according to its constituting materials (objects). Nevertheless, due to the variability of the content of documents and the variability of the writing materials used to produce them, it is tough to guess prior information about these objects' signatures, shapes and texture, and number, not their locations. Thus, unsupervised methods, especially blind source separation (BSS) approaches, are more practical to address this problem.

A set of MS images is generally modelled as a static linear mixture model (LMM) (Settle & Drake, 1993) of the form

$$\mathbf{X} = \mathbf{M}\mathbf{A} + \mathbf{N}, \tag{4.1}$$

where $\mathbf{X} \in \mathbb{R}_+{}^{b\times n}$ denotes an MS data matrix consisting of b spectral bands and n pixels, $\mathbf{M} \in \mathbb{R}_+{}^{b\times k}$ represents the spectral endmembers responses, $\mathbf{A} \in \mathbb{R}_+{}^{k\times n}$ represents the fraction abundance matrix, $\mathbf{N} \in \mathbb{R}^{b\times n}$ is the noise matrix, and k is the number of materials present in the image scene. In practice, the data matrix $\mathbf{X}$ is obtained by converting the corresponding MS data-cube into a 2D matrix, where the spectral images from the MS data-cube are vectorized and stacked together to form the rows of the matrix $\mathbf{X}$.

Nonnegative matrix factorization (NMF) is one of the famous BSS techniques. It aims to factorize an input nonnegative data matrix, $\mathbf{X}$, into the product of two nonnegative low-rank matrices, that is $\mathbf{X} \approx \mathbf{M}\mathbf{A}$. NMF fits exactly the model in (4.1), where the matrices $\mathbf{M}$ and $\mathbf{A}$ and the number of materials $k \ll \{b, n\}$ are all unknown. Formally, the NMF problem can be written as the following optimization problem

$$\min_{\mathbf{M},\mathbf{A}} \frac{1}{2}\|\mathbf{X} - \mathbf{M}\mathbf{A}\|^2 \qquad \text{s.t. } \mathbf{M}, \mathbf{A} \geq 0, \tag{4.2}$$

The standard objective in (4.2) accounts only for nonnegativity of the low-rank factors. This condition is primordial for positive data, and it is very useful for results interpretation. However, the nonnegativity alone is not enough to obtain a unique solution as NMF is known to be an ill-posed problem (Vavasis, 2010), and its objective function is non-convex (Lee & Seung, 1999). This issue can be addressed by imposing additional constraints and regularization terms to the model, such as the orthogonality and sparsity (Hoyer, 2004; Choi, 2008), to narrow down the solutions space.

Due to the vectorization of MS images required by NMF, the intrinsic geometrical structure of input images is affected, making the standard NMF incapable of preserving such a spatial structure. Indeed, texts on document images are generally sparse and form small shapes (characters) that are structured in lines. This issue can be addressed by incorporating a graph regularization term into the NMF model to preserve intrinsic structure of data. Although graph NMF variants can outperform standard NMF, most existing approaches are restricted to two graph signal priors, namely the graph Laplacian (Cai, He, Han & Huang, 2010; Pei, Wu & Chen, 2014) and the similarity matrix (Kuang, Ding & Park, 2012).

Furthermore, the objective in (4.2) ignores the correlation between the spectral bands, which is generally present in spectral data. Kernel-based variants in (Buciu *et al.*, 2008; Zafeiriou & Petrou, 2009; Pan *et al.*, 2011) have been proposed to handle such a nonlinear correlation. Although the success of kernel models and their ability to discover more latent features compared to standard NMF, most models (Zhang & Liu, 2009; An *et al.*, 2011; Tolić *et al.*, 2018) suffer from the so-called pre-images problem (Kwok & Tsang, 2004), i.e., they cannot estimate the basis matrix in its initial space. Besides, some models involve explicit computation of a huge Gram matrix and suffer from the scalability issue when the number of samples is large (Rahiche & Cheriet, 2021b).

Thus, in this work, we propose a new NMF model that simultaneously addresses the three issues mentioned above: spectral non-linearity, uniqueness of the solution via spatial orthogonality, and spatial local structure preservation. The resulting optimization problem is solved using the Alternating Direction Method of Multipliers (ADMM) (Boyd *et al.*, 2011).

4.1.1 Preliminaries on graphs and notations

A graph representing N pixels of a multispectral data cube can be denoted by a tuple $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{W}\}$, where $\mathcal{V} = \{v_1, \cdots, v_n\}$ denotes a set of n vertices whose elements are the set of MS image pixels, $\mathcal{E} = \{e_{ij} | i \sim j\}$ is a set of edges, where e_{ij} represents an edge spanning two connected vertices v_i and v_j (denoted $v_i \sim v_j$), and $\mathcal{W} \in \mathbb{R}^{n \times n}$ is the corresponding adjacency matrix. For the points p_i, p_j, the non-negative edge weight w_{ij} can be calculated using a Gaussian kernel: $w_{ij} = \exp(-\gamma \|p_i - p_j\|_2^2)$, where $\gamma = 1/(2\sigma^2)$ denotes the precision and σ is the kernel's bandwidth.

In contrast to existing studies that have focused on using the graph Laplacian as a quadratic (Tikhonov) regularizer (Cai *et al.*, 2010; Tolić *et al.*, 2018), in our model, we incorporate the spatial information via the total variation (TV) of the graph data that allows quantifying the smoothness of the graph and promotes sparse graph gradient. This GTV regularization can be expressed by (Hütter & Rigollet, 2016; Berger, Hannak & Matz, 2017):

$$\|\mathbf{A}\|_{TV} = \|\mathbf{\Gamma} \mathbf{A}^T\|_1, \tag{4.3}$$

where $\|\mathbf{A}\|_{TV}$ denotes the TV of $\mathbf{A}$ w.r.t. graph $\mathcal{G}$, $\mathbf{\Gamma} \in \mathbb{R}^{|\mathcal{E}| \times n}$ denotes the incidence matrix of $\mathcal{G}$, and $\|.\|_1$ is the $L1$-norm.

4.1.2 Contribution

The contributions of this work can be summarized as follows:

1) We propose a novel kernel-based NMF model that does not suffer from the pre-image problem.
2) We incorporate the orthogonality as an optimization problem over the Stiefel manifold.
3) We include a GTV regularizer to preserve the intrinsic structure of data while promoting sparse graph gradients.
4) We devise an efficient ADMM-based algorithm to solve the proposed model.
5) We address the problem of MS document image decomposition.

4.2 Proposed Orthogonal Non-linear NMF with Graph-based Total Variation regularization

In this section, we describe the main ingredients of the proposed model we called GTV-ONNMF.

To incorporate a spectral non-linearity into the model in (4.1), we consider a nonlinear mapping of the original data matrix $\mathbf{X}$ into a higher dimensional feature space $\mathbf{\Phi}(\mathbf{X}) = (\mathbf{\Phi}(\mathbf{x_1}), \mathbf{\Phi}(\mathbf{x_2}), \cdots, \mathbf{\Phi}(\mathbf{x_n})) \in \mathbb{R}^{d\times n}$, with $d \gg n$. In contrast to previous studies (Zhang & Liu, 2009; An *et al.*, 2011; Tolić *et al.*, 2018), to avoid the pre-image problem we also consider a nonlinear mapping of the basis matrix $\mathbf{M}$ into another feature space of higher dimensions $\mathbf{\Phi}(\mathbf{M}) = (\mathbf{\Phi}(\mathbf{m_1}), \mathbf{\Phi}(\mathbf{m_2}), \cdots, \mathbf{\Phi}(\mathbf{m_k})) \in \mathbb{R}^{d\times k}$, with $d \gg k$. Thus, problem in (4.2) can be reformulated as

$$\min_{\mathbf{M},\mathbf{A}} \|\mathbf{\Phi}(\mathbf{X}) - \mathbf{\Phi}(\mathbf{M})\mathbf{A}\|_F^2 \quad \text{s.t.} \quad \mathbf{M}, \mathbf{A} \geq 0, \tag{4.4}$$

By replacing the Frobenius norm with the trace operator, Eq.(4.4) can be equivalently written as

$$\min_{\mathbf{M},\mathbf{A}} \frac{1}{2}\mathrm{Tr}(\mathbf{K}_{xx} - 2\mathbf{K}_{xm}\mathbf{A} + \mathbf{A}^T\mathbf{K}_{mm}\mathbf{A}) \quad \text{s.t.} \quad \mathbf{M}, \mathbf{A} \geq 0, \tag{4.5}$$

where $\mathbf{K}_{xx} = \langle\, \mathbf{\Phi}(\mathbf{X}), \mathbf{\Phi}(\mathbf{X})\rangle$ is a kernel matrix of size $n \times n$ associated to the matrix $\mathbf{X}$, $\mathbf{K}_{mm} = \langle\, \mathbf{\Phi}(\mathbf{M}), \mathbf{\Phi}(\mathbf{M})\rangle$ is a kernel matrix of size $k \times k$ associated to the matrix $\mathbf{M}$, and $\mathbf{K}_{xm} = \langle\, \mathbf{\Phi}(\mathbf{X}), \mathbf{\Phi}(\mathbf{M})\rangle$ is a kernel matrix of size $n \times k$.

Similar to (Rahiche & Cheriet, 2021b,a), we address the uniqueness issue via spatial orthogonality over Stiefel manifold $\mathcal{M}$. This condition is imposed on the abundance matrix $\mathbf{A}$, such that

$$\mathbf{A}^T \in \mathcal{M} := St(k, n) = \{\mathbf{A}^T \in \mathbb{R}^{n\times k} | \mathbf{A}\mathbf{A}^T = \mathbf{I}_k\} \tag{4.6}$$

In order to preserve the spatial structure lost by the vectorization, we add the graph-based total variation regularizer defined in (4.3). This term promotes sparse graph gradients.

We finally obtain the following optimization problem

$$\begin{aligned} &\min_{\mathbf{M},\mathbf{A}} \frac{1}{2}\mathrm{Tr}(-2\mathbf{K}_{xm}\mathbf{A} + \mathbf{A}^T\mathbf{K}_{mm}\mathbf{A}) + \lambda\|\mathbf{\Gamma}\mathbf{A}^T\|_1 + \imath_{R_+}(\mathbf{A}) \\ &\qquad\qquad \text{s.t.} \quad \mathbf{M} \geq 0, \mathbf{A}^T \in \mathcal{M} \end{aligned} \tag{4.7}$$

where $\lambda > 0$ is a regularization parameter. Note that the term $\mathrm{Tr}(\mathbf{K}_{xx})$ in (4.5) is constant w.r.t. the two variables $\mathbf{M}$, and $\mathbf{A}$. Therefore, we drop it from the objective function (4.7). Moreover, the nonnegativity constraint imposed on $\mathbf{A}$ is replaced by the following indicator function

$$\imath_{R_+}(\mathbf{A}) = \begin{cases} 0, & \text{if} \quad \mathbf{A} \geq 0. \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.8}$$

We can do the same with the factor $\mathbf{M}$, but we found that we can handle the nonnegativity differently, see the next section.

4.3 Optimization using ADMM

The objective function in Eq.4.7 is bi-convex and contains non-differential terms that cannot be optimized efficiently with a direct approach. Therefore, we apply the ADMM technique to split the problem into four distinct sub-problems that are easier to solve individually and use proximal operators to address the non-differential terms. Thus, by incorporating two auxiliary variables $\mathbf{Y} \in \mathbb{R}^{|\mathcal{E}|\times k}$ and $\mathbf{Z} \in \mathbb{R}^{k\times n}$ into the objective function (4.7), the problem then can be reformulated as a minimization problem of four variables

$$\begin{aligned}&\min_{\mathbf{M},\mathbf{A}} \frac{1}{2}\mathrm{Tr}(-2\mathbf{K}_{xm}\mathbf{A} + \mathbf{A}^T\mathbf{K}_{mm}\mathbf{A}) + \lambda\|\mathbf{Y}\|_1 + \iota_{R_+}(\mathbf{Z}) \\ &\text{s.t.} \quad \mathbf{M} \geq 0, \mathbf{A}^T \in \mathcal{M}, \mathbf{Y} = \mathbf{\Gamma}\mathbf{A}^T, \mathbf{Z} = \mathbf{A}\end{aligned} \tag{4.9}$$

The augmented Lagrangian for the above problem is given by

$$\begin{aligned}\mathcal{L}(\mathbf{M}, \mathbf{A}, \mathbf{Y}, \mathbf{Z}, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) &= \frac{1}{2}\mathrm{Tr}(\mathbf{A}^T\mathbf{K}_{mm}\mathbf{A} - 2\mathbf{K}_{xm}\mathbf{A}) \\ &+ \lambda\|\mathbf{Y}\|_1 + \iota_{R_+}(\mathbf{Z}) + \langle \mathbf{\Lambda}_1, \mathbf{\Gamma}\mathbf{A}^T - \mathbf{Y}\rangle + \frac{\rho}{2}\|\mathbf{\Gamma}\mathbf{A}^T - \mathbf{Y}\|_2^2 \\ &+ \langle \mathbf{\Lambda}_2, \mathbf{A} - \mathbf{Z}\rangle + \frac{\rho}{2}\|\mathbf{A} - \mathbf{Z}\| \quad \text{s.t.} \quad \mathbf{M} \geq 0, \mathbf{A}^T \in \mathcal{M},\end{aligned} \tag{4.10}$$

where $\mathbf{\Lambda}_1 \in \mathbb{R}^{|\mathcal{E}|\times n}$ and $\mathbf{\Lambda}_2 \in \mathbb{R}^{k\times n}$ are the Lagrangian variables, ρ is the penalty parameter, and $|\mathcal{E}|$ is the cardinality of $\mathcal{E}$. Using ADMM to solve this problem yields the following optimization sub-problems

$$\begin{cases}\mathbf{M}^{t+1} := & \arg\min_{\mathbf{M}\geq 0} \frac{1}{2}\mathrm{Tr}(\mathbf{A}^T\mathbf{K}_{mm}\mathbf{A} - 2\mathbf{K}_{xm}\mathbf{A}), \\ \mathbf{A}^{t+1} := & \arg\min_{\mathbf{H}\mathbf{H}^T=\mathbf{I}} \frac{1}{2}\mathrm{Tr}(\mathbf{A}^T\mathbf{K}_{mm}^{t+1}\mathbf{A} - 2\mathbf{K}_{xm}^{t+1}\mathbf{A}) + \\ & \frac{\rho}{2}\|\mathbf{\Gamma}\mathbf{A}^T - \mathbf{Y} + \frac{\mathbf{\Lambda}_1}{\rho}\|_2^2 + \frac{\rho}{2}\|\mathbf{A} - \mathbf{Z} + \frac{\mathbf{\Lambda}_2}{\rho}\|_2^2, \\ \mathbf{Y}^{t+1} := & \arg\min_{\mathbf{Y}} \lambda\|\mathbf{Y}\|_1 + \frac{\rho}{2}\|\mathbf{\Gamma}(\mathbf{A}^{t+1})^T - \mathbf{Y} + \frac{\mathbf{\Lambda}_1}{\rho}\|_2^2, \\ \mathbf{Z}^{t+1} := & \arg\min_{\mathbf{Z}\geq 0} \iota_{R_+}(\mathbf{Z}) + \frac{\rho}{2}\|(\mathbf{A}^{t+1} + \frac{\mathbf{\Lambda}_2}{\rho}) - \mathbf{Z}\|_2^2.\end{cases} \tag{4.11}$$

4.3.1 Optimize M for fixed A, Y and Z

The minimization of the first sub-problem in (4.11) involves calculating its partial derivative w.r.t. **M** and setting it equal to 0. For radial basis function kernel (RBF), the final expression of the gradient of $\mathcal{L}(\mathbf{M}, \mathbf{A}, \mathbf{Y}, \mathbf{Z}, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ w.r.t. $\mathbf{M}$ is given by

$$\nabla_M \mathcal{L} = \frac{1}{\sigma^2}(\mathbf{M} \odot (\mathbf{J}\mathbf{B}) - \mathbf{X}\mathbf{B}) + 4\mathbf{M}(\mathrm{Diag}(\mathbf{R}\mathbb{1}) - \mathbf{R}), \tag{4.12}$$

where $\mathbf{B} = \mathbf{K}_{xm} \odot \mathbf{A}^T$, $\mathbf{R} = (\frac{-1}{2\sigma^2})(\mathbf{K}_{mm} \odot \mathbf{A}\mathbf{A}^T)$, Diag is the diagonal operator, $\mathbf{J}$ is a matrix of ones shaped like $\mathbf{B}^T$, $\mathbb{1}$ is a k-by-1 vector of ones, and $\odot$ is the Hadamard product.

Due to the kernel matrices present in Eq.4.12, this equation has no analytical nor a closed-form solution. Thus, we adopt a projected gradient descent (PGD) method (Lin, 2007) to optimize $\mathbf{B}$, which allows us to handle the nonnegativity via projection onto a positive set. Our projected gradient scheme[20] then reads:

$$\mathbf{M}^{t+1} = \mathrm{Proj}(\mathbf{M}^t - \alpha_t \nabla_M \mathcal{L}(\mathbf{M}^t)), \tag{4.13}$$

where, $\mathrm{Proj}(\cdot) = \max(\cdot, 0)$ is the elements-wise proximal operator that ensures the non-negativity of the elements of $\mathbf{M}^{t+1}$, t is the iterations index, α_t is an adaptive step size that satisfies Armijo's condition (Lin, 2007) given by

$$\mathcal{L}(\mathbf{M}^{t+1}) - \mathcal{L}(\mathbf{M}^t) \leq \epsilon \langle \nabla_M \mathcal{L}(\mathbf{M}^t), \mathbf{M}^{t+1} - \mathbf{M}^t \rangle, \tag{4.14}$$

where ϵ is a predefined stopping tolerance. Here, the value of α_t is increased when there is a decrease in the objective function, i.e., $\alpha_t \leftarrow \alpha_t \times \beta$, and decreased otherwise, i.e., $\alpha_t \leftarrow \alpha_t / \beta$ (Lin & Jorge, 1999), where β is a constant with $0 < \beta < 1$. The related algorithm are given in Section S.II of the supplementary material.

4.3.2 Optimize A for fixed M, Y and Z

The second sub-problem in (4.11) is treated as an optimization problem over the Stiefel manifold. To update the abundance matrix $\mathbf{A}$ while being restricted on the Stiefel manifold, we employ a gradient-based optimization approach on the Riemannian space instead of the Euclidian space.

[20] Due to the space limit, the corresponding algorithm is omitted here and is given in Appendix II.

Similar to traditional gradient-based optimization methods on the natural Euclidean space, at each iteration, the algorithm finds the search direction by calculating the natural gradient, i.e., $\nabla_A \mathcal{L}$, of the objective function in the Euclidean space, which in our case is given by

$$\nabla_A \mathcal{L} = \mathbf{K}_{mm} H - \mathbf{K}_{xm}^T + \rho[(\mathbf{\Gamma} \mathbf{A}^T - \mathbf{Y} + \frac{\mathbf{\Lambda}_1}{\rho})^T \mathbf{\Gamma} + (\mathbf{A} - \mathbf{Z} + \frac{\mathbf{\Lambda}_2}{\rho})], \tag{4.15}$$

Then, it calculates the corresponding Riemannian gradient by projecting $\nabla_A \mathcal{L}$ onto the tangent space $T_X \mathcal{M}$ of the manifold space $\mathcal{M}$ via the following operator

$$\text{grad} \mathcal{L}(\mathbf{A}) = \mathcal{P}_{T_X \mathcal{M}}(\nabla_A \mathcal{L}), \tag{4.16}$$

which is defined by

$$\mathcal{P}_{T_X \mathcal{M}}(\nabla_A \mathcal{L}) = (\mathbf{I}_k - \mathbf{A}\mathbf{A}^T) \nabla_A \mathcal{L} + \frac{1}{2} \mathbf{A}(\mathbf{A}^T \nabla_A \mathcal{L} - \nabla_A \mathcal{L}^T \mathbf{A}) \tag{4.17}$$

This projection is mapped back from the tangent plane $T_X \mathcal{M}$ to the manifold $\mathcal{M}$ via a retraction operator $R_{\mathbf{A}}$ as follows

$$\mathbf{A}^{(t+1)} = R_{\mathbf{A}}(-\alpha^{(t)} \text{grad} \mathcal{L}(\mathbf{A}^{(t)})), \tag{4.18}$$

where $\alpha^{(t)}$ is the step length. The operator $R_{\mathbf{A}}$ is defined by

$$R_{\mathbf{A}}(\mathbf{U}) = (\mathbf{A} + \mathbf{U})(\mathbf{I}_k + \mathbf{U}^T \mathbf{U})^{-1/2}, \tag{4.19}$$

where $\mathbf{U}$ satisfies $\mathbf{A}^T \mathbf{U} + \mathbf{U}^T \mathbf{A} = 0$. Detailed explanation about the optimization over the Stiefel manifold is given in (Absil *et al.*, 2009; Rahiche & Cheriet, 2021a).

It is worth noting that in our implementation we employed Pymanopt, an efficient manifold optimization solver proposed in (Townsend *et al.*, 2016).

4.3.3 Optimize Y for fixed M, A and Z

The third sub-problem in (4.11) is a least squares optimization with $L1$-norm regularization. It consists of a differentiable quadratic term and the $L1$-norm term, which is not continuously differentiable. This problem has a closed-from solution that is given by the following soft-thresholding proximal operator

$$\mathrm{Prox}_{\lambda}(\mathbf{D}) = \mathrm{sign}(\mathbf{D}) \odot \max(|\mathbf{D}| - \lambda/\rho, 0), \tag{4.20}$$

where $\mathbf{D} = \mathbf{\Gamma}\mathbf{A}^T + \mathbf{\Lambda}_1/\rho$, $|.|$ is the absolute value operator and max is the element-wise maximum operator.

4.3.4 Optimize Z for fixed M, A and Y

The last sub-problem in (4.11) involves a non-linear term and a quadratic expression. This problem can also be solved using a closed-form solution of the form

$$\mathrm{Prox}_{\iota_{R_+}}(\mathbf{B}) = \max(\mathbf{B}, 0), \quad \text{where } \mathbf{B} = \mathbf{A} + \mathbf{\Lambda}_2/\rho. \tag{4.21}$$

As for the two Lagrange multipliers, i.e., $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, their respective update formulas are given by

$$\begin{aligned} \mathbf{\Lambda}_1^{t+1} :=& \mathbf{\Lambda}_1 + \rho(\mathbf{\Gamma}(\mathbf{A}^{t+1})^T - \mathbf{Y}^{t+1}), \\ \mathbf{\Lambda}_2^{t+1} :=& \mathbf{\Lambda}_2 + \rho(\mathbf{A}^{t+1} - \mathbf{Z}^{t+1}). \end{aligned} \tag{4.22}$$

Arranging the formulas obtained above to allow sequential updates of all the different variables yields the iterative steps summarized in Algorithm 4.1.

4.3.5 Time Complexity

In our implementation, we used some recent libraries to speed up matrix operations involved. We employed the FLANN library (Fast Library for Approximate Nearest Neighbors) (Muja & Lowe, Algorithm 4.1 Pseudo code of GTV-ONNMF

```
1  Input: $\mathbf{X}$, $k$, $\rho$, maxiter
2  Initialization: $M^0$, $A^0$, $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = 0$, $\mathbf{Y} = \mathbf{\Gamma} A^T$, $\mathbf{Z} = A$, $\tau = 1.01$
3  for $t = 1$ to maxiter do
4      Update $\mathbf{M}$ using a PGD algorithm
5      Update $\mathbf{A}$ using Eq.(4.15) via Pymanopt solver (Townsend et al., 2016)
6      Update $\mathbf{Y}$ using Eq.(4.20)
7      Update $\mathbf{Z}$ using Eq.(4.21)
8      Update $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ using Eq.(4.22)
9      Update $\rho = \tau \times \rho$
10     if stopping condition is met then
11         break.
12     end if
13 end for
14 return $\mathbf{M}$, $\mathbf{A}$
```

2014) to calculate the weight matrix W, whose cost is $O(n^2 \log n)$. This matrix is calcultaed one time. Using the *KeOps* toolbox (Kernel Operations on the GPU) (Charlier, Feydy, Glaunès, Collin & Durif, 2021), the largest kernel matrice, which is $\mathbf{K}_{xm}$, costs $O(bn)$ without memory overflows (Feydy, Glaunès, Charlier & Bronstein, 2020). The overall per-iteration complexity of Algorithm 4.1 is $O(t_2|\mathcal{E}|kn)$, where t_2 is the number of iterations required for step 3. Note that $|\mathcal{E}| = $ nbr of neighbors $\times\, n$ and $k \ll n$. Thus, the complexity can be expressed by $O(t_2 n^2)$.

4.4 Experimental results

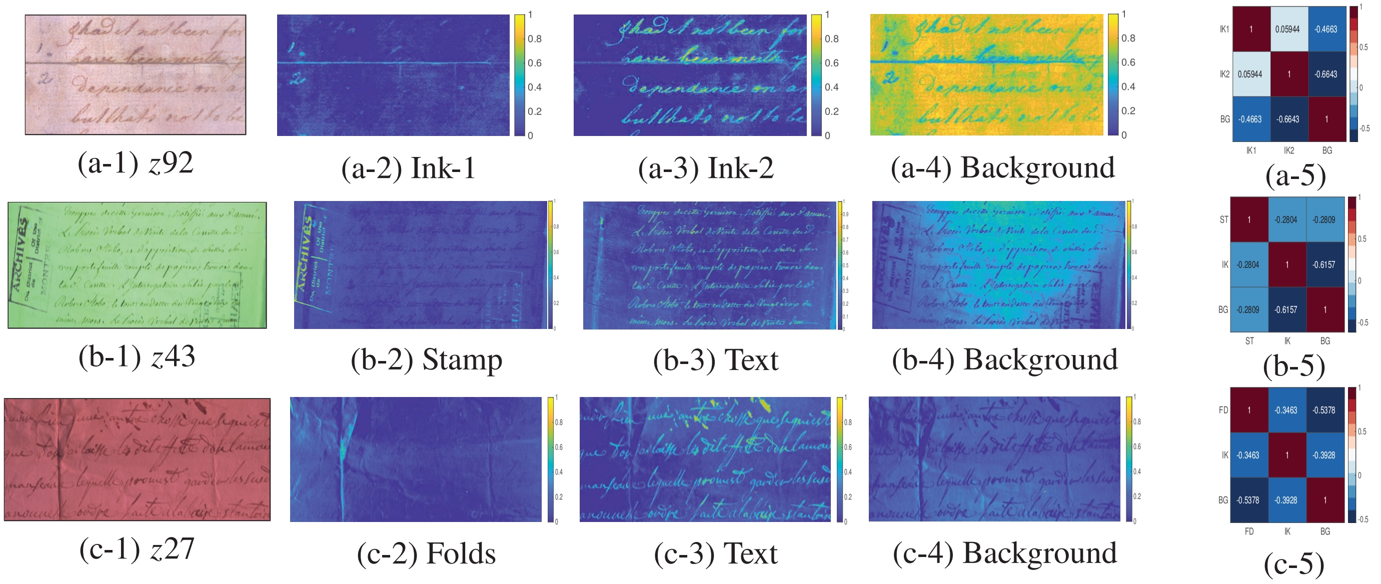
To assess the effectiveness of our model, we addressed the task of blind source separation in real MS images. Specifically, we validated our model on a collection of MS images of 10 real ancient hand-written documents, called MSTEx-2, collected by Synchromedia Lab[21] (Hedjam *et al.*, 2015). The experiments were divided into two practical use-cases: full source separation and text extraction from MS images.

[21] available at http://www.synchromedia.ca/databases/MSI-HISTODOC

In our experiments, an eight-neighborhood graph is constructed for each MS data matrix, where each pixel p_i represents a node and the spectral bands represent the features. The precision parameter γ is selected in the range of $\{10^3, \cdots, 10^{-3}\}$ using a grid search method. As for our model hyper-parameters, the best parameters were specified for each sample due to the variability of the content of MS images and their spectral signatures. The number of components k (rank of factorization) is set manually between 3 and 6 depending on the number of materials. The model is initialized (i.e., **M, A**) using an SVD decomposition. The remaining parameters, i.e., ρ, λ, and the variance parameter σ of the kernel RBF are tuned using a grid search method. We set $\rho \in \{10^{-1}, \cdots, 10^{-4}\}$, $\lambda \in \{10^{-2}, \cdots, 10^{-4}\}$, and $\sigma \in \{10^{2}, \cdots, 10^{-4}\}$.

4.4.1 Blind source separation of MS document images

In this experiment, we performed a complete MS document images decomposition into their constituting components. An illustration is given in Fig. 4.1 to show some obtained results. Besides, the separation quality is quantitatively measured using the Pearson Correlation Coefficient (PCC). Thus, we calculated the correlation between the obtained sources. The results show that, even with the presence of overlapped objects in the three samples, the proposed approach was able to separate between them efficiently. The corresponding PCC matrices show negative coefficients between different components. For instance, up to -0.66 over -1 between the text and background components was obtained with the sample $z92$, demonstrating our approach's effectiveness and generalization power.



(a-1) *z*92 (a-2) Ink-1 (a-3) Ink-2 (a-4) Background (a-5)

(b-1) *z*43 (b-2) Stamp (b-3) Text (b-4) Background (b-5)

(c-1) *z*27 (c-2) Folds (c-3) Text (c-4) Background (c-5)

Figure 4.1 Qualitative illustrations of sources extracted from samples of the MSTEx-2 dataset using GTV-ONNMF. The right column shows the heat maps of the corresponding CCP matrices of the extracted sources, where IK, BG, ST, and FD denote ink, background, stamp, and folds, respectively

4.4.2 Text extraction from MS document images

In this experiment, we show how the proposed model can be used effectively to improve text extraction from MS document images. This task is quite important for subsequent processing of document images, such as optical character recognition (OCR). To compare the results of our model with the state-of-the-art, we consider the task of text binarization in MS document images. Clearly, text extraction quality influences the binarization quality, i.e., a good separation of the text from other components yields a good binarization and vice-versa. To this end, we added a post-processing step to convert the abundance maps obtained by our model to binary maps. In our experiments, we adopted Howe's method (Howe, 2013), one of the best image binarization techniques. Otherwise, any other technique can be used. For the sake of comparative performance evaluation, we compared our model against six approaches. As a baseline, we considered the two conventional binarization methods Howe's (Howe, 2013) and Sauvola's (Howe, 2013); two other approaches devoted to MS images, namely, SKKHM[22] (Li *et al.*, 2007) and GMM (Hollaus *et al.*, 2018b); and two factorization approaches, namely, MA-ONTF (Rahiche & Cheriet, 2021a) and KONMF (Rahiche & Cheriet, 2021b). To assess the binarization quality, we adopted the F-measure (FM), distance reciprocal distortion (DRD), negative rate metric (NRM), and peak signal-to-noise ratio (PSNR) metrics that are widely used in the literature of text binarization.

Table 4.1 Average scores in the binarization of the MS-TEx-2 dataset and comparison with state-of-the-art methods

Method	FM	DRD	NRM	PSNR
Howe (Howe, 2013)	70.41	8.58	12.06	-
Sauvola (Sauvola & Pietikäinen, 2000)	60.92	8.21	24.01	-
SKKHM (Li *et al.*, 2007)	62.68	17.16	16.33	13.29
GMM (Hollaus *et al.*, 2018b)	82.95	4.16	8.47	16.26
MA-ONTF (Rahiche & Cheriet, 2021a)	83.86	3.72	8.92	-
KONMF (Rahiche & Cheriet, 2021b)	84.73	3.76	**6.51**	**17.42**
GTV-ONNMF (proposed)	**85.69**	**3.24**	7.03	17.41

[22] Code obtained at http://utopia.duth.gr/~nmitiano/

Table 4.1 shows that the proposed model GTV-ONNMF and KONMF achieved the best scores in terms of all metrics. The proposed GTV-ONNMF outperformed all the methods in terms of FM and DRD metrics and achieved a comparable score in terms of PSNR. These results (see also Fig. II-1 of Appendix II) shows the considerable improvement gained by incorporating the spectral non-linearity, the intrinsic structure of data, and the spatial orthogonality.

4.5 Conclusion

In this letter, we have presented an effective GTV-ONNMF-based framework for the blind decomposition of MS document images. Our GTV-ONNMF model does not suffer from the scalability issue. Using efficient kernel operations toolboxes allowed us to process larges data images without memory overflows. Based only on the spectral signatures of writing materials, our model shows efficient decomposition performance while outperforming conventional and some state-of-the-art factorization techniques.

CHAPTER 5

VARIATIONAL BAYESIAN ORTHOGONAL NONNEGATIVE MATRIX FACTORIZATION OVER THE STIEFEL MANIFOLD

Abderrahmane Rahiche[a] , Mohamed Cheriet[a]

[a] Department of Systems Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article submitted to IEEE Transaction on Image Processing, December, 2021.

Abstract

Nonnegative matrix factorization (NMF) is one of the best-known multivariate data analysis techniques. Imposing an additional orthogonality constraint on one of its latent factors provides sparser part-based representations. This property has demonstrated remarkable performance in clustering and source separation tasks. However, existing orthogonal NMF algorithms rely mainly on non-probabilistic frameworks that ignore the noise inherent in real-life data and lack variable uncertainties. Thus, in this work, we investigate a new probabilistic formulation of orthogonal NMF (ONMF). In the proposed model, we impose the orthogonality through a directional prior distribution defined on the Stiefel manifold called von Mises-Fisher distribution. This manifold consists of a set of directions that comply with the orthogonality condition that arises in many applications. Moreover, our model involves an automatic relevance determination (ARD) prior to inferring the factorization rank. We devise an efficient variational Bayesian inference algorithm to solve the proposed ONMF model, which allows fast processing of large datasets. We evaluated the proposed model, called VBONMF, on the task of blind decomposition of real-world multispectral images of ancient documents. The numerical experiments demonstrate its efficiency and competitiveness compared to the state-of-the-art approaches.

Keywords

Blind source separation, Multispectral images, Orthogonal nonnegative matrix factorization, Stiefel manifold, Variational Bayes, Von Mises-Fisher distribution.

5.1 Introduction

Nonnegative matrix factorization (NMF) is one of the powerful data decomposition tools. From data clustering to finding latent features in datasets, NMF produces a parts-based representation of input data and makes the interpretation of the outcomes easier without further processing. Formally, NMF seeks to factorize a nonnegative data matrix, $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ ($\mathbb{R}_+$ denotes the set of nonnegative reals), into the product of two nonnegative low-rank factors, a basis matrix, $\mathbf{W} \in \mathbb{R}_+^{m\times k}$, and a coefficient matrix, $\mathbf{H} \in \mathbb{R}_+^{k\times n}$, so that $\mathbf{X} \approx \mathbf{WH}$, where the rank $k \ll \min(m, n)$ is an unknown integer.

To assess the quality of factorization, several cost functions can be used. The Frobenius norm of the difference between the data and its approximation is a common choice in the literature to measure the reconstruction error. The problem is then formulated as an optimization problem of the form

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2, \quad \text{such that } \mathbf{W},\ \mathbf{H} \geq 0 \tag{5.1}$$

Unlike unconstrained matrix factorization approaches, such as singular values decomposition, that can be solved exactly, NMF is an NP-hard (Vavasis, 2010) problem. This makes it difficult to obtain a global minimum point to the corresponding bi-convex objective function (i.e., Eq.5.1). Moreover, because of the ill-posed nature of NMF, the solution to problem (5.1) is not unique (Laurberg, 2007; Huang *et al.*, 2013). Indeed, it is easy to see that it is possible to obtain an equivalent solution, in terms of another pair of factors $\tilde{\mathbf{W}} = \mathbf{WQ}$ and $\tilde{\mathbf{H}} = \mathbf{Q}^{-1}\mathbf{H}$, to problem in (5.1) such that $\mathbf{X} \approx \tilde{\mathbf{W}}\tilde{\mathbf{H}} = (\mathbf{WQ})(\mathbf{Q}^{-1}\mathbf{H}) = \mathbf{W}(\mathbf{QQ}^{-1})\mathbf{H} = \mathbf{WH}$, where $\mathbf{Q}$ is an invertible $k \times k$ positive matrix. In order to remove ambiguities related to the choice of $\mathbf{Q}$ and reduce the volume of solution space, additional constraints can be imposed on the latent factors $\mathbf{W}$ and $\mathbf{H}$

such as the sparsity (Hoyer, 2004), smoothness (Yokota *et al.*, 2015), and orthogonality (Choi, 2008).

Orthogonal NMF (ONMF) is an extension of the standard NMF in which the orthogonality condition is imposed on one of the two factor matrices, i.e, considering $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{k\times k}$ or $\mathbf{H}\mathbf{H}^T = \mathbf{I}_{k\times k}$, or on both of them (Ding *et al.*, 2006; Rahiche & Cheriet, 2021a). ONMF has demonstrated its efficiency for clustering applications and source separation tasks. In most existing ONMF models, the orthogonality constraint is generally added as a penalty term to the objective function (Ding *et al.*, 2006). By doing so, the algorithm is optimized in hopes of generating a solution that respects the orthogonality condition. Alternatively, the orthogonality can be imposed as a part of the optimization problem over a set of orthogonal matrices called the Stiefel manifold (Rahiche & Cheriet, 2021a), without resorting to additional penalty terms. In this strategy, the orthogonality is verified at each iteration until the algorithm's convergence, and it has been shown to be more effective than other strategies.

Existing ONMF models rely mostly on deterministic formulations that do not take the noise inherent in real-life data into account. This issue is usually addressed through a probabilistic reformulation of problem (5.1). Probabilistic NMF models allow accounting for the uncertainty of noise inherent in the datasets (Schmidt, Winther & Hansen, 2009), called *aleatoric* uncertainty (Kendall & Gal, 2017), and the *epistemic* uncertainty (Kendall & Gal, 2017) of the model parameters. Moreover, the model order selection (rank) is another open problem in deterministic NMF. However, this issue can be addressed in probabilistic NMF using dedicated techniques such as the automatic relevance determination (ARD) (Tan & Févotte, 2012) and minimum description length (Squires, Prügel-Bennett & Niranjan, 2017) principle. Nevertheless, despite the benefits of probabilistic settings over deterministic formulations, almost all existing ONMF models are non-probabilistic.

In this work, we propose a new variational Bayesian orthogonal NMF model that addresses the uniqueness issue by imposing orthogonality on one factor. We refer to our proposed model as VBONMF. The proposed model accounts for nonnegativity and orthogonality simultaneously.

The latter is incorporated as a uniform distribution over a set of orthogonal matrices called the Stiefel manifold (Chikuse, 1990), which yields a posterior distribution called *von Fisher-Misses* distribution (Khatri & Mardia, 1977) or also *Matrix Langevin* distribution. The Bayesian setting allows us to consider the noise and include an additional prior to control the relevance of the latent components via the ARD principle. For the sake of computation efficiency, we devised a variational Bayesian algorithm to solve the proposed model.

The specific contributions of this paper are summarized as follows:

1) A new Bayesian orthogonal NMF model over the Stiefel manifold.
2) A variational Bayesian-based inference algorithm to approximate the parameters of the proposed generative model.
3) Application of the proposed variational Bayesian orthogonal NMF for the task of multispectral (MS) document images decomposition.
4) An extensive comparison between the proposed VBONMF model and some existing Bayesian NMF models.
5) An extensive comparison between the proposed VBONMF model and some existing deterministic orthogonal NMFs over the Stiefel manifold. To the best of our knowledge such a comparison has not been done before in previous studies.

The remaining parts of this paper are organized as follows. Section 5.2 gives a brief overview of existing works on this topic. Section 5.3 describes the proposed model and the corresponding inference solution. Section 5.5 presents the validations performed and numerical results obtained. Section 5.6 discusses the results and concludes the paper.

5.2 Related work

In contrast to deterministic NMF models, the probabilistic framework treats the two latent factor matrices, i.e., the basis matrix $\mathbf{W}$ and the coefficient matrix $\mathbf{H}$, as two independent random variables. It places prior distributions over them that act as regularization terms. The loss function in (5.1) is also replaced by a suitable likelihood function $p(\mathbf{X}|\mathbf{W},\mathbf{H})$. The problem then becomes a Bayesian inference problem that aims to calculate the posterior $p(\mathbf{W}, \mathbf{H}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{W}, \mathbf{H})p(\mathbf{W})p(\mathbf{H})$.

The design of Bayesian NMF models is based on three main ingredients: 1) the likelihood function, 2) the prior distributions, and 3) the inference technique used to estimate the model's parameters. The best choice of each component depends on the nature of the data and the targeted application. A non-appropriate choice of these components can affect the model's performance and influence its behaviour. A good review of existing Bayesian matrix factorization methods and the trade-offs between likelihood and choices of priors is given in (Brouwer & Lió, 2017).

As for the likelihood function, assigning a normal function to it is a common choice in the literature on Bayesian matrix factorization to capture noise existing in data (Moussaoui, Brie, Mohammad-Djafari & Carteret, 2006; Salakhutdinov & Mnih, 2008; Schmidt *et al.*, 2009; Schmidt & Mohamed, 2009; Tichỳ & Šmídl, 2015; Brouwer & Lió, 2017). To ensure the nonnegativity of low-rank factors, the choice is, generally, limited to the exponential distribution (Schmidt *et al.*, 2009; Brouwer, Frellsen & Lio, 2016), the gamma distribution (Moussaoui *et al.*, 2006), or a distribution derived from a normal distribution, such as the truncated normal distribution (Hinrich & Mørup, 2018; Tichỳ, Bódiová & Šmídl, 2019), or the half-normal distribution (Psorakis, Roberts, Ebden & Sheldon, 2011; Tan & Févotte, 2012).

The main existing Bayesian NMF models are summarized in Table 5.1.

Table 5.1 Overview of main existing Bayesian NMF models

Model	Likelihood	Basis prior	Coefficient prior	Noise prior	Rank selection	Inference	Application
BPSS (Moussaoui *et al.*, 2006)	$\mathcal{N}(\mathbf{X}_{ij}; (\mathbf{AS})_{ij}, \sigma^2)$	$\mathcal{G}(\mathbf{A}; \alpha, \beta)$	$\mathcal{G}(\mathbf{S}; \gamma, \lambda)$	-	-	MCMC	Chemical spectroscopy
BNMF (Schmidt & Mohamed, 2009)	$\mathcal{N}(\mathbf{X}_{ij}; (\mathbf{AB})_{ij}, \sigma^2)$	$\mathcal{E}(\mathbf{A}; \alpha)$	$\mathcal{E}(\mathbf{B}; \beta)$	$\sigma \sim \mathcal{G}^{-1}(k, \theta)$	BIC	Gibbs	UMIST Face
BNMF-ARD (Brouwer, Frellsen & Lió, 2017)	$\mathcal{N}(\mathbf{X}_{ij}; (\mathbf{UV})_{ij}; \tau^{-1})$	$\mathcal{E}(\mathbf{U}; \lambda_k)$	$\mathcal{E}(\mathbf{V}; \lambda_k)$	$\tau \sim \mathcal{G}(\alpha_\tau, \beta_\tau)$	ARD, $\lambda_k \sim \mathcal{G}(\alpha_\lambda, \beta_\lambda)$	VB	Drug sensitivity
NMF-APC (Tichý *et al.*, 2019)	$\mathcal{N}(\mathbf{D}_{ij}; (\mathbf{AX})_{ij}, w^{-1})$	$\mathcal{T}\mathcal{N}(\mathbf{A}; 0, \xi_k^{-1})$	$\mathcal{T}\mathcal{N}(\mathbf{X}; 0, \Sigma_k^{-1})$	$w \sim \mathcal{G}(\vartheta_0, \rho_0)$	ARD, $\xi_k \sim \mathcal{G}(\alpha_\xi, \beta_\xi)$	VB	Dynamic renal scintigraphy
psNMF (Hinrich & Mørup, 2018)	$\mathcal{N}(\mathbf{X}_{ij}; (\mathbf{WH})_{ij}, \tau^{-1})$	$\mathcal{T}\mathcal{N}(\mathbf{W}; 0, \lambda_d^{-1}, 0, \infty)$	$\mathcal{T}\mathcal{N}(\mathbf{X}; 0, \lambda_d^{-1}, 0, \infty)$	$\tau \sim \mathcal{G}(\alpha_\tau, \beta_\tau)$	ARD, $\lambda_d \sim \mathcal{G}(\alpha_\lambda, \beta_\lambda)$	VB	Swimmer, CBCL Face, MNIST

In (Moussaoui *et al.*, 2006), the authors proposed a probabilistic source separation NMF (BPSS) to handle the task of source separation of the spectroscopic signal of chemical substances. Only one-dimensional signals are considered in that study. Also, neither the noise nor the rank selection is addressed in it. Both (Schmidt & Mohamed, 2009) and (Brouwer *et al.*, 2017) considered the nonnegativity property solely. While in (Schmidt & Mohamed, 2009) the challenge of automatic feature extraction from human faces is addressed, the model proposed in (Brouwer *et al.*, 2017) is applied on a drug sensitivity dataset. In (Tichỳ *et al.*, 2019; Hinrich & Mørup, 2018), in addition to the nonnegativity, the sparsity is enforced by placing an ARD prior on each element of the factors (**W**, **H**).

As for the orthogonality constraint in Bayesian frameworks, it has been addressed only in Bayesian variants of singular value decomposition (Hoff, 2007) and principal component analysis (PCA) models (Šmídl & Quinn, 2007; Dobigeon & Tourneret, 2010) by assigning a uniform distribution on the Stiefel manifold. These models did not involve the nonnegativity and were designed mainly to work with real-valued data, which makes them not suitable to handle nonnegative data as demonstrated in many studies (Lee & Seung, 1999). Recently, the orthogonality property has gained more importance in statistical applications. For instance, in (Lin, Rao & Dunson, 2017; Pal, Sengupta, Mitra & Banerjee, 2020; Jauch, Hoff & Dunson, 2021), statistical analyses of orthogonal matrices on the Stiefel manifold have been performed. Nevertheless, these models did not consider the problem of data factorization nor the nonnegativity condition.

In summary, despite its widespread use in non-probabilistic NMF models, to the best of our knowledge, imposing orthogonality over the Stiefel manifold within a Bayesian NMF framework has not been studied before in any existing model. Moreover, a comparative study between deterministic orthogonal NMF and probabilistic orthogonal NMF frameworks is still lacking.

5.3 Bayesian Orthogonal NMF framework

For applications like multichannel image decomposition, spectral image data are usually modelled as a linear mixture model (LMM) (Settle & Drake, 1993) of the form

$$\mathbf{X} = \mathbf{M}\mathbf{A}^T + \mathbf{E}, \tag{5.2}$$

where $\mathbf{X} \in \mathbb{R}^{b \times n}$ is the matrix of observed spectral images, $\mathbf{M} \in \mathbb{R}^{b \times k}$ is the endmembers matrix, $\mathbf{A} \in \mathbb{R}^{n \times k}$ is the abundance matrix, $\mathbf{E} \in \mathbb{R}^{b \times n}$ represents the noise matrix, b denotes the number of observed channels, n is the number of pixel in each image, and k is the number of mixed sources.

When tackling this problem via a non-probabilistic NMF, the noise is generally ignored, i.e. assuming that $\mathbf{X} \approx \mathbf{M}\mathbf{A}^T$, which makes the model less robust to noise. Therefore, in this paper, we propose a probabilistic NMF formulation that allows incorporating a prior knowledge about the noise.

The graphical model representing our VBONMF is shown in Fig.5.1.

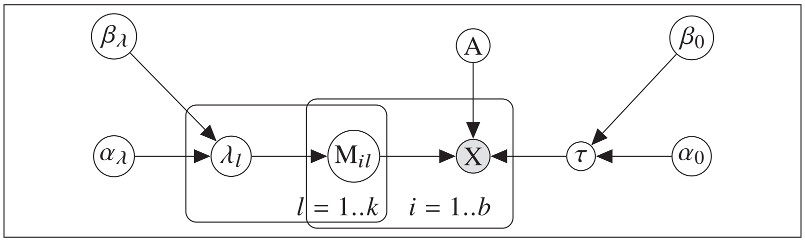


Figure 5.1 Graphical model of the proposed VBONMF.

The proposed generative model can be expressed by the joint distribution of all the random variables as

$$p(\mathbf{X}, \mathbf{M}, \mathbf{A}, \boldsymbol{\lambda}, \tau) = p(\mathbf{X}|\mathbf{M}, \mathbf{A}, \tau)p(\mathbf{A})p(\mathbf{M}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})p(\tau) \tag{5.3}$$

Next, we describe the likelihood function and the prior distributions placed over the model parameters $\mathbf{M}$, $\mathbf{A}$, τ, and λ.

5.3.1 Likelihood

For our model, the normal function is adopted as a likelihood function to replace the objective in equation (5.1). Moreover, by assuming that the residuals in (5.2), $\mathbf{E}_{i,j}$, are independent and identically distributed (i.i.d.) from a zero mean normal with precision $\tau = 1/\sigma^2$, the likelihood takes the following form

$$p(\mathbf{X}|\mathbf{M}, \mathbf{A}) \propto \prod_{i=1}^{b} \prod_{j=1}^{n} \mathcal{N}(\mathbf{X}_{ij}, (\mathbf{M}\mathbf{A}^T)_{ij}, \tau^{-1}), \tag{5.4}$$

In addition to the additive noise property, this function allows controlling the noise parameter and using conjugate priors that are easier for calculating a closed analytical form for the posterior. A natural choice of the appropriate conjugate prior for the noise precision is the gamma distribution, which has the formula

$$\tau \sim \mathcal{G}(\tau|\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tau^{(\alpha_0 - 1)} e^{(-\beta_0 \tau)} \tag{5.5}$$

where $\Gamma(x) = \int_0^\infty x^{t-1} e^{-x}$ is the gamma function, and $\alpha_0, \beta_0 > 0$ are the corresponding shape and scale parameters, respectively.

5.3.2 Prior distributions

The VBONMF model assumes that the columns of the coefficient (abundance) matrix $\mathbf{A}$ form an orthogonal set. This corresponds to the assumption that the coefficient columns are independent. Accordingly, we incorporate this condition by placing a uniform prior distribution that is defined on a feasible set of orthogonal axes, called Stiefel manifold[23], of the form

$$f(\mathbf{A}) = Vol^{-1}(\mathcal{V}_{n,k}) \mathbb{I}_{\mathcal{V}_{n,k}}(\mathbf{A}), \tag{5.6}$$

[23] Named in honor of Eduard L. Stiefel.

where $\mathcal{V}_{n,k} \triangleq \{\mathbf{A} \in \mathbb{R}^{n\times k} | \mathbf{A}^T\mathbf{A} = \mathbf{I}_k\}$ is the set of orthonormal $n \times k$ matrices called the Stiefel manifold, whose dimension is $dim(\mathcal{V}_{n,k}) = nk - \frac{1}{2}k(k+1)$, and $\mathbb{I}_{\mathcal{V}_{n,k}}(.)$ is an indicator function of the set $\mathcal{V}_{n,k}$, that is defined as

$$\mathbb{I}_{\mathcal{V}_{n,r}}(\mathbf{A}) = \begin{cases} 1 & \text{if} \quad \mathbf{A} \in \mathcal{V}_{n,k}, \\ 0 & \text{otherwise,} \end{cases} \tag{5.7}$$

Vol is the volume (or the surface area) of the Stiefel manifold, which is given by (Muirhead, 1982; Chikuse, 1990; Gupta & Nagar, 1999)

$$Vol(\mathcal{V}_{n,k}) = \frac{2^k \pi^{nk/2}}{\Gamma_k(n/2)}, \qquad k < n, \tag{5.8}$$

$\Gamma_k(n/2) = \pi^{k(k-1)/4} \prod_{i=1}^{k} \Gamma[\frac{n}{2} - \frac{i-1}{2}]$.

To ensure the nonnegativity of the elements of the basis matrix (called also endmembers matrix in the field of remote sensing). The prior distributions over the latent variable $\mathbf{M}$ is assumed to be an exponential function, which has the form

$$\mathcal{E}(\mathbf{M}|\lambda) = \lambda \exp(-\lambda \mathbf{M}) u(\mathbf{M}), \tag{5.9}$$

where λ is the shape parameter and $u(\mathbf{M})$ is the unit step function defined as

$$u(\mathbf{M}) = \begin{cases} 1 & \text{if} \quad \mathbf{M} \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5.10}$$

Moreover, in our VBONMF model, we consider the rank factorization as an unknown parameter. Thus, we place a further Gamma prior over the column-dependent relevance weights

$$\lambda_l \sim \mathcal{G}(\lambda_l | \alpha_\lambda, \beta_\lambda) \tag{5.11}$$

The idea behind ARD is to control the relevance of each column of $\mathbf{M}$ on the basis of the value of the corresponding λ_l, where a close to zero value of λ_l yields turning-off the corresponding l column and vice-versa.

Note that it is also possible to place other conjugate priors over the parameters α_0, β_0, α_λ, and β_λ. However, in our case, we stop at this level of hierarchy, and we set them manually.

Table 5.2 Approximate posteriors distributions and their shaping parameters

$q(\mathbf{M}_{il}) \sim \mathcal{TN}(\mathbf{M}_{il}\|\mu_{il}, \tau_{il})$	$\mu_{il} = \frac{1}{\tau_{il}}[-\mathbb{E}[\lambda_l] + \mathbb{E}[\tau] \sum_{j=1}^{n} (\mathbf{X}_{ij} - \sum_{l' \neq l}^{k} \mathbb{E}[\mathbf{M}_{il'}]\mathbb{E}[\mathbf{A}_{jl'}])\mathbb{E}[\mathbf{A}_{jl}]]$	$\tau_{il} = \mathbb{E}[\tau] \sum_{j=1}^{n} \mathbb{E}[\mathbf{A}_{il}^2]$
$q(\mathbf{A}) \sim \mathcal{VMF}(\mathbf{Z}_A)$	$\mathbf{Z}_A = \mathbb{E}[\tau]\mathbf{X}^T\mathbb{E}[\mathbf{M}]$	-
$q(\tau) \sim \mathcal{G}(\tau\|\alpha_\tau, \beta_\tau)$	$\alpha_\tau = \alpha_0 + \frac{b \times n}{2}$	$\beta_\tau = \beta_0 + \frac{1}{2}\mathbb{E}[\|\mathbf{X} - \mathbf{MA}\|^2]$
$q(\lambda_l) \sim \mathcal{G}(\lambda_l\|\alpha_\lambda, \beta_\lambda)$	$\alpha_\lambda = \alpha_{\lambda 0} + b$	$\beta_\lambda = \beta_{\lambda 0} + \sum_{i=1}^{b} \mathbb{E}[\mathbf{M}_{il}]$

5.4 Variational Bayes optimization

For ease of notation, we set $\theta = \{\mathbf{M}, \mathbf{A}, \tau, \lambda\}$ to denote the set of the unobserved latent variables and parameters of the model. The exact Bayesian inference of the posterior distribution $p(\theta|\mathbf{X}) = \frac{p(\mathbf{X},\theta)}{\int p(\mathbf{X},\theta)d\theta}$ is intractable. Since sampling methods such as Gibbs sampling used in (Dobigeon & Tourneret, 2010) are slow and computationally demanding, we adopt the mean-field variational Bayes method to perform the inference by approximating the true posterior distribution with a simpler variational distribution $q(\theta|\mathbf{X}) \approx p(\theta|\mathbf{X})$. Specifically, we seek an optimal variational distribution that minimizes the reversed Kullback-Leibler (KL) divergence defined as $\mathcal{D}_{\mathrm{KL}}(q(\theta)\|p(\theta|\mathbf{X})) = \mathbb{E}_\theta[\frac{q(\theta)}{p(\theta|\mathbf{X})}]$. Furthermore, we assume that for $q(\theta|\mathbf{X})$ all variables are independent, and the variational distribution factorizes completely, i.e.,

$$p(\theta|\mathbf{X}) \approx q(\theta) = \prod_{\theta_i \in \theta} q(\theta_i). \tag{5.12}$$

For each one of the variational variables, the optimal distribution q_{θ_i} that minimizes the KL-divergence can be expressed as

$$\ln q(\theta_i) \propto \mathbb{E}_{\theta \backslash \theta_i}[\ln p(\mathbf{X}, \theta)] + const. \tag{5.13}$$

The operator $\mathbb{E}_{\theta \backslash \theta_i}$ means expectation with respect to all variables in θ, except for the current θ_i.

Proposition 3. *Given the prior distribution defined in (5.6) and the normal likelihood function in (5.4), the corresponding variational posterior distribution of the latent factor* **A** *has the form of a von Mises-Fisher distribution ($\mathcal{VMF}$), that is,*

$$ln\, q(A) \propto \mathcal{VMF}(\mathbf{Z}), \tag{5.14}$$

which has the form

$$\mathcal{VMF}(\mathbf{A}|\mathbf{Z}) = \frac{exp^{Tr(\mathbf{Z}^T\mathbf{A})}}{{}_0F_1\left(\frac{1}{2}n, \frac{1}{4}\mathbf{Z}^T\mathbf{Z}\right) Vol(\mathcal{V}_{n,k})}, \tag{5.15}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times k}$ is a random variable and ${}_0F_1\left(\frac{1}{2}n, \frac{1}{4}\mathbf{Z}^T\mathbf{Z}\right)$ is the hypergeometric function of matrix argument $\mathbf{Z}^T\mathbf{Z}$.

Proof. Considering the Eq.(5.13), and by substituting the formula of the joint distribution (5.3) into (5.13) and excluding terms that do not depend on **A**, we can write

$$\begin{aligned} \ln q(\mathrm{A}) &\propto \mathbb{E}_{\theta \backslash \mathbf{A}}[\ln p(\mathbf{X}|\mathbf{M}, \mathbf{A}, \tau) + \ln p(\mathbf{A})] + cst. \\ &\propto \mathbb{E}_{\theta \backslash \mathbf{A}}\left[\tau \mathrm{Tr}(\mathbf{X}^T\mathbf{M}\mathbf{A}^T)\right] + cst. \\ &\propto \mathrm{Tr}\left(\mathbb{E}[\tau]\mathbf{X}^T\mathbb{E}[\mathbf{M}]\mathbf{A}^T\right) \\ &\propto \mathrm{Tr}\left(\mathbf{Z}\mathbf{A}^T\right) = \mathrm{Tr}\left(\mathbf{Z}^T\mathbf{A}\right) \end{aligned} \tag{5.16}$$

where $\mathbf{Z} = \mathbb{E}[\tau]\mathbf{X}^T\mathbb{E}[\mathbf{M}]$. We observe that this expression consists of the trace (Tr) of the product of **A** and another matrix term. Exponentiating both sides, we can identify $q(\mathbf{A})$ as a von Mises-Fisher distribution ($\mathcal{VMF}$). Q.E.D.

The derivation of $q(\mathbf{M})$, $q(\tau)$, and $q(\boldsymbol{\lambda})$ is similar to the above. We omit the details for the sake of space. The formulas of the posterior distributions and their corresponding parameters are summarized in Table 5.2.

It is worth noting that with either an exponential, truncated normal, or a half-normal prior distribution, the resulted variational distribution, i.e., $q(\mathbf{M})$ here, is always a truncated normal. The choice between the three prior distributions affects only the expressions of the corresponding shaping parameters μ and τ of the resulting truncated normal distribution.

The moments of the different variational distributions involved in the inference process of VBONMF are: $\mathbb{E}[\tau], \mathbb{E}[\lambda_l], \mathbb{E}[\mathbf{M}]$, and $\mathbb{E}[\mathbf{A}]$. The expression of each variable is given as follows:

5.4.1 Moments of gamma distribution

As mentioned above, we have found that both τ and λ both follow a gamma distribution. Thus, it is easy to show that the first moment (expectation) has the form

$$\mathbb{E}[\tau] = \frac{\alpha_\tau}{\beta_\tau}, \tag{5.17}$$

and

$$\mathbb{E}[\lambda_l] = \frac{\alpha_\lambda}{\beta_\lambda}. \tag{5.18}$$

5.4.2 Moments of truncated normal distribution

The expectation of $\mathbf{M}$, that follows a truncated normal distribution, is given by

$$\mathbb{E}[\mathbf{M}] = \mu_m + \frac{1}{\sqrt{\tau_m}} g(-\mu_m \sqrt{\tau_m}). \tag{5.19}$$

The variance (second moment) has the following form:

$$\mathrm{Var}[\mathbf{M}] = \frac{1}{\tau_m} \left[1 - h(-\mu_m \sqrt{\tau_m}) \right], \tag{5.20}$$

where:

$$\begin{aligned} g(x) &= \frac{\Phi(x)}{1-\Phi(x)} \\ h(x) &= g(x)\left[g(x)-x\right] \\ \Phi(x) &= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right) \end{aligned} \tag{5.21}$$

5.4.3 Moments of Von Mises-Fisher distribution

Because of the hypergeometric function involved in the expression of the $\mathcal{VMF}$ distribution (Eq.5.14), calculating its first two moments is not trivial. Hence, an approximation method is adopted to derive the required moments.

Proposition 4. *Let* **Z** *be a von Mises-Fisher random matrix, that is,*

$$\boldsymbol{Z} \sim \mathcal{VMF}(\boldsymbol{F}),$$

and $\boldsymbol{F}$ *can be expressed as*

$$\boldsymbol{F} = \boldsymbol{BDY}, \tag{5.22}$$

where $\boldsymbol{B}$ *and* **Y** *are two orthogonal matrices, i.e.,* $\boldsymbol{BB}^T = I_k$ *and* $\boldsymbol{YY}^T = I_k$*. Then, the moment of* $\boldsymbol{F}$ *can be expressed as* $\mathbb{E}[\boldsymbol{F}] = \boldsymbol{B}\mathbb{E}[\boldsymbol{D}]Y$.

Proof. Using singular values decomposition (SVD) of **F**, we can write

$$\mathbf{F} = \mathbf{USV}^T, \tag{5.23}$$

where $\mathbf{S} = \mathrm{diag}(s_{11}, s_{22}, \cdots, s_{kk})$,**U** and **V** are two orthogonal matrices.

Following the results shown in (Khatri & Mardia, 1977), it has been shown that $\mathbb{E}[\mathbf{F}_{ij}] = \mathbb{E}[\mathbf{S}_{ij}]$ and $\mathbb{E}[\mathbf{F}_{ij}\mathbf{F}_{i'j'}] = \mathbb{E}[\mathbf{S}_{ij}\mathbf{S}_{i'j'}]$. According to these results, the first moment of **F** can be expressed by

$$\mathbb{E}[\mathbf{F}] = \mathbf{UG}(\mathbf{S})\mathbf{V}^T. \tag{5.24}$$

and the second moment by

$$\mathrm{Var}[\mathbf{F}] = \mathbf{U}\mathbf{H}(\mathbf{S})\mathbf{V}^T, \tag{5.25}$$

Using the moment generating function (MGF) of the variable **F**, defined by

$$M_{\mathbf{F}}(\mathbf{T}) = \mathbb{E}[\exp{(\mathbf{T}\mathbf{F}^T)}], \tag{5.26}$$

it is possible to find the expressions of $\mathbf{G}(\mathbf{S})$ and $\mathbf{H}(\mathbf{S})$ by taking the first and second derivatives of Eq. 5.26 w.r.t. **T**, respectively, which yield

$$\begin{aligned} \mathbf{G}(\mathbf{S}) &= \mathrm{diag}(\Psi_1, \Psi_2, \cdots, \Psi_k), \\ \text{where } \ \Psi_i &= \frac{\delta}{\delta \mathbf{S}_Z} ln \; {}_0F_1(\frac{1}{2}p, \frac{1}{4}\mathbf{S}_Z), \end{aligned} \tag{5.27}$$

and

$$\begin{aligned} \mathbf{H} &= (\phi_{ij}), \\ \text{where } \ \phi_{ij} &= \frac{\delta^2}{\delta^2 \mathbf{S}_Z} ln \; {}_0F_1(\frac{1}{2}p, \frac{1}{4}\mathbf{S}_Z) \end{aligned} \tag{5.28}$$

This end the proof. Q.E.D.

In our implementation, we employed a numerical approximation[24] discussed in (Šmídl & Quinn, 2007) to approximate the two functions $\mathbf{G}(\mathbf{S})$ and $\mathbf{H}(\mathbf{S})$.

Arranging the formulas of Table 5.2 to allow sequential updates of the different variables yields the iterative steps summarized in Algorithm 5.1:

5.4.4 Computational complexity

The proposed VBONMF model involves updates of simple analytical forms. This is thanks to the deterministic behaviour of variational Bayesian inference that makes obtaining analytical closed forms formulas, which are easier to compute, instead of time-consuming sampling techniques.

[24] We thank the authors for sharing the code http://staff.utia.cas.cz/smidl/files/mat/OVPCA.zip

Algorithm 5.1 VBONMF algorithm

```
**Input**: $\mathbf{X}$
**Set constants:** $\alpha_0, \beta_0, \alpha_\lambda, \beta_\lambda, k$
**repeat**
    **for** $l = 1$ *to* $k$ **do**
        Evaluate $q(\lambda_l)$
    **end for**
    **for** $l = 1$ *to* $k$ **do**
        Evaluate $q(m_{il})$
    **end for**
    Evaluate $q(A)$
    Set $\mathbb{E}[\mathbf{A}] \leftarrow \max(\mathbb{E}[\mathbf{A}], 1e^{-10})$
    Set $\mathrm{Var}[\mathbf{A}] \leftarrow \max(\mathrm{Var}[\mathbf{A}], 1e^{-10})$
    Evaluate $q(\tau)$
**until** *convergence is reached*;
Set $\mathbb{E}[\mathbf{M}] \leftarrow \mathbb{E}[\mathbf{M}]$ with zero columns removed
Set $\mathbb{E}[\mathbf{A}] \leftarrow \mathbb{E}[\mathbf{A}]$ with zero rows removed
**Return** $\mathbb{E}[\mathbf{M}], \mathbb{E}[\mathbf{A}]$
```

Formally, the time complexity of the proposed VBONMF model is $O(k^2bn)$ per iteration. We suppose that Algorithm 5.1 stops after t iterations; therefore, the overall cost becomes $O(tk^2bn)$.

5.5 Experiments

The performance of the proposed model was thoroughly evaluated using real-world datasets of multispectral (MS) images. In our experiments, we focused on the task of text extraction in MS document images, which is also called binarization. This task is one of the central and challenging tasks in document images analysis pipelines. The rationale here is that the decomposition quality of the document images greatly affects the quality of text extraction. That is, a good separation of text from other components corresponds to generating a clean version of the input image, which, when fed into any binarization method, yields improved results and vice-versa.

In view of that, we compared our results with the best results of some existing approaches (whose code is made available) under the same evaluation protocol. For a fair comparison, we performed two sets of experiments in order to answer the following two questions that arise here:

1) *Can we gain more from enforcing orthogonality constraints in Bayesian NMF models?*
2) *As for the orthogonality over the Stiefel manifold, between the probabilistic and non-probabilistic frameworks, which one is better for this task?*

To answer the first question, the proposed model is compared with two existing Bayesian NMF approaches to benchmark the orthogonality condition effects on Bayesian models. In order to evaluate the orthogonality over the Stiefel manifold strategy in different settings and respond to the second question, the proposed VBONMF is evaluated against other non-probabilistic ONMF models with different forms of orthogonality conditions.

5.5.1 Datasets

We validated our model on three existing datasets of real-world MS images of ancient documents, namely:

1) MSTEx-1 (Hedjam *et al.*, 2015): a collection of 21 MS data-cubes of real ancient documents images. This collection was built by the Synchromedia research group[25]. Each data-cube consists of 8 images taken from eight different bands, ranging from the ultra-violet (UV) spectrum (340*nm*) and visible spectrum (Blue, 500*nm*, Green, 600*nm*, and Red, 700*nm*), to the infrared spectrum (IR) (800*nm*, 900*nm*, 1000*nm*, and 1100*nm*).
2) MSTEX-2 (Hedjam *et al.*, 2015): another collection of ten MS data cubes of ten different ancient document scenes provided by the same research group.
3) HISTOODOC1 (Hedjam & Cheriet, 2013b) consists of a subset of 9 MS data cubes of 9 images taken from the MSTEx-1 (Hedjam *et al.*, 2015) dataset, in which the original images were cropped[26]. The size of these copies was reduced in order to meet the computational

[25] available at http://www.synchromedia.ca/databases/MSI-HISTODOC

[26] available at http://www.synchromedia.ca/databases/HISTODOC1

requirements of different algorithms used in (Hedjam & Cheriet, 2013b). Thus, instead of processing the entire images, only regions of interest were considered. For the same reason, we chose this dataset to benchmark our model against some existing Bayesian NMF models.

The corresponding ground-truth (GT) in these dataset are all binary images, the pixels being labeled into text and non-text classes only. Thus, each pixel of the GT images is either one (background) or zero (text).

5.5.2 Evaluation metrics

A set of four metrics widely used to measure the quality of document image binarization (as in DIBCO contests) were used to objectively assess the performance of the proposed model and quantitatively compare it with state-of-the-art methods. This set includes the F-Measure (FM), distance reciprocal distortion (DRD), negative rate metric (NRM), and the peak signal-to-noise ratio (PSNR), whose definitions are given in Appendix III. The values of FM are expressed in percentage (%) and PSNR in decibels (dB), whereas the values of DRD and the NRM range from 0 to 1. The higher the F-Measure and PSNR values are, the better the binarization quality is. In contrast, the lowest is the value of NRM or DRD, the best is the quality of the binarization.

5.5.3 Convergence

The empirical convergence rate of VBONMF and BNMF_ARD (defined in the next section) algorithms in terms of mean square error (MSE) is shown in Fig.5.2. The two algorithms converge within a few tens of iterations. The curve of BNMF_ARD goes under that of VBONMF, reaching a somewhat lower error rate.

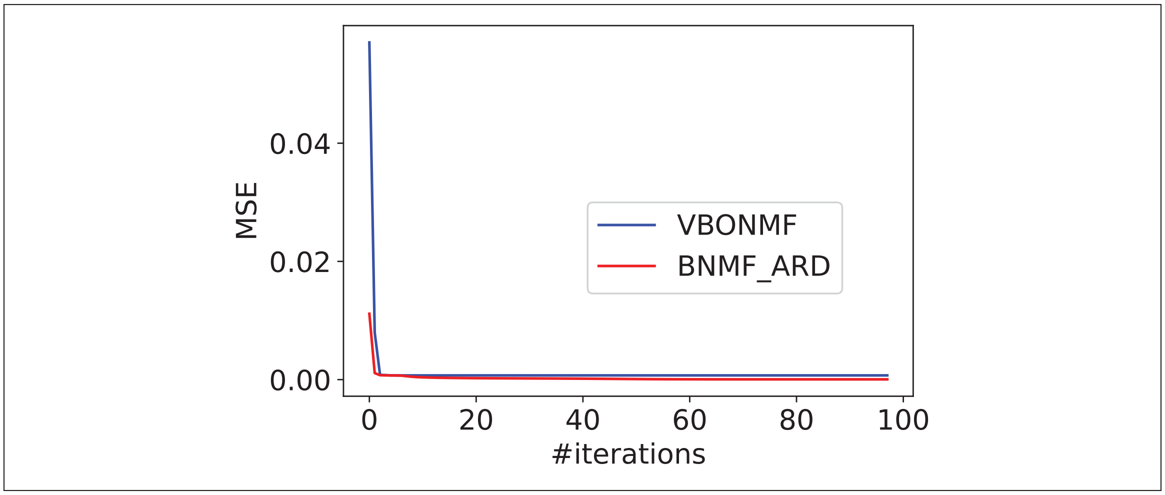


Figure 5.2 Typical convergence of VBONM and BNMF_ARD in terms of MSE. The experiment has been carried out on ẑ60 sample from the HISTODOC1 dataset

5.5.4 Text extraction in the HISTODOC1 dataset

In this experiment, we compared our model against three Bayesian NMF models, namely: the Bayesian nonnegative matrix factorization model (BNMF-ARD)[27] (Brouwer *et al.*, 2017), the probabilistic sparse non-negative matrix factorization model (psNMF)[28] (Hinrich & Mørup, 2018), and the Bayesian nonnegative matrix factorization model with adaptive sparsity and smoothness prior model (NMF-APC)[29] (Tichỳ *et al.*, 2019). As mentioned above, to cope with the computational requirements of these methods, the validation was performed on HISTODOC1. We also considered the standard NMF[30] to show the improvement made by the Bayesian framework.

However, the issue that arises here is that all factorization models generate positive real-valued outputs that are not binary. Therefore, we resorted to another post-processing step to convert them to the needed binary form. For this step, we used two state-of-the-art document image

[27] https://github.com/ThomasBrouwer/BNMTF_ARD

[28] https://github.com/JesperLH/psNMF-LVA2018

[29] http://www.utia.cz/AS/softwaretools/image_sequences/.

[30] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition

binarization methods, namely: Howe (Howe, 2013), which is one of the best documents binarization algorithms in the literature, and the supervised Convolutional Neural Networks (CNN) based approach, called Selectional Auto-encoder (SAE)[31] (Calvo-Zaragoza & Gallego, 2019).

Furthermore, to make the comparison meaningful, we structured the results as follows: 1) Howe's method alone, 2) SAE alone[32], 3) factorization + SAE, and finally 4) factorization + Howe. On the one hand, this allows us to compare the performance of our proposed method against other Bayesian and non-probabilistic NMF models. On the other hand, it allows us to evaluate the separation effects on the performance of state-of-the-art binarization methods.

As for the parameter setting of these models, we used the SAE model in a transfer learning fashion. i.e., it was trained on a dataset of grey-level images from DIBCO, and we used its weights without retraining on MS images. For VB-NMF and VBONMF, we used a grid search method in the range of $\{10^{-5}, \cdots, 10^{5}\}$ to tune the associated parameters. For psNMF (Hinrich & Mørup, 2018) and NMF-APC (Tichỳ *et al.*, 2019), we employed the authors' default parameters.

As for the rank k (i.e., number of objects), for a fair comparison, we set its value manually for all these methods. However, for our approach it is possible to use the estimated value. Fig.5.3 shows a plot of MSE reached by VBONMF as a function of different rank values. The estimated number of objects corresponds to the minimum value of MSE. In this example, the lowest error value is obtained with $k = 4$ and he visually evaluated value is 3. The value $k = 3$ gives also a low MES value, which means that, even without reaching the lowest error point, in practice, the ARD technique still can give a close estimate of the number of sources.

Table 5.3 shows the detailed performance of the proposed model and the compared algorithms. In particular, we can observe that i) using factorization as a decomposition step improved the final binarization results of both Howe's and SAE methods, except for the NMF-APC model. The decomposition helped separate the text from the unwanted components (i.e., degradation,

[31] https://github.com/ajgallego/document-image-binarization

[32] We applied both methods on a single input image from the full stack of images.

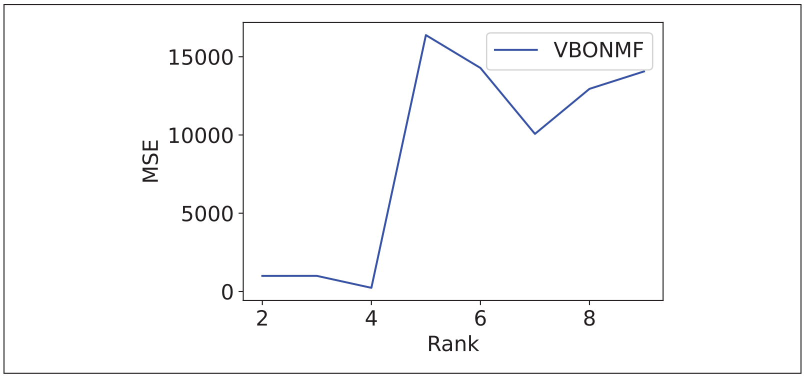


Figure 5.3 Plot of MSE as a function of the rank value. The estimated value is $k = 4$ and the current value is $k = 3$

paper, etc.), which helped boost their binarization performance; ii) the orthogonality yields a better components separation, which allowed the proposed model to outperform all other models; iii) the results also confirm that the training phase significantly affects the behaviours of deep learning approaches. Here, even when pre-trained on other document images, SAE could not provide good results on samples from the HISTODOC1 dataset that had not been seen before.

A qualitative comparison is illustrated in Fig.5.4 for a visual assessment of the obtained results. Fig.5.4 shows that the process involving VBONMF and BNMF-ARD generates binary images that are comparable to the quality of GT images. The standard NMF and NMF-APC are the most affected by the degradation present on the samples ($\hat{z}30$) and ($\hat{z}76$).

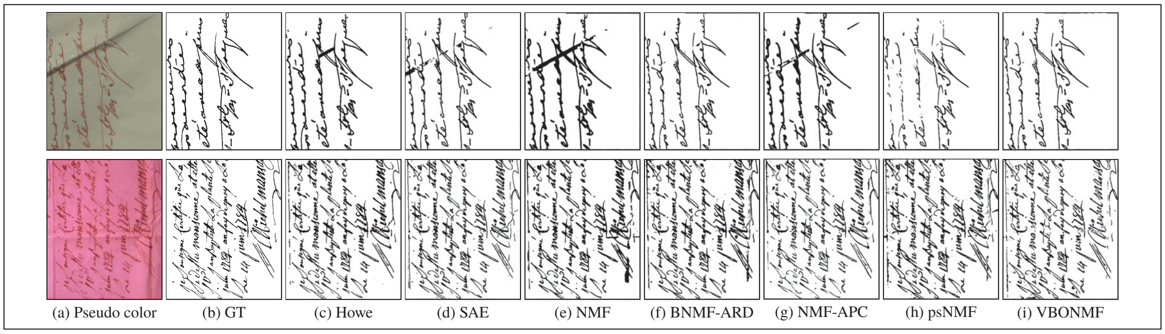


Figure 5.4 Qualitative comparison between different methods for the task of document binarization. The sample $\hat{z}30$, its GT image, and the related results are illustrated in the 1^{st} row, $\hat{z}76$ and the corresponding results are shown in 2^{nd} row

Table 5.3 Results of various methods in the task of HISTODOC1 dataset binarization. Best values are highlighted in bold and underlined, second best are highlighted in bold only

data	Metric	Howe	SAE	Factorization + SAE					Factorization + Howe				
				NMF	BTNMF	NMF-APC	psNMF	VBONMF	NMF	BTNMF	NMF-APC	psNMF	VBONMF
$\hat{z}$30	FM↑	89.79	82.49	77.26	85.53	82.14	75.72	88.58	91.33	**92.33**	81.24	92.09	**92.54**
	DRD↓	3.43	6.04	10.32	4.66	7.22	7.25	3.76	2.63	**2.23**	7.51	2.28	**2.13**
	NRM↓	2.72	8.60	3.13	6.88	2.38	17.90	3.30	2.84	**1.80**	2.19	2.86	**2.14**
	PSNR↑	17.32	15.16	13.06	15.99	14.37	14.54	16.82	18.13	**18.61**	14.08	18.58	**18.78**
$\hat{z}$36	FM↑	83.54	71.26	80.04	80.01	61.66	83.12	83.35	83.47	**86.02**	2.78	72.61	**87.14**
	DRD↓	4.55	5.99	5.92	5.95	11.59	4.01	3.59	4.47	**3.12**	12.59	5.96	**2.52**
	NRM↓	4.19	17.7	**3.04**	**3.02**	16.95	5.21	7.93	4.23	4.71	49.32	16.39	6.57
	PSNR↑	16.33	14.74	15.15	15.14	12.60	16.29	16.66	16.31	**17.24**	12.02	14.86	**17.86**
$\hat{z}$58	FM↑	76.62	52.39	68.69	74.53	12.04	69.46	68.84	75.03	**85.34**	48.29	74.36	**85.03**
	DRD↓	6.87	8.80	11.74	8.37	11.56	12.07	10.66	7.42	**2.85**	11.47	8.45	**3.11**
	NRM↓	6.92	29.54	6.15	**5.59**	46.76	**4.12**	8.69	7.65	7.14	29.62	6.10	6.25
	PSNR↑	15.67	14.17	13.77	15.02	13.04	13.73	14.08	15.36	**18.26**	13.39	15.03	**18.05**
$\hat{z}$60	FM↑	80.72	78.28	80.74	68.61	82.47	81.60	82.19	83.73	**87.84**	83.84	83.95	**85.79**
	DRD↓	5.32	5.26	5.42	8.39	3.54	4.63	5.11	4.24	**2.86**	3.43	4.64	**3.77**
	NRM↓	7.44	10.95	8.44	15.53	9.84	7.13	5.78	6.06	**3.51**	10.23	**4.24**	4.38
	PSNR↑	15.99	15.73	16.11	14.04	16.79	16.22	16.25	16.77	**18.01**	17.27	16.65	**17.31**
$\hat{z}$67	FM↑	33.99	58.16	52.54	64.82	20.97	63.03	65.68	71.67	**72.38**	13.55	27.85	**71.33**
	DRD↓	8.50	8.93	13.76	7.92	17.09	9.99	7.70	**5.56**	**5.36**	84.46	51.11	5.59
	NRM↓	38.75	19.02	17.83	13.61	41.13	**11.64**	13.55	13.14	12.78	36.79	21.16	**12.07**
	PSNR↑	15.10	14.96	13.73	15.50	12.72	14.91	15.66	**16.87**	**16.98**	6.99	9.03	16.79
$\hat{z}$68	FM↑	85.41	74.97	74.21	75.09	15.12	72.63	83.03	82.84	83.94	75.97	**86.93**	**88.67**
	DRD↓	3.75	7.26	8.96	8.69	12.17	7.58	3.67	5.05	3.57	7.47	**3.07**	**2.24**
	NRM↓	**2.30**	8.08	6.18	4.83	45.79	12.24	8.29	2.36	7.42	6.90	**1.98**	4.53
	PSNR↑	20.21	17.84	17.46	17.53	15.55	17.77	20.03	19.37	20.22	17.95	**20.74**	**21.69**
$\hat{z}$70	FM↑	76.80	62..64	65.13	74.57	71.86	69.52	73.06	77.35	**86.74**	72.86	83.85	**87.20**
	DRD↓	6.36	14.49	13.11	7.31	6.01	10.48	5.85	6.47	**2.68**	7.22	3.66	**2.58**
	NRM↓	5.81	4.61	5.32	3.01	11.10	4.43	11.57	4.14	**2.76**	8.73	**2.51**	3.74
	PSNR↑	18.70	15.56	16.11	17.89	18.13	16.92	18.45	18.68	**21.39**	18.08	20.36	**21.67**
$\hat{z}$76	FM↑	**87.88**	81.03	77.75	78.44	84.19	77.84	79.98	87.16	**88.02**	84.95	86.59	87.22
	DRD↓	**3.85**	7.50	10.05	9.72	5.72	10.01	8.08	4.29	**3.74**	5.23	4.57	4.05
	NRM↓	4.39	6.28	4.97	4.80	5.99	5.60	7.61	**4.06**	4.37	5.18	**4.29**	5.24
	PSNR↑	**14.96**	12.74	11.63	11.80	13.73	11.73	12.58	14.62	**15.02**	13.90	14.42	14.78
$\hat{z}$95	FM↑	77.44	68.88	13.85	72.74	37.08	69.67	**82.58**	15.01	76.43	32.19	15.05	**80.68**
	DRD↓	4.99	6.84	16.33	7.74	14.49	7.07	**4.27**	65.00	228.46	5.24	228.55	**4.83**
	NRM↓	15.95	21.02	46.42	13.37	36.80	18.67	**10.13**	49.45	16.47	23.65	49.33	**11.75**
	PSNR↑	14.91	13.68	10.43	13.38	11.07	13.45	**15.61**	0.46	14.73	5.63	0.47	**15.23**
Average	FM↑	76.91	70.01	65.58	74.93	51.95	73.62	78.59	74.18	**84.34**	55.07	69.25	**85.07**
	DRD↓	5.29	7.90	10.62	7.64	9.93	8.12	5.85	29.84	**3.52**	22.71	34.70	**3.42**
	NRM↓	9.83	13.98	11.28	7.85	24.08	9.66	8.54	10.44	**6.77**	19.18	12.10	**6.41**
	PSNR↑	16.58	14.95	14.16	15.14	14.22	15.06	16.24	15.17	**17.83**	13.26	14.46	**18.02**

To highlights the gained improvement in terms of FM and PSNR metrics, we provide in Fig. 5.5 bar chart representations of the scores obtained by all the models. As it can be seen, the worst results were obtained by NMF-APC. However, both models, BNMF-ARD and VBONMF, obtained the top scores in terms of FM and PSNR metrics. Significant performance gains were obtained with SAE as well, but less than the gain obtained with Howe's method. Compared with Howe's method, FM is boosted up to **4.31**% with BNMF-ARD factorization and to **8.16**% with VBONMF. As for SAE, using BNMF-ARD increased the FM score by **14.33**%, and **15.06**% is gained with VBONMF.

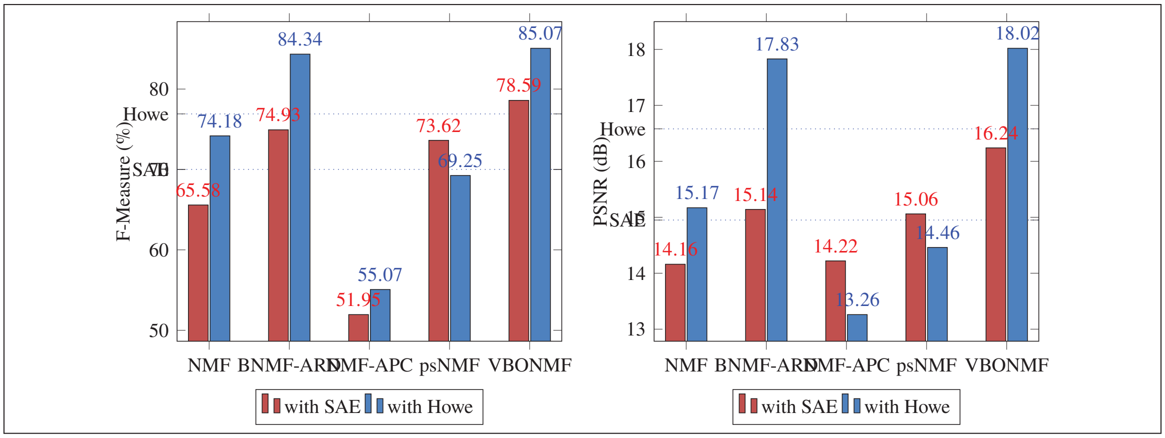


Figure 5.5 Results improvement in terms of FM and PSNR metrics for NMF, BNMF-ARD, NMF-APC, psNMF, and VBONMF, respectively. The two horizontal dotted lines denote the average scores of Howe's and SAE methods alone

5.5.5 Text extraction in the MSTEX-2 dataset

In this experiment, the proposed model is evaluated on the MSTEx-2 dataset and compared with 11 methods, including seven ONMF models, namely: EM-ONMF[33] (Expectation-Maximization Orthogonal NMF) (Pompili *et al.*, 2014), ONP-MF[33] (Orthogonal Nonnegatively Penalized Matrix Factorization) (Pompili *et al.*, 2014), OMNTF[34] (Orthogonal Nonnegative Matrix Tri-Factorization) (Yoo & Choi, 2010), M-ONMF (uni-orthogonal NMF) (Rahiche & Cheriet, 2021a), A-ONMF (uni-orthogonal NMF) (Rahiche & Cheriet, 2021a), MA-ONTF (bi-orthogonal nonnegative matrix Tri-factorization) (Rahiche & Cheriet, 2021a), and KONMF (Kernel Orthogonal ONMF model) (Rahiche & Cheriet, 2021b). In addition to Howe's (Howe, 2013) and SAE (Calvo-Zaragoza & Gallego, 2019) methods employed in the previous experiment, two other binarization approaches including SKKHM[35] (Spatial Kernel K-Harmonic Means) (Li *et al.*, 2007), and GMM (Gaussian Mixture Model) (Hollaus *et al.*, 2018b) were also considered.

[33] Available at https://github.com/filippo-p/onmf

[34] In NMFlib https://www.ee.columbia.edu/~grindlay/code.html

[35] Code is available at http://utopia.duth.gr/~nmitiano/

As in the previous experiment, the Howe's method is used on top of these models to generate the targeted binary images.

Furthermore, this comparison allowed us also to evaluate the orthogonality constraint over the Stiefel manifold when applied in a Bayesian and deterministic framework. It also allowed us to compare the orthogonality over the Stiefel manifold strategy with other strategies.

Table 5.4 reports the numerical results of all the methods evaluated on the MSTEx-2 dataset. In this experiment, we observe that the proposed VBONMF achieves the best scores in terms of almost all the metrics, except for the NRM metric, where KONMF obtained the best (lowest) value, i.e., **6.51** (versus **7.38** obtained by VBONMF). We observed that three orthogonal models, namely, AM-ONTF, KONMF, and VBONMF, have performed very well in this experiment.

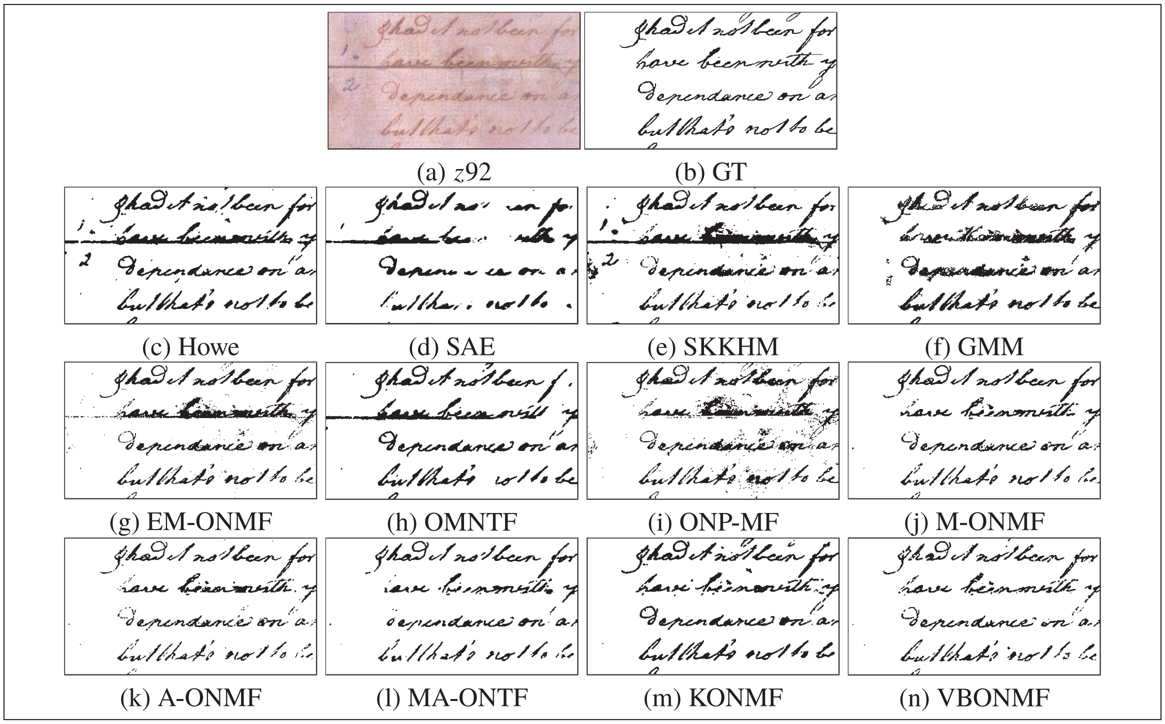
Table 5.4 Results of various methods for the binarization of the HISTODOC1 dataset

Ms-cube	Metric	Howe	SAE	SKKHM	GMM	EM-ONMF	OMNTF	ONP-MF	M-ONMF	A-ONMF	MA-ONTF	KONMF	VBONMF
z92	FM	72.11	63.88	69.88	73.12	78.42	77.33	78.87	**82.72**	81.09	80.32	**82.62**	82.48
	DRD	8.68	9.55	10.74	5.98	6.14	6.57	6.30	**3.56**	4.64	**3.76**	4.40	4.29
	NRM	10.46	17.81	8.93	9.70	7.75	7.18	7.37	9.55	**6.26**	12.26	**4.99**	6.55
	PSNR	-	11.49	11-59	11.94	-	-	-	-	-	-	**14.48**	**14.61**
z58-2	FM	19.93	7.21	18.13	77.36	76.81	**79.87**	74.80	77.61	78.22	76.86	78.47	**85.25**
	DRD	11.51	12.19	71.34	6.88	7.22	5.30	8.30	6.05	5.12	7.03	**5.03**	**3.09**
	NRM	44.38	48.13	36.54	9.01	**8.77**	9.46	12.16	11.12	13.06	9.13	12.96	**8.56**
	PSNR	-	13.19	6.89	14.86	-	-	-	-	-	-	**16.89**	**18.43**
z27	FM	78.57	49.24	70.59	81.97	79.76	82.04	78.33	84.36	80.62	**85.08**	**85.35**	83.64
	DRD	6.96	9.73	10.35	5.52	5.28	4.88	5.45	4.57	5.18	**4.26**	**4.18**	4.42
	NRM	10.20	33.29	14.50	9.57	13.17	9.70	14.73	6.84	11.85	**6.76**	**6.57**	9.08
	PSNR	-	10.59	10.99	13.34	-		-	-	-	-	**14.16**	**13.87**
z65	FM	84.35	81.56	74.05	85.34	72.98	84.99	70.12	85.49	78.71	**87.97**	85.30	**88.03**
	DRD	4.21	4.35	6.55	3.74	5.10	4.32	11.40	3.91	5.85	**2.45**	4.05	**2.38**
	NRM	6.68	6.27	14.61	8.01	19.62	**4.83**	8.97	5.58	10.37	7.20	**5.41**	7.19
	PSNR	-	15.01	14.03	16.11	-	-	-	-	-	-	**16.15**	**17.40**
z59-2	FM	81.53	78.43	72.92	**90.21**	80.11	81.77	74.81	86.65	85.75	83.70	86.42	**88.76**
	DRD	5.51	5.69	10.08	**1.89**	6.01	5.54	9.25	3.49	3.29	3.75	3.67	**2.41**
	NRM	6.42	8.94	7.09	6.92	7.61	6.01	6.01	**4.67**	6.97	9.43	**3.93**	5.26
	PSNR	-	16.36	14.79	**20.16**	-	-	-	-	-	-	18.31	**19.40**
z82-2	FM	84.64	81.67	79.09	**87.39**	15.38	84.31	84.83	86.49	86.80	**87.48**	86.55	85.76
	DRD	2.64	3.42	3.93	2.36	144.93	2.79	2.66	**2.29**	2.32	**2.21**	2.35	2.64
	NRM	10.41	9.44	14.13	**6.47**	30.30	10.20	10.73	9.24	8.27	8.86	8.85	**8.34**
	PSNR	-	18.12	17.96	**19.70**	-	-	-	-	-	-	**19.64**	19.33
z80-2	FM	81.28	62.04	81.52	83.68	73.82	82.82	**88.69**	90.15	87.27	89.28	88.68	**89.03**
	DRD	2.35	20.57	5.25	3.88	4.79	5.09	**2.35**	2.66	2.93	**2.39**	3.28	3.06
	NRM	7.31	4.38	10.28	11.80	20.70	5.43	8.50	4.43	8.11	5.79	**2.16**	**2.96**
	PSNR	-	13.68	18.76	**21.50**	-	-	-	-	-	-	20.45	**20.68**
z90	FM	84.64	68.38	51.31	**87.39**	56.63	76.07	56.57	78.89	67.41	86.22	85.96	**86.92**
	DRD	2.64	7.72	26.81	**2.36**	19.52	9.23	24.45	7.47	13.76	3.36	3.35	**3.19**
	NRM	10.41	18.74	15.43	**6.47**	16.77	6.93	10.98	7.16	11.10	6.67	7.45	**6.20**
	PSNR	-	15.51	11.67	13.83	-	-	-	-	-	-	**18.74**	**18.96**
z31	FM	66.22	63.43	59.34	**85.60**	65.93	67.48	67.48	68.05	68.28	80.84	**82.54**	79.66
	DRD	12.48	14.52	13.01	**2.96**	1023	10.22	10.22	11.00	10.93	4.70	3.64	**3.56**
	NRM	9.29	8.64	14.23	**7.63**	15.07	13.42	13.42	10.46	10.36	**8.43**	9.01	12.91
	PSNR	-	13.82	14.12	16.30	-	-	-	-	-	-	**18.37**	**17.99**
z43	FM	68.05	70.55	50.01	85.01	78.57	55.30	78.44	65.59	68.64	80.81	**85.49**	**85.36**
	DRD	10.78	11.74	13.56	3.33	5.07	11.68	5.09	11.22	8.32	**3.25**	3.72	**3.16**
	NRM	13.33	6.53	27.58	7.84	12.74	25.22	12.87	12.99	17.09	14.71	**3.85**	**6.79**
	PSNR	-	13.17	12.19	14.89	-	-	-	-	-	-	**16.96**	**17.22**
Average	FM	70.41	62.64	62.68	82.95	65.61	79.16	75.29	80.64	78.44	83.86	**84.73**	**85.49**
	DRD	8.09	9.95	17.16	**4.16**	22.09	6.55	8.55	5.73	6.13	3.72	3.76	**3.26**
	NRM	12.56	16.22	16.33	8.47	16.48	**7.33**	10.57	7.88	10.67	8.92	**6.51**	7.38
	PSNR	-	14.09	13.29	16.26	-	-	-	-	-	-	**17.42**	**17.79**

To observe the significant performance gains obtained by the three orthogonal models, VBONMF, KONMF, and AM-ONTF over Howe's, SAE, SKKHM, and GMM, Table 5.5 highlights the differences between the scores of these methods in terms of FM, DRD, NRM, and PSNR. It can be seen that VBONMF increased the FM score up to **22.85**% and **22.81**% compared to SAE and SKKHM, respectively. As stated before, the KONMF model was able achieve the lowest values in terms of NRM; up to **-9.82** gain was obtained compared to the SKKHM method.

Table 5.5 Gained performance by VBONMF, KONMF and AM-ONTF over Howe, SAE, SKKHM, and GMM in the task of text extraction

	MA-ONTF				KONMF				VBONMF			
	FM	DRD	NRM	PSNR	FM	DRD	NRM	PSNR	FM	DRD	NRM	PSNR
Howe	+13.45	-4.37	-3.64	-	+14.32	-4.33	**-6.05**	-	**+15.08**	**-4.83**	-5.18	-
SAE	+21.22	-6.23	-7.30	-	+22.09	-6.19	**-9.71**	+3.33	**+22.85**	**-6.69**	-8.84	**+3.70**
SKKHM	+21.18	-13.44	-7.41	-	+22.05	-13.40	**-9.82**	+4.13	**+22.81**	**-13.90**	-8.95	**+4.50**
GMM	+0.91	-0.44	+045	-	+1.78	-0.40	**-1.96**	+1.16	**+2.54**	**-0.90**	-1.09	**+1.53**



(a) $z92$ (b) GT

(c) Howe (d) SAE (e) SKKHM (f) GMM

(g) EM-ONMF (h) OMNTF (i) ONP-MF (j) M-ONMF

(k) A-ONMF (l) MA-ONTF (m) KONMF (n) VBONMF

Figure 5.6 Visual comparison of the binarization results of different models on the sample $z92$ from the MSTEx-2 dataset

For the sake of qualitative comparison, the next two figures, i.e., Figs. 5.7 and 5.6, compare the quality of final outputs obtained by all the methods and the GT. The results illustrated in Fig.

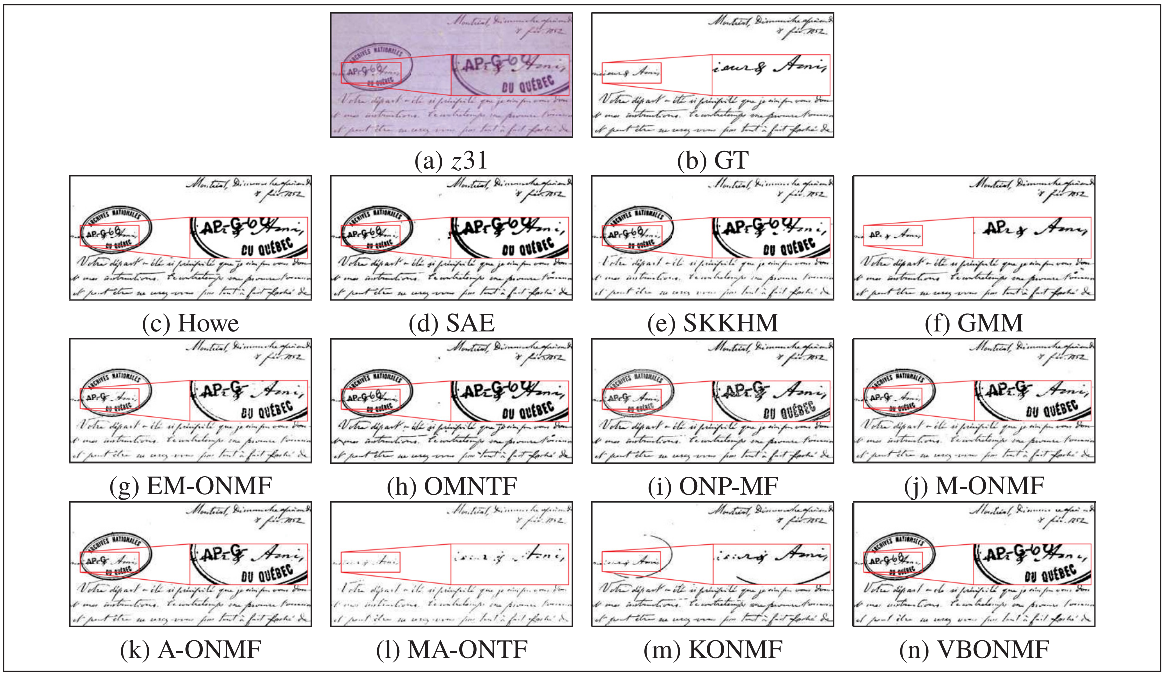


Figure 5.7 Qualitative comparison of the binarization outputs of different approaches on the sample z31 from the MSTEx-1 dataset

5.6 (sample z92) witness to the effectiveness of the proposed model. Because of the degradation present in this MS image, conventional approaches were not able separate the two text from the degraded background. However, the results in Fig 5.7 show a case were the proposed VBONMF fails in performing a good separation of all the objects from each other. In this example, a stamp, handwritten digits, and a line of the text overlap. Most approaches failed to separate between the overlapped objects. Only GMM, MA-ONTF and KONMF were able to perform acceptable separation of the four components (i.e., stamp, digits, text, and paper).

5.5.6 Comparison with winner methods in MSTex2015 contest

Since the MSTEx-2 collection was officially adopted for the MSTEx 2015 Competition (Hedjam *et al.*, 2015), we also compared the results of our approach with the two winning methods submitted to that contest, namely, MSIO (Diem *et al.*, 2016) and Hollaus *et al.* (Hedjam *et al.*, 2015). In addition to the algorithms considered in Table.5.4, we also included the results of a recent Convolutional Neural Network (CNN) based method developed by Hollaus *et al*. (Hollaus *et al.*, 2019). Table 5.6 reports the average scores obtained by all these methods.

Table 5.6 Comparison of mean scores of different methods against winners of MSTEx2015 contest in the binarization of the MSTEx-2 dataset

Method	FM (%)	DRD ($\times 10^{-3}$)	NRM ($\times 10^{-2}$)
1st rank of MSTEx2015 (MSIO (Diem *et al.*, 2016))	83.33	4.24	9.25
2nd rank of MSTEx2015 (Hollaus *et al*. (Hedjam *et al.*, 2015))	81.90	4.74	10.10
EM-ONMF (Pompili *et al.*, 2014)	65.51	22.09	16.48
SKKHM (Li *et al.*, 2007)	62.68	17.16	16.33
Howe (Howe, 2013)	70.41	8.58	12.06
ONP-MF (Pompili *et al.*, 2014)	75.19	8.55	10.57
MA-ONTF (Rahiche & Cheriet, 2021a)	83.86	3.72	8.92
GMM (Hollaus *et al*.)(Hollaus *et al.*, 2018b)	82.95	4.16	8.47
M-ONMF (Rahiche & Cheriet, 2021a)	80.64	5.73	7.88
Hollaus *et al*. (CNN)(Hollaus *et al.*, 2019)	79.90	4.71	9.44
ONMTF (Yoo & Choi, 2010)	79.16	6.55	7.33
A-ONMF (Rahiche & Cheriet, 2021a)	78.44	6.13	10.67
KONMF (Rahiche & Cheriet, 2021b)	84.73	3.76	**6.51**
Proposed VBONMF	**85.49**	**3.26**	7.38

It can be observed that the scores obtained by the proposed VBONMF outperformed the state-of-the-art approaches, including two deep learning techniques. As depicted in Table 5.6, VBONMF obtained the best scores in terms of FM and DRD metrics and the second-best score in terms of NRM.

5.5.7 Text extraction in the MSTEX-1 dataset

In this experiment, we validated our approach on a subset of 11 MS data-cubes from the MSTEx-1 dataset. We considered the same 11 methods employed in the previous experiment.

The detailed quantitative results obtained for all methods on each MS data-cube are reported in Table 5.7. It can be also observed that all orthogonal NMF models outperform standard binarization methods in the text extraction task. The three methods MA-ONTF, KONMF, and

VBONMF demonstrated their ability to separate text from other components and achieved the best scores in terms of all the metrics. The SAE method obtained the lowest scores with respect to all the metrics.

Table 5.7 Binarization results of different methods on MS-TEx-2 dataset. In bold and underlined, best performance; bold only, second-best scores. FM is expressed in %, DRD ×10^{-3}, and NRM ×10^{-2}. Size of MS cubes reads bands×width× height

Ms-cube	Metric	Howe	SAE	SKKHM	GMM	EM-ONMF	OMNTF	ONP-MF	M-ONMF	A-ONMF	MA-ONTF	KONMF	VBONMF
z64 (8x1470x686)	FM	82.71	72.43	82.21	**87.28**	**86.65**	81.78	84.67	85.38	66.53	82.93	83.32	82.23
	DRD	4.48	9.24	3.62	**3.14**	**3.33**	5.01	3.76	3.47	11.99	3.92	3.72	4.53
	NRM	**4.14**	3.70	12.24	5.97	5.93	**3.52**	6.66	7.14	8.97	7.93	6.89	4.58
	PSNR	-	12.66	**16.22**	-	-	-	-	-	-	-	**15.93**	15.45
z37 (8x1638x305)	FM	84.80	75.21	87.22	83.74	83.13	85.99	87.35	**88.31**	85.52	**87.95**	84.61	86.43
	DRD	4.37	10.13	**2.98**	3.86	3.64	4.02	3.17	**2.80**	3.52	3.12	4.13	3.39
	NRM	5.98	**5.47**	9.54	12.43	14.16	**4.59**	7.25	7.44	9.69	7.44	7.32	7.14
	PSNR	-	10.74	**15.07**	-	-	-	-	-	-	-	13.79	**14.45**
z35 (8x690x773)	FM	70.30	66.14	91.07	78.75	**91.97**	73.32	90.06	89.51	91.01	**92.81**	91.29	91.61
	DRD	17.33	10.01	2.32	5.99	**2.14**	14.61	2.63	2.78	2.38	**1.92**	2.42	2.36
	NRM	4.55	22.86	3.73	15.06	5.03	4.23	5.75	6.14	4.08	3.26	**2.16**	**2.61**
	PSNR	-	13.32	**18.22**	-	-	-	-	-	-	-	18.18	**18.39**
z80 (8x1627x523)	FM	78.97	79.62	71.61	82.45	58.55	79.32	**91.43**	88.98	89.55	88.28	**89.63**	82.31
	DRD	10.18	6.83	10.51	4.63	15.38	10.04	**2.45**	3.62	**3.01**	3.19	3.26	7.93
	NRM	4.91	9.29	15.87	14.45	25.29	4.86	5.54	**4.54**	6.93	7.67	**4.32**	4.56
	PSNR	-	14.37	13.21	-	-	-	-	-	-	-	**17.39**	**14.63**
z38 (8x1603x264)	FM	83.74	13.37	82.17	55.26	**87.48**	84.89	85.96	**87.34**	86.65	86.53	86.02	85.92
	DRD	3.91	14.73	3.92	14.36	**2.92**	4.13	3.25	**3.05**	3.34	3.39	3.84	3.94
	NRM	**5.08**	46.45	13.35	26.04	7.36	5.92	8.58	6.34	6.38	5.98	4.72	4.44
	PSNR	-	8.84	13.47	-	-	-	-	-	-	-	**13.67**	**13.60**
z68 (8x1506x448)	FM	84.24	74.26	81.09	83.49	83.58	83.51	81.05	**86.31**	85.73	85.07	83.85	**86.57**
	DRD	4.26	9.01	6.16	3.61	4.99	4.63	5.72	**2.90**	3.34	3.91	4.20	**2.92**
	NRM	**2.50**	3.91	3.46	9.96	3.53	**2.44**	5.16	8.20	6.70	3.30	4.91	6.22
	PSNR	-	14.12	15.86	-	-	-	-	-	-	-	**16.86**	**17.96**
z70 (8x1532x448)	FM	79.86	66.99	74.57	80.99	79.99	79.49	78.07	80.44	78.62	79.44	82.99	**82.46**
	DRD	5.01	12.76	7.55	5.19	5.77	5.53	6.18	4.54	4.74	4.34	**3.81**	**3.82**
	NRM	7.66	5.01	8.09	**3.56**	6.04	6.08	6.96	9.43	11.18	10.10	**6.56**	7.98
	PSNR	-	13.02	15.03	-	-	-	-	-	-	-	**17.13**	**17.14**
z76 (8x1319x748)	FM	86.46	80.78	85.16	87.02	86.42	85.45	85.55	86.69	82.59	**87.90**	86.85	**88.04**
	DRD	5.10	7.60	5.13	3.87	4.58	5.11	4.95	**3.82**	6.17	3.96	4.62	**3.73**
	NRM	**3.70**	7.23	6.48	8.05	6.25	4.25	6.65	9.02	8.28	4.22	**4.19**	4.54
	PSNR	-	12.95	14.31	-	-	-	-	-	-	-	**14.69**	**15.22**
z82 (8x1892x583)	FM	82.26	68.74	80.57	77.74	**84.72**	82.27	83.88	84.13	73.84	8**4.66**	83.95	84.56
	DRD	4.03	11.81	3.58	5.75	**2.78**	4.12	3.47	3.35	4.17	3.30	3.56	**3.28**
	NRM	7.23	4.17	13.65	9.60	10.81	6.80	8.92	6.50	20.39	6.38	**5.67**	**5.67**
	PSNR	-	13.99	17.92	-	-	-	-	-	-	-	**18.03**	**18.33**
z58 (8x1435x269)	FM	80.21	59.43	69.79	84.01	82.12	82.03	78.91	84.28	**85.19**	**85.41**	84.66	85.15
	DRD	4.98	7.56	10.05	3.75	4.84	4.32	5.79	**3.01**	**2.88**	3.33	3.59	3.25
	NRM	7.56	26.09	9.70	5.98	6.57	6.68	7.86	10.02	9.31	**4.80**	4.76	5.79
	PSNR	-	15.39	15.02	-	-	-	-	-	-	-	**18.39**	**18.66**
z100 (8x386x1066)	FM	84.60	75.62	73.01	77.24	88.95	84.68	85.90	86.45	86.09	88.57	**89.55**	**89.66**
	DRD	6.12	7.26	4.49	6.59	3.90	6.15	4.77	3.79	4.39	3.63	**3.06**	**3.30**
	NRM	4.25	16.37	17.58	16.57	4.22	**3.86**	6.42	9.62	8.28	6.10	5.44	**4.00**
	PSNR	-	14.10	14.69	-	-	-	-	-	-	-	**17.32**	**17.22**
Average	FM	81.83	66.60	79.86	79.82	83.05	82.07	84.80	**86.17**	82.85	**86.32**	**86.07**	85.90
	DRD	6.34	9.72	5.48	5.52	4.93	6.15	4.19	**3.38**	4.54	**3.46**	3.66	3.86
	NRM	5.23	13.69	10.34	11.61	8.65	**4.84**	6.89	7.67	9.11	6.11	**5.18**	5.23
	PSNR	-	13.05	15.04	-	-	-	-	-	-	-	**16.49**	**16.46**

Another qualitative evaluation (sample $z37$) is given in Fig 5.8. The results shown in this figure can be interpreted as follow:

1) The results show that the orthogonal NMF models (except ONMTF), especially the M-ONMF, A-ONMF, MA-ONTF, KONMF, VBONMF models (all imposing the orthogonality over the Stiefel manifold), besides performing a good text separation, were able to distinguish between two the types of ink present in the document $z31$. The word *verif* is a note that was written with a different ink at another time. All the traditional binarization techniques were not able to distinguish between the two texts.
2) Considering the results mentioned in 1), it is difficult to achieve higher scores on this sample. We recall that the GT images of MSTEx-1 and MSTEx-2 were generated for the task of text/non-text separation only. Thus, no differentiation between different texts was considered at that time. Moreover, the method used for the binarization (Nafchi *et al.*, 2014) is a traditional technique that does not allow for such differentiation.

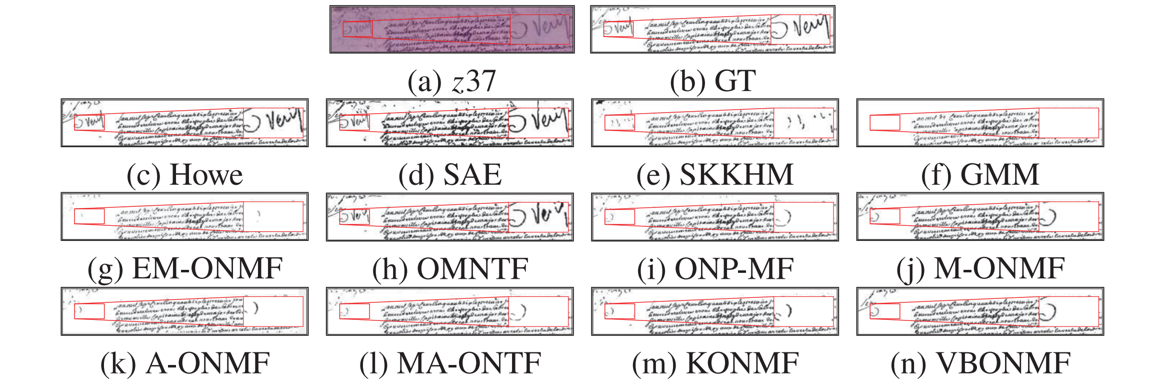
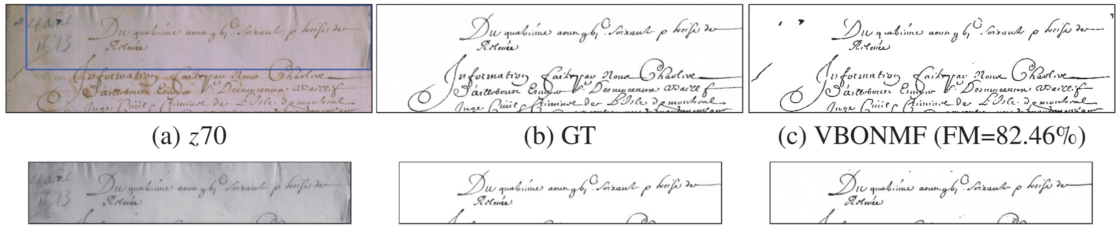

(a) $z37$ (b) GT
(c) Howe (d) SAE (e) SKKHM (f) GMM
(g) EM-ONMF (h) OMNTF (i) ONP-MF (j) M-ONMF
(k) A-ONMF (l) MA-ONTF (m) KONMF (n) VBONMF

Figure 5.8 Visual comparison of the results of different methods for the task of MS text binarization on the sample $z37$ from MSTEx-1 dataset

Moreover, it is interesting to note that for VBONMF, we observed that four MS samples ($z58$, $z68$, $z70$, and $z76$) from the MSTEx-1 dataset are also part of the HISTODOC1 dataset ($\hat{z}58$, $\hat{z}68$, $\hat{z}70$, and $\hat{z}76$). We recall that the HISTODOC1 dataset contains a subset of modified MS images from the MSTEx-1 dataset. Therefore, the question that one may ask here is *does the size of input images affect the performance of the results?* To answer this question, we compared the performance of VBONMF on both subsets. The obtained results are shown in Table 5.8. The results indicate that, on average, working on small patches gives better results than processing whole MS samples. An illustrative example that shows the results of VBONMF on the sample $z70$ and a on region cropped from it $\hat{z}70$ is given for qualitative evaluation in Fig.5.9.



(a) $z70$ (b) GT (c) VBONMF (FM=82.46%)

(d) $\hat{z}70$ (e) GT (f) VBONMF (FM=87.20%)

Figure 5.9 VBONMF text extraction results on the sample $z70$ from the MSTEx-1 dataset and the sample $\hat{z}70$ from the HISTODOC1 dataset. The blue box in (a) shows from where the image in (d) was cropped, approximately. (b) and (e) are the corresponding GT for (a) and (d), respectively, (c) and (f) are the final binary images

Table 5.8 Results of VBONMF on original samples from the MSTEX-1 dataset and their modified versions for the HISTODOC1 datasets

MS ID	Original sample from MSTEx-1 (z)				Modified sample from HISTODOC1 ($\hat{z}$)			
	FM	DRD	NRM	PSNR	FM	DRD	NRM	PSNR
Sample 58	85.15	3.25	5.79	18.66	85.03	3.11	6.25	18.05
Sample 68	86.57	2.92	6.22	17.96	88.67	2.24	4.53	21.69
Sample 70	82.46	3.82	7.98	17.14	87.20	2.58	3.74	21.67
Sample 76	88.04	3.73	4.54	15.22	87.22	4.05	5.24	14.78
Average	85.56	3.43	6.13	17.25	87.03	2.99	4.94	19.05

5.5.8 Further discussion

Based on our extensive experiments and observations, we can conclude the following points:

1) In response to the first question asked at the beginning of Section 5.5, we can say that, yes, we gain great deal from imposing orthogonality conditions on NMF. The above experiments have demonstrated that ONMF models also outperform other non-orthogonal NMF and traditional binarization techniques. The quality of obtained binary images, which is a very important parameter, in most cases was close to the GT.

2) For the second question, the proposed VBONMF achieved the top scores in terms of all metrics. This allows us to say that in this experiment, the Bayesian formulation has a positive impact on its performance.

3) The results show that working on small patches from MS samples gives results that are higher than the results obtained when processing the whole MS samples. This finding is interesting, as the processing time required can be dramatically reduced by parallelizing the treatment of several patches.

5.6 Conclusion

In this paper, we propose VBONMF, a new data-driven Bayesian orthogonal NMF model. In this generative model, the estimate of parameters is not point-wise as in deterministic NMF models but rather model-based. The parameters of the different distributions involved are estimated from the data using variational Bayesian inference. An extensive validation of our approach on text extraction on real MS document images was performed. The empirical evaluations show that the proposed framework outperforms several conventional and deep learning-based binarization approaches and helpes to improve their binarization performance.

CHAPTER 6

GENERAL DISCUSSION

This research work has addressed the problem of multichannel document image decomposition from a blind source separation perspective. The main goal, as stated before, was to develop a unified NMF based framework for the unsupervised decomposition of MS document images in an attempt to simplify their analysis and expand their scope of applications. Considering existing approaches' related challenges and limitations, we have defined three specific research objectives in Chapter 2. The investigation of these research objectives has resulted in the development of five novel NMF models. These new advanced MS document images processing approaches, their motivation, rationale development, and evaluations are discussed in Chapters 3, 4, and 5 of this thesis. In the following, we briefly discuss the strengths and limitations of the proposed methods and their contributions to the state-of-the-art of multichannel document image analysis.

6.1 Efficient orthogonality over the Stiefel manifold

Our first objective aimed to develop an NMF framework for the unsupervised decomposition of multichannel document images. Besides the nonnegativity, we investigated the orthogonality constraint as a key property for our framework. Accordingly, in Chapter 3 (Rahiche & Cheriet, 2021a), we have presented three new orthogonal NMF models to investigate the effect of orthogonality on each factor. The orthogonality constraint over the Stiefel manifold has yielded better results, enabled natural interpretations of outcomes, and allowed higher separation performances.

Besides being one of the first studies that address the blind decomposition problem of MS document images, our comparison results show that the proposed approaches:

1) Exploit orthogonality over the Stiefel manifold in NMF that yields better results than other strategies of incorporating the orthogonality condition.
2) Outperform conventional and state-of-the-art methods in the task of text extraction and binarization.

3) Significantly improve the performance of conventional methods when combined with them.

Despite our proposed models' capability to perform well in complete separation and text extraction tasks in MS images across different datasets, these approaches still face barriers and limitations. The first one is related to the assumption that the relationship between features in the spectral space is linear, which may not fit the nature of some real-world data. The second one concerns the need to manually select the number of sources, i.e., factorization rank. Moreover, as shown in Chapter 3, the NMF formulation adopted here ignores the noise present in real-world datasets. All those issues have been tackled in the last two objectives of the thesis. The former is tackled in the second paper, while the latter two limitations are addressed in the third paper.

6.1.1 Non-linearity mapping using Kernel methods with graph-based total variation regularization

In the second contribution, in Chapter 4, we combined the kernel trick, the orthogonality constraint over the Stiefel manifold, and the graph-based regularization. The orthogonality condition yields sparser solutions. The kernel mapping allows accounting for the non-linear relationship between features for better data modelling, which has had a significant impact on our model in terms of performance improvement. The graph regularization helped preserve the geometrical structure of texts in our document image data. Altogether, the proposed model yields a significant separation performance improvement.

However, in this model, the rank selection issue and the noise were not considered. This is because it is very challenging to address them deterministically. Thus, both limitations are left to the last contribution. In the proposed model, we investigated the Gaussian kernel function only. Other kernel functions are left for future improvements.

6.1.2 Variational generative model for orthogonal NMF

In the last contribution, Chapter 5, we addressed the two issues left from the two previous contributions. Thus, we proposed a probabilistic generative model that accounts for data noise and allows performing model order selection. Taking the orthogonality key feature into account, we developed an efficient variational Bayesian orthogonal NMF model.

Our proposed method was evaluated on the same datasets. The results show that the proposed Bayesian model outperforms several state-of-the-art approaches. In fact, such behaviour was not surprising. The Bayesian formulation helped build a fully nonlinear model to capture complex relationships between model elements. Incorporating the noise into account helped strengthen the model estimation. Involving prior information about the factors yields a better data representation.

As all the proposed models exploit the orthogonality condition of one or more factors, it is judicious to compare them from a conceptual point of view. Thus, we close this discussion by comparing the proposed algorithms in terms of the objective function, orthogonal factor, computational complexity, number of hyperparameters, their tuning method, and the rank selection strategy used in each algorithm.

Table 6.1 recapitulates the main features of each model[35]. In addition to what we have mentioned above and in the related articles, the main observations are as follows:

Table 6.1 Comparison between the proposed ONMF algorithms in terms of multiple criterion

	ONMF models				
Criterion	A-ONMF	M-ONMF	MA-ONTF	GTV-ONNMF	VBONMF
Objective function	$\|\mathbf{Y}-\mathbf{MA}\|_F^2$	$\|\mathbf{Y}-\mathbf{MA}\|_F^2$	$\|\mathbf{Y}-\mathbf{MA}\|_F^2$	$\|\mathbf{\Phi}(\mathbf{X})-\mathbf{\Phi}(\mathbf{M})\mathbf{A}\|_F^2+\lambda\|\mathbf{\Gamma}\mathbf{A}^T\|_1$	$\|\mathbf{Y}-\mathbf{MA}^T\|_F^2$
Orthogonal factor	$\mathbf{M}^T\mathbf{M}=\mathbf{I}_k$	$\mathbf{AA}^T=\mathbf{I}_k$	$\mathbf{M}^T\mathbf{M}=\mathbf{I}_k, \mathbf{AA}^T=\mathbf{I}_k$	$\mathbf{AA}^T=\mathbf{I}_k$	$\mathbf{A}^T\mathbf{A}=\mathbf{I}_k$
Graph regularization	No	No	No	Graph-based Total Variation	No
Optimization strategy	ADMM	ADMM	ADMM	ADMM + PG	VB
Time complexity	$O(t_1bkn)$	$O(t_2kn^2)$	$O(t_2kn^2)$	$O(t_2kn^2)$	$O(k^2bn)$
Hyperparameters	ρ_x, ρ_z	ρ_u, ρ_v	ρ_f, ρ_g, ρ_h	$\lambda, \gamma, \sigma, \rho, \beta$	$\alpha_0, \beta_0, \alpha_\lambda, \beta_\lambda$
Tuning method	Grid search	Grid search	Grid search	Grid search	Grid search
Rank estimation	Manual	Manual	Manual	Manual	ARD
GPU	No	No	No	Yes	No

1) **Objective function:** as an impact of the modularity of NMF and the orthogonality over the Stiefel manifold, the objective function in all models remains the same. This opens the

[35] For the different symbols, see the corresponding papers and the list of symbols provided at the beginning of this document.

door to either investigate other loss functions or integrate other specific regularization terms (priors) to improve the efficiency of the models.

2) **Orthogonal factor:** based on our experimental validations, combining orthogonality constraint with non-linearity in NMF gives better results. Imposing the orthogonality on both factors led to superior results in our experiments. As the behavior of these models may change with the application at hand, we can not confirm whether this observation will remain valid for other types of data or not.
3) **Graph regularization:** Only the TGV-ONNMF model that involves a graph regularization. Despites the cost induced by the graph, the results indicates a significant gain in performance.
4) **Optimization strategy:** ADMM was adopted for all non-probabilistic models, while a variational Bayesian inference was adopted for VBONMF. ADMM was very practical in simplifying the corresponding optimization problems to deal with more than one constraint efficiently. VB provides a powerful inference method to approximate intractable terms of the proposed VBONMF model.
5) **Time complexity:** A-ONMF, MA-ONTF, and TGV-ONNMF have similar time complexities and are also the most time-consuming approaches due to the requirements for performing optimization over the Stiefel manifold with w.r.t a $k \times n$ matrix. But, in general, some ADMM iteration updates are independent of each other and can therefore be updated in parallel, which can reduce their computational complexity. The VBONMF model has deterministic updating forms, which resulted in a simpler and less time-consuming algorithm.
6) **Hyperparameters:** all models involve related hyperparameters, for which the grid search method was adopted for selecting their optimal values. In ADMM based methods, while the convergence of ADMM itself is less sensitive to the value of the augmented Lagrangian parameter, this value has an important influence on the convergence speed of the algorithm.
7) **Rank estimation:** Handling the issue of source separation (rank estimation) was possible with the Bayesian model. This makes it very attractive and more practical. All other models require a manual setting of this parameter.

8) **GPU based implementation:** due to the incorporation of the kernel, only the GTV-ONNMF model requires GPU resources. However, in contrast to some existing kernel models in the literature, this GPU-based implementation allowed us to process large images.

Finally, a trade-off between the accuracy and the time complexity should be made to select the model that fits the needs of the application at hand. For instance, the GTV-ONNMF model requires GPU resources and several parameters to tune. However, it demonstrates very efficient performances.

CONCLUSION AND RECOMMENDATIONS

This research aimed to address the problem of spectral document image decomposition from a blind source separation perspective. We demonstrated the power of NMF models for tackling this challenging problem. We have investigated the orthogonality over the Stiefel manifold as a key feature for developing efficient ONMF models in deterministic and probabilistic settings. In total, we have presented five new orthogonal NMF algorithms, namely A-ONMF, M-ONMF, MA-ONTF, GTV-ONNMF, and VBONMF algorithms. We performed extensive experiments to assess the proposed models and provided evidence of their superior performance with respect to existing methods.

We started by studying the orthogonality over the Stiefel manifold as an efficient alternative to existing strategies and investigating the effect of on which factor we impose this condition. This led to the development of very efficient source separation methods.

We then addressed the problem of non-linearity in spectral image data and the geometrical structure of data lost by the vectorization of MS images. To account for the existing non-linear spectral relationship between features, we proposed a new kernel-based formulation that overcomes the pre-image problem inherited from the kernel machines. We combine the power of non-linear kernel mapping, orthogonality, and graph regularization to develop a more robust NMF model.

We finally reformulated the problem as a Bayesian graphical model. We involved the orthogonality as a uniform distribution over the Stiefel manifold, and we incorporated the ARD technique to automatically estimate the number of sources. Several comparative experiments were conducted to evaluate the proposed models and compare their efficiency against conventional techniques, deep learning methods, and existing competitive NMF models.

Through several quantitative and qualitative analyses, we demonstrated the superiority of the proposed algorithms over existing conventional techniques. We have also shown that the developed approaches can reveal other interesting latent information to an image analyst than any other traditional approach.

Recommendations for future work

Although the outstanding results obtained, the proposed methods are not perfect, and rooms for improvements are still open:

1) In the previous chapter, we discussed some limitations of our proposed models. One of the possible future directions is to investigate another optimization technique over the Stiefel manifold is another direction to improve the models.
2) The need for tuning the corresponding hyper-parameters is another challenge that we did not address in this study. For VBONMF, adding another hierarchical level to the model by assigning some priors to the different hyper-parameters would help to avoid tuning them.
3) For the deterministic models, the ADMM optimization scheme is parallelizable. Hence, reimplementing the models while considering this property may benefit the models to reduce their computational time.
4) In GTV-ONNMF model, we restricted our study to the RBF kernel function, though, other kernel functions can be investigated. A joint optimization and parameters estimation is another future direction to investigate in order to overcome the shortcoming of manual parameters tuning.

APPENDIX I

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

This Appendix is modified from the Supplementary Material of our paper in (Rahiche & Cheriet, 2021a). It provides additional details and results for the main paper that were omitted due to space restrictions. Section 1 gives a summary that highlights the difference between the proposed approaches and the main existing ONMF models. Section 2, provides further details about the optimization of the M-ONMF, A-ONMF, and MA-ONTF models, respectively. Section 3 provides the definitions of the metrics used in the paper for evaluating the proposed models. Section 4 provided additional qualitative and quantitative evaluations of our models. Section 5 reports the values of the different hyper-parameters used in our experiments.

1. Main ONMF approaches

Table. I-1 gives a summary of the main ONMF approaches with their corresponding formulation[36], the optimization method used, and the application targeted. In this brief summary, we didn't include models that adopt the same optimization strategies, didn't investigate new techniques, or are just extensions of existing models with some additional regularization terms.

2. Optimization solutions

2.1 M-ONMF optimization

Here, we can minimize each subproblem of Eq. 3.23 as an unconstrained problem with an easier formulation in an iterative way.

2.1.1 Optimization of the M-subproblem

[36] For the meanings of the different variables and parameters in each method, we refer the reader to the corresponding reference. The abbreviations used read MU: Multiplicative updates, EM: Expectation Maximization, PG: Proximate Gradient, HALS: Hierarchical Alternating Least Squares, APG: Accelerated Proximate Gradient, GD like: Gradient Descent like.

The M-step minimization involves minimizing a smooth function on the Stiefel manifold which is carried out using gradient-descent like optimization algorithms, as described in Section 3.2.4. In our case, this subproblem is solved using Algorithm 3.2.

The Euclidian gradient of $\mathcal{L}_1$ with respect to $\mathbf{M}$ is given by:

$$\begin{aligned}\nabla_M \mathcal{L}_1 &= \frac{\partial}{\partial \mathbf{M}}(\frac{1}{2}\|\mathbf{Y}-\mathbf{MA}\|_F^2 + \frac{\rho_x}{2}\|\mathbf{M}-\mathbf{X}+\mathbf{\Lambda_1}\|_2^2) \\ &= -\mathbf{YA^T} + \mathbf{MAA^T} + \rho_x(\mathbf{M}-\mathbf{X}+\mathbf{\Lambda_1}) \qquad \text{(A I-1)}\end{aligned}$$

The Riemannian gradient of $\mathcal{L}_1$ with respect to $\mathbf{M}$ on $\mathcal{M}_1$, denoted by grad$M$, is calculated using Eq. 3.8.

Table-A I-1 Variables update formulas/schemes used to handle the orthogonality constraint in some existing ONMF models

Method		Update formula/scheme	Optimization
BiOR-NM3F [24]	$\|\mathbf{X}-\mathbf{FSG}^T\|_F^2 + \mathrm{Tr}(\mathbf{\Lambda}^T(\mathbf{F}^T\mathbf{F}-\mathbf{I}))$ $+\mathrm{Tr}(\mathbf{\Gamma}^T(\mathbf{GG}^T-\mathbf{I}))$	$G_{jk} \leftarrow G_{jk}\sqrt{\frac{(X^TFS)_{jk}}{(GG^TX^TFS)_{jk}}}, F_{ik} \leftarrow F_{ik}\sqrt{\frac{(XGS^T)_{ik}}{(FF^TXGS^T)_{ik}}},$ $S_{ik} \leftarrow S_{ik}\sqrt{\frac{(F^TXG)_{ik}}{(F^TFSG^TG)_{ik}}}$	MU
OPNMF [23]	$\|\mathbf{X}-\mathbf{WW}^T\mathbf{X}\|_F$	$W \leftarrow W \odot \frac{XX^TW}{WW^TXX^T + XX^TWW^TW}$	MU
ONMF [26]	$\frac{1}{2}\|\mathbf{X}-\mathbf{UV}^T\|^2$	$V \leftarrow V \odot \frac{X^TV}{VU^TXV}$	MU
ONMF-A [28]	$\frac{1}{2}\|\mathbf{X}-\mathbf{AS}\|^2$	$A \leftarrow A \odot \frac{XS^T}{ASX^TA}$	MU
ONMF-S [28]	$\frac{1}{2}\|\mathbf{X}-\mathbf{AS}\|^2$	$S \leftarrow S \odot \frac{A^TX}{SX^TAS}$	MU
ONMTF[27]	$\|\mathbf{X}-\mathbf{USV}^T\|^2$	$U \leftarrow U \odot \frac{XVS^T}{USV^TXU}, V \leftarrow V \odot \frac{X^TUS}{VS^TU^TXV},$ $S \leftarrow S \odot \frac{U^TXV}{U^TUSV^TV}$	PG
ONP-MF [29]	$\frac{1}{2}\|\mathbf{M}-\mathbf{UV}\|_F^2 + \langle\mathbf{\Lambda}, -\mathbf{V}\rangle + \frac{\rho}{2}\|\min(\mathbf{V},0)\|_F^2$	$V \leftarrow Proj_{St}(V - \beta\nabla_V L_\rho(U,V,\Lambda)),$ $where\ Proj_{St}(V) = \frac{V}{\|VV^T\|}$	PG
Our approach	$\frac{1}{2}\|\mathbf{X}-\mathbf{MA}\|_F^2$	Algorithm2	GD like

2.1.2 Optimization of the A-subproblem

In order to minimize Eq. 3.25, we first compute the derivative of $\mathcal{L}_1(\mathbf{M}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{\Lambda_1}, \mathbf{\Lambda_2})$ w.r.t **A**, which is given as

$$\begin{aligned} \frac{\partial \mathcal{L}_1}{\partial \mathbf{A}} &= \frac{\partial}{\partial \mathbf{A}}(\frac{1}{2}\|\mathbf{Y} - \mathbf{MA}\|_F^2 + \frac{\rho_z}{2}\|\mathbf{A} - \mathbf{Z} + \mathbf{\Lambda_2}\|_2^2) \\ &= -\mathbf{M^T Y} + \mathbf{M^T MA} + \rho_z \mathbf{\Lambda_2} + \rho_z \mathbf{A} - \rho_z \mathbf{Z} \end{aligned} \tag{A I-2}$$

Then, by setting the first derivative equation $\frac{\partial \mathcal{L}_1}{\partial \mathbf{A}} = 0$ and solve it for **A**, we obtain the following **A**-update formula

$$\mathbf{A} = (\mathbf{M^T M} + \rho_z I)^{-1}(\mathbf{M^T Y} + \rho_z \mathbf{Z} - \rho_z \mathbf{\Lambda_2}) \tag{A I-3}$$

2.1.3 Optimization of the X-subproblem and the Z-subproblem

It is easy to show that the X-subproblem

$$\mathbf{X} = \arg\min_{\mathbf{X} \geqslant 0} \frac{\rho_x}{2}\|\mathbf{X} - (\mathbf{M} + \mathbf{\Lambda_1})\|_2^2, \tag{A I-4}$$

and the Z-subproblem

$$\mathbf{Z} = \arg\min_{\mathbf{Z} \geqslant 0} \frac{\rho_z}{2}\|\mathbf{Z} - (\mathbf{A} + \mathbf{\Lambda_2})\|_2^2, \tag{A I-5}$$

each has a solution in the direction of $\mathbf{M} + \mathbf{\Lambda_1}$ and $\mathbf{A} + \mathbf{\Lambda_2}$, respectively. Thus, closed-form solutions of these two sub-problems exist and are given by:

$$\mathbf{X} \leftarrow \text{prox}_{\iota_S}(\mathbf{M}^{t+1} + \mathbf{\Lambda}_1) \tag{A I-6}$$

$$\mathbf{Z} \leftarrow \text{prox}_{\iota_S}(\mathbf{M}^{t+1} + \mathbf{\Lambda}_2) \tag{A I-7}$$

where $\text{prox}_{\iota_S}(.) = \max(., 0)$ is the element-wise projection on positive set.

2.2 A-ONMF optimization

2.2.1 Optimization of the M-subproblem

After eliminating the unrelated and constant terms from Eq. 3.34 and taking the derivative w.r.t **M** we get

$$\begin{aligned}\frac{\partial \mathcal{L}_2}{\partial \mathbf{M}} &= \frac{\partial}{\partial \mathbf{M}}(\frac{1}{2}\|\mathbf{Y} - \mathbf{M}\mathbf{A}\|_F^2 + \frac{\rho_u}{2}\|\mathbf{M} - \mathbf{U} + \mathbf{\Pi_1}\|_2^2) \\ &= -\mathbf{Y}\mathbf{A}^{\mathbf{T}} + \mathbf{M}\mathbf{A}\mathbf{A}^{\mathbf{T}} + \rho_u\mathbf{\Pi_1} + \rho_u\mathbf{M} - \rho_u\mathbf{U}\end{aligned} \tag{A I-8}$$

Then, by setting Eq.(A I-8) to zero, we can obtain the following update formula for **M**

$$\mathbf{M} = (\mathbf{Y}\mathbf{A}^{\mathbf{T}} + \rho_u\mathbf{U} - \rho_u\mathbf{\Pi_1})(\mathbf{A}\mathbf{A}^{\mathbf{T}} + \rho_u I)^{-1} \tag{A I-9}$$

2.2.2 Optimization of the A-subproblem

We follow the same steps used in the M-ONMF for the minimization of Eq. 3.35, which involves minimizing a smooth function on the Stiefel manifold that can be optimized using gradient-descent like optimization algorithms, as summarized in Algorithm 3.2. The gradient of $\mathcal{L}_2$ with respect to **A** is given by

$$\begin{aligned}\nabla_A \mathcal{L}_2 &= \frac{\partial}{\partial \mathbf{A}}(\frac{1}{2}\|\mathbf{Y} - \mathbf{MA}\|_F^2 + \frac{\rho_v}{2}\|\mathbf{A} - \mathbf{V} + \mathbf{\Pi_2}\|_2^2) \\ &= \mathbf{M^T MA} - \mathbf{M^T Y} + \rho_v(\mathbf{A} - \mathbf{V}) + \rho_v \mathbf{\Pi}_2 \qquad \text{(A I-10)}\end{aligned}$$

2.2.3 Optimization of the U-subproblem and the V-subproblem

Considering the following corresponding subproblems:

$$\mathbf{U^{t+1}} = \arg\min_{\mathbf{U} \geqslant 0} \frac{\rho_u}{2}\|\mathbf{U} - (\mathbf{M^{t+1}} + \mathbf{\Pi_1})\|_2^2, \qquad \text{(A I-11)}$$

$$\mathbf{V^{t+1}} = \arg\min_{\mathbf{V} \geqslant 0} \frac{\rho_v}{2}\|\mathbf{V} - (\mathbf{A^{t+1}} + \mathbf{\Pi_2})\|_2^2, \qquad \text{(A I-12)}$$

As in the previous model, the closed-form solution for each variable is given by the following updates:

$$\mathbf{U} \leftarrow \text{prox}_{\iota_S}(\mathbf{M}^{t+1} + \mathbf{\Pi}_1) \qquad \text{(A I-13)}$$

$$\mathbf{V} \leftarrow \text{prox}_{\iota_S}(\mathbf{A}^{t+1} + \mathbf{\Pi}_2) \qquad \text{(A I-14)}$$

2.3 MA-ONTF

2.3.1 Optimization of the M-subproblem and the A-subproblem

As in the M-ONMF and A-ONMF models, the M-subproblem and the A-subproblem are solved using Algorithm 3.2, as described in Section 3.2.4.

2.3.2 Optimization of the Q-subproblem

Given the following subproblem:

$$\mathbf{Q^{t+1}} := \arg\min_{\mathbf{Q}} \frac{\rho_f}{2} \|\mathbf{MQ} - \mathbf{F} + \mathbf{\Gamma_1}\|_2^2. \tag{A I-15}$$

The solution is obtained by taking the first order derivative and solving for **Q**, thus

$$\mathbf{Q^{t+1}} := (\mathbf{M^T M})^{-1}(\mathbf{M^T}(\mathbf{F} - \mathbf{\Gamma_1})). \tag{A I-16}$$

2.3.3 Optimization of the F-subproblem

The corresponding subproblem is written as:

$$\mathbf{F^{t+1}} := \arg\min_{\mathbf{F} \geqslant 0} \frac{1}{2} \|\mathbf{Y} - \mathbf{FA^{t+1}}\|_F^2 + \frac{\rho_f}{2} \|\mathbf{M^{t+1} Q^{t+1}} + \mathbf{\Gamma_1} - \mathbf{F}\|_2^2,$$

After taking the derivative and solving the obtained equation for **F**, we can obtain the following updating formula:

$$\mathbf{F^{t+1}} := (\mathbf{YA^T} + \rho_f(\mathbf{MQ} - \mathbf{\Gamma_1}))(\mathbf{AA^T} + \rho_f \mathbf{I})^{-1} \tag{A I-17}$$

2.3.4 Optimization of the G-subproblem and the H-subproblem

The following subproblems:

$$\mathbf{G^{t+1}} := \arg\min_{\mathbf{G}\geqslant 0} \frac{\rho_g}{2}||\mathbf{M^{t+1}} + \mathbf{\Gamma_2} - \mathbf{G}||_2^2,$$
$$\mathbf{H^{t+1}} := \arg\min_{\mathbf{H}\geqslant 0} \frac{\rho_h}{2}||\mathbf{A^{t+1}} + \mathbf{\Gamma_3} - \mathbf{H}||_2^2,$$

have closed-form solutions given as follows:

$$\mathbf{G} \leftarrow \text{prox}_{\iota_S}(\mathbf{M}^{t+1} + \mathbf{\Gamma}_1) \tag{A I-18}$$
$$\mathbf{H} \leftarrow \text{prox}_{\iota_S}(\mathbf{A}^{t+1} + \mathbf{\Gamma}_2) \tag{A I-19}$$

As fo the stopping criterions used in the three algorithms, we have: $r_1 = \max(||\mathbf{M} - \mathbf{X}||_F, ||\mathbf{A} - \mathbf{Z}||_F)$, $r_2 = \max(||\mathbf{M} - \mathbf{U}||_F, ||\mathbf{A} - \mathbf{V}||_F)$, and $r_3 = \max(||\mathbf{MQ} - \mathbf{F}||_F, ||\mathbf{M} - \mathbf{G}||_F, ||\mathbf{A} - \mathbf{H}||_F)$

3. Evaluation metrics

For the sake of completeness, we state here the definition of all metrics used in the paper, namely, PCC, FM, NRM, DRD, and ACC.

1) ***The PCC*** Person's Correlation Coefficient measures the correlation between two images. It consists of the covariance between the two images, normalized by the product of their standard deviations. The PCC between two images is described as follows:

$$PCC = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{A I-20}$$

where $\bar{x}, \bar{y}$ represent the average intensity of the first and second image respectively, x_i and y_i denote the intensity of the i^{th} pixel from the first and second image, respectively.

2) ***The F1-score or F-measure (FM)*** is a weighted mean of the precision and recall that measures the accuracy of the prediction.

$$FM = \frac{2 \times Recall \times Precision}{Recall + Precision}, \tag{A I-21}$$

where $Recall = \frac{TP}{TP+FN}$, $Precision = \frac{TP}{TP+FP}$, and TP, FP, FN denote the True Positive, False Positive, and False Negative predictions respectively.

3) ***The Negative Rate Metric (NRM)*** measures the mismatches between the prediction and the ground-truth (GT) on the pixel level.

$$NRM = \frac{NR_{FN} + NR_{FP}}{2}, \tag{A I-22}$$

with $NR_{FN} = \frac{FN}{FN+TP}$, and $NR_{FP} = \frac{FP}{FP+TN}$

4) ***The Distance Reciprocal Distortion (DRD)*** measures the distortion between two binary images. It is defined as follows (Lu *et al.*, 2004):

$$DRD = \frac{\sum_{k=1}^{N} DRD_k}{NUBN}, \tag{A I-23}$$

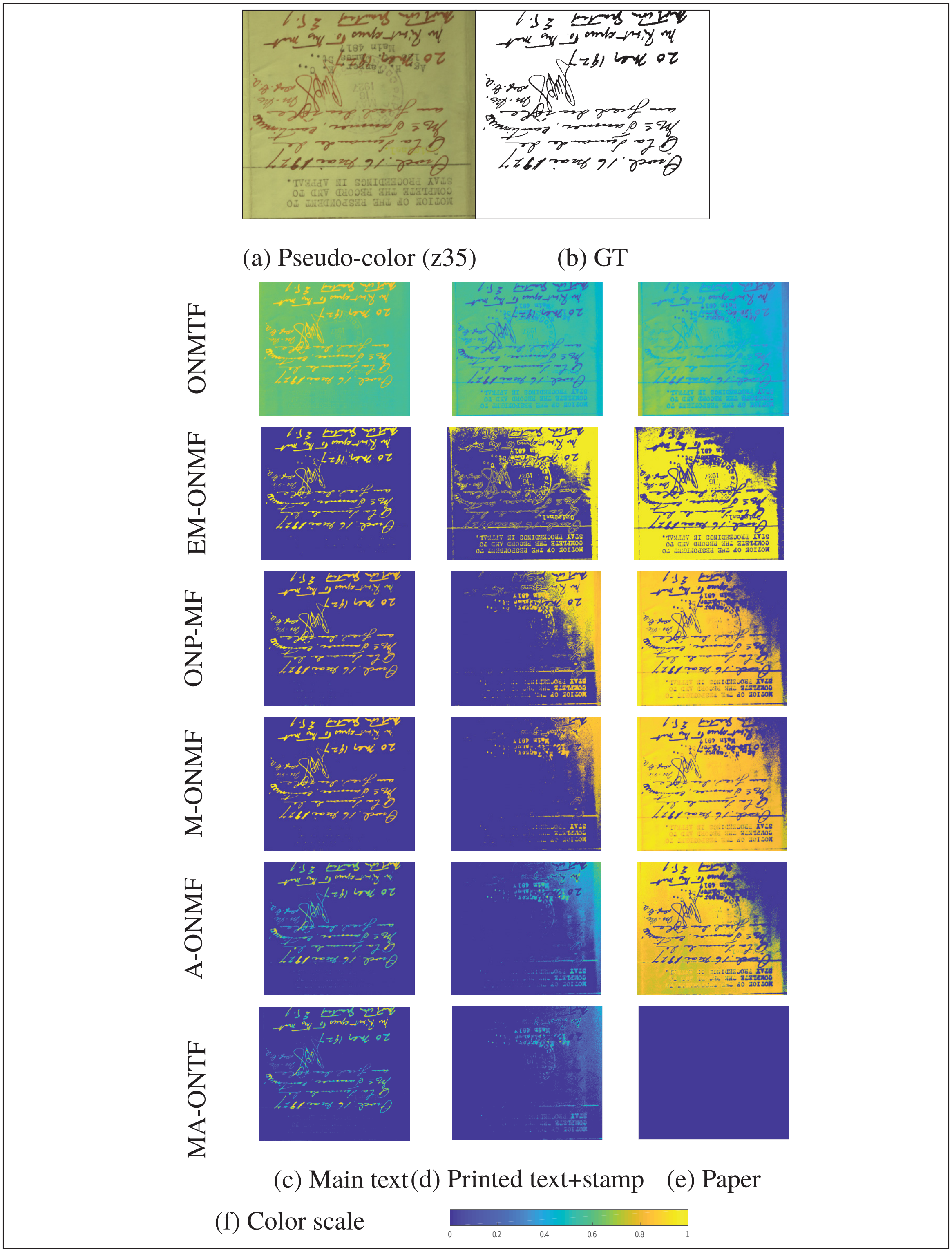
where DRD_k is the distortion of the k^{th} flipped pixel and is defined as the weighted sum of the pixels in the 5×5 block of the GT image that differs from the value flipped pixel $B(x, y)_k$ in the predicted image. $NUBN$ is defined as the number of non-uniform (not all black or white pixels) 8x8 blocks in the GT image. For further details about this metric see (Lu *et al.*, 2004).

5) ***The accuracy (ACC)*** is defined as the ratio between all correct predictions and the total number of samples, which is given by:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{A I-24}$$

4. Blind decomposition of MSTEx1 dataset

Fig. I-1 illustrates an example of the results obtained by different approaches in the decomposition task of a sample (z35) from the MSTEx-1.



(a) Pseudo-color (z35) (b) GT

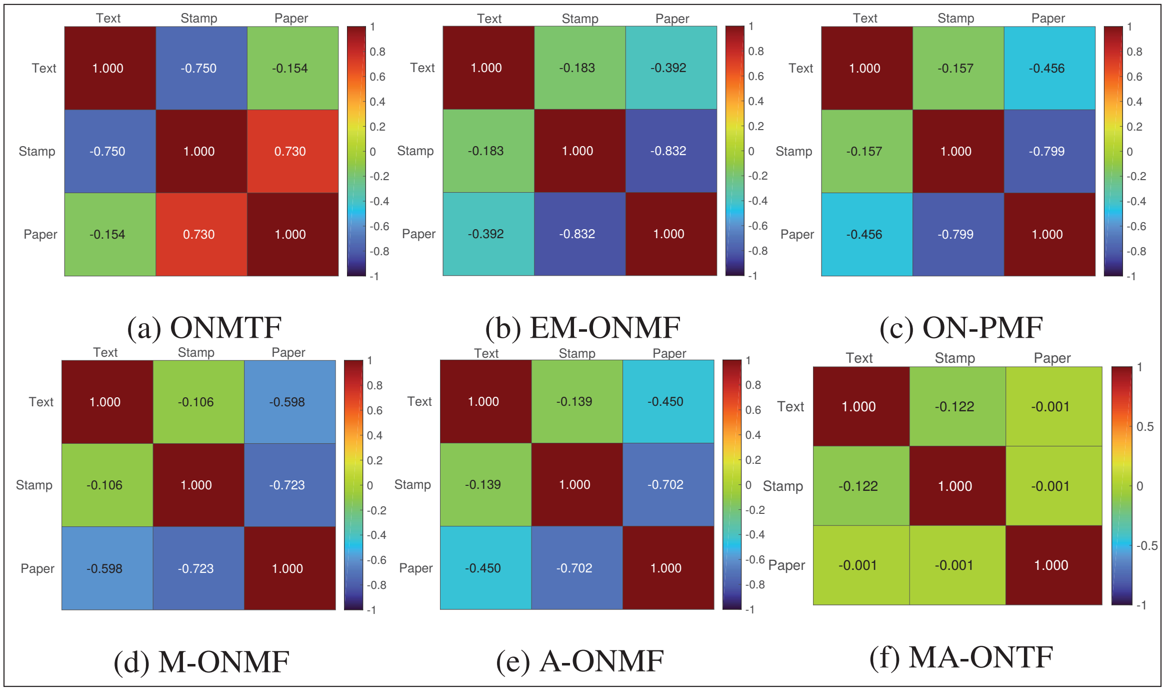
(c) Main text (d) Printed text+stamp (e) Paper

(f) Color scale

Figure-A I-1 Abundance maps of different materials decomposed by different ONMF methods applied on a sample from the MSTEx-2 dataset. (a) pseudo-colour of the image *z*35 generated from 3 channels of the visible band, (b) its corresponding GT image, (c) handwritten text, (d) stamp and printed text, and (e) paper component

In Fig. I-1, we can see that among all methods, MA-ONTF, A-ONMF, M-ONMF, and ONP-MF were able to separate the two overlapping texts (printed from handwritten). Another interesting observation is that our approaches, especially the MA-ONTF and A-ONMF models, preserve the original appearance of materials (in terms of the pixels' intensity) in the new representation with a high fidelity. For example, we can see that on the abundance map representing the ink in Fig. I-1, the pixels with the highest coefficients (intensity) values represent zones of inks with high intensities (density) in the original images and vice versa. On the contrary, EM-ONMF does not preserve the initial intensity of the pixels.

To assess the quality of the decomposition, Fig. I-2 shows the heat maps representation of the corresponding different Pearson's correlation coefficient matrices.



(a) ONMTF

(b) EM-ONMF

(c) ON-PMF

(d) M-ONMF

(e) A-ONMF

(f) MA-ONTF

Figure-A I-2 Heat maps representation of Pearson's correlation coefficient matrices. Pairwise PCC is calculated between the three abundance images corresponding to text, stamp, and paper

Similarly to the previous validation carried out on MSTEx-2, Table I-2 compares the performances of the proposed methods with EM-ONMF and ONP-MF methods for the binarization of the

MS document images of the MSTEx-1 dataset. We measure the binarization results through four metrics: F1-score, DRD, NRM, and ACC as well. The results shown in Table I-2 confirm our observations mentioned above. In general, the proposed MA-ONTF model with Howe binarization performs better than the remaining models. The EM-ONMF method, which is a k-means-like method, obtained the lowest average scores of all the metrics, although it achieved competitive results in some cases. In Table I-2 FM is expressed in %, DRD $\times 10^{-3}$, and NRM $\times 10^{-2}$. Size of MS cubes reads bands×width× height.

Table-A I-2 Binarization results of different methods on MSTEx-2 dataset. In bold and underlined, best performance; bold only, second-best scores

MS-cube (size)	Metric	RICA	Howe	SKKHM	GMM	EM-ONMF	Factorization + Threshold				Factorization + Howe				
							ONP-MF	M-ONMF	A-ONMF	MA-ONMF	ONMF	ONP-MF	M-ONMF	A-ONMF	MA-ONMF
z64 (8x1470x686)	FM	76.85	82.71	82.21	**87.28**	86.65	86.13	**86.69**	69.56	78.58	81.78	84.67	85.38	66.53	82.93
	DRD	7.07	4.48	3.62	**3.14**	3.33	3.55	**3.27**	10.58	6.57	5.01	3.76	3.47	11.99	3.92
	NRM	6.53	**4.14**	12.24	5.97	5.93	5.44	6.34	8.84	5.14	**3.52**	6.66	7.14	8.97	7.93
	ACC	96.06	95.85	97.61	**98.12**	**98.01**	97.89	98.03	94.47	96.32	96.91	97.69	97.84	93.59	97.44
z37 (8x1638x305)	FM	65.49	84.80	87.22	83.74	83.13	**88.66**	**89.12**	86.25	88.78	85.99	87.35	88.31	85.52	87.95
	DRD	7.10	4.37	2.98	3.86	3.64	2.92	**2.66**	3.76	2.82	4.02	3.17	**2.80**	3.52	3.12
	NRM	25.28	5.98	9.54	12.43	14.16	6.08	6.84	**4.47**	6.82	**4.59**	7.25	7.44	9.69	7.44
	ACC	93.19	95.66	96.89	96.17	96.21	97.02	**97.22**	96.39	**97.12**	96.00	96.71	97.03	96.39	96.76
z35 (8x690x773)	FM	67.51	70.30	91.07	78.75	91.97	**92.04**	89.72	88.99	88.85	73.32	90.06	89.51	91.01	**92.81**
	DRD	9.92	17.33	2.32	5.99	2.14	**2.12**	2.74	3.47	2.78	14.61	2.63	2.78	2.38	**1.92**
	NRM	21.06	4.55	3.73	15.06	5.03	3.52	5.96	**2.95**	8.87	4.23	5.75	6.14	4.08	**3.26**
	ACC	95.31	93.19	98.49	96.87	**98.70**	98.67	98.33	98.05	98.30	94.16	98.38	98.29	98.49	**98.80**
z80 (8x1627x523)	FM	71.24	78.97	71.61	82.45	58.55	**93.86**	80.52	90.23	82.81	79.32	**91.43**	88.98	89.55	88.28
	DRD	14.71	10.18	10.51	4.63	15.38	**1.76**	10.17	2.73	4.56	10.04	**2.45**	3.62	3.01	3.19
	NRM	8.88	4.91	15.87	14.45	25.29	**3.89**	**3.34**	7.31	13.46	4.86	5.54	4.54	6.93	7.67
	ACC	93.98	95.73	95.23	97.42	93.77	**98.97**	95.99	98.43	97.41	95.82	**98.57**	98.06	98.28	98.18
z38 (8x1603x264)	FM	58.25	83.74	82.17	55.26	87.48	87.51	**87.63**	**87.76**	86.94	84.89	85.96	87.34	86.65	86.53
	DRD	8.24	3.91	3.92	14.36	**2.92**	**2.93**	3.08	2.99	2.96	4.13	3.25	3.05	3.34	3.39
	NRM	29.25	**5.08**	13.35	26.04	7.36	7.15	**5.71**	6.26	8.56	5.92	8.58	6.34	6.38	5.98
	ACC	91.65	95.64	95.50	87.55	**96.52**	**96.51**	96.40	96.47	96.46	95.42	96.13	96.36	96.13	96.04
z68 (8x1506x448)	FM	50.26	84.24	81.09	83.49	83.58	82.77	**87.59**	**87.78**	78.49	83.51	81.05	86.31	85.73	85.07
	DRD	21.28	4.26	6.16	3.61	4.99	5.39	**2.66**	**2.79**	5.60	4.63	5.72	2.90	3.34	3.91
	NRM	20.18	**2.50**	3.46	9.96	3.53	3.35	7.55	6.28	12.66	**2.44**	5.16	8.20	6.70	3.30
	ACC	92.45	97.89	97.41	98.14	97.83	97.65	**98.59**	**98.57**	97.58	97.76	97.50	98.44	98.30	98.05
z70 (8x1532x448)	FM	53.20	79.86	74.57	80.99	79.99	79.39	**82.28**	**81.35**	79.44	79.49	78.07	80.44	78.62	79.44
	DRD	13.26	5.01	7.55	5.19	5.77	6.01	4.35	**4.02**	4.86	5.53	6.18	4.54	4.74	**4.34**
	NRM	23.15	7.66	8.09	**3.56**	6.04	**5.64**	7.78	11.13	9.26	6.08	6.96	9.43	11.18	10.10
	ACC	94.66	97.68	96.86	97.62	97.60	97.48	**98.03**	**98.08**	97.71	97.53	97.36	97.86	97.71	97.76
z76 (8x1319x748)	FM	44.70	86.46	85.16	**87.02**	86.42	84.96	86.32	82.13	86.44	85.45	85.55	86.69	82.59	**87.90**
	DRD	51.85	5.10	5.13	**3.87**	4.58	5.22	3.95	6.38	3.92	5.11	4.95	**3.82**	6.17	3.96
	NRM	19.07	**3.70**	6.48	8.05	6.25	6.78	9.33	8.47	9.25	4.25	6.65	9.02	8.28	**4.22**
	ACC	73.42	96.44	96.29	**96.99**	96.66	96.26	96.91	95.56	96.93	96.16	96.43	**96.98**	95.68	96.93
z82 (8x1892x583)	FM	78.06	82.26	80.57	77.74	84.72	**85.12**	**84.88**	75.44	82.49	82.27	83.88	84.13	73.84	84.66
	DRD	4.01	4.03	3.58	5.75	**2.78**	3.28	3.06	3.94	**2.91**	4.12	3.47	3.35	4.17	3.30
	NRM	15.94	7.23	13.65	9.60	10.81	8.03	9.51	19.46	12.36	6.80	8.92	**6.50**	20.39	**6.38**
	ACC	98.24	98.29	98.38	97.84	**98.70**	98.64	**98.67**	98.17	98.63	98.27	98.54	98.48	98.07	98.53
z58 (8x1435x269)	FM	68.12	80.21	69.79	84.01	82.12	80.54	**85.78**	**85.71**	85.57	82.03	78.91	84.28	85.19	85.41
	DRD	7.59	4.98	10.05	3.75	4.84	5.65	3.01	**2.89**	3.26	4.32	5.79	3.01	**2.88**	3.33
	NRM	17.97	7.56	9.70	5.98	6.57	**5.92**	8.19	8.91	7.30	6.68	7.86	10.02	9.31	**4.80**
	ACC	97.33	98.14	96.86	98.52	98.33	98.11	98.78	**98.80**	98.73	98.33	98.00	98.69	**98.76**	98.64
z100 (8x386x1066)	FM	51.25	84.60	73.01	77.24	**88.95**	87.56	86.83	87.17	**88.86**	84.68	85.90	86.45	86.09	88.57
	DRD	12.36	6.12	4.49	6.59	3.90	4.52	3.76	4.08	**3.73**	6.15	4.77	3.79	4.39	**3.63**
	NRM	31.45	4.25	17.58	16.57	**4.22**	**4.22**	9.24	8.07	5.28	**3.86**	6.42	9.62	8.28	6.10
	ACC	93.61	96.98	97.07	96.49	97.96	97.66	97.81	97.80	**98.00**	96.97	97.43	97.75	97.59	**97.98**
Average	FM	62.27	81.83	79.86	79.82	83.05	**86.23**	86.12	83.85	84.30	82.07	84.80	86.17	82.85	**86.32**
	DRD	14.31	6.34	5.48	5.52	4.93	3.94	3.88	4.33	4.01	6.15	4.19	**3.38**	4.54	**3.46**
	NRM	19.89	**5.23**	10.34	11.61	8.65	5.46	7.25	8.38	9.00	**4.84**	6.89	7.67	9.11	6.11
	ACC	92.72	96.50	96.96	96.52	97.30	97.71	97.70	97.34	97.56	96.67	97.52	**97.80**	97.18	**97.74**

5. Hyperparametrs settings

The values of different hyperparameters used for each sample in our work are given in Table I-3.

Table-A I-3 Hyperparameters settings for each sample from the three datasets used in this work

Dataset		M-ONMF			A-ONMF			MA-ONTF				ONP-MF				SKKHM		
		ρ_x	ρ_z	thresh	ρ_u	ρ_v	thresh	ρ_f	ρ_g	ρ_h	thresh	c	α	ρ	thresh	Q	σ	β
Hubble		10^{-5}	10^5	-	10^5	10^5	-	10^{-3}	10^5	10^{-3}	-	1.01	10^2	10^{-2}	-	3×3	500	0.5
MS-TEx2	z27	10^4	5.10^4	0.15	5.10^5	5.10^5	0.01	10	10^2	10^3	0.15	1.01	10^2	10^{-2}	0.01	3×3	100	0.5
	z31	10^4	5.10^4	0.36	10^3	10^4	0.01	10^3	5.10^3	5.10^3	0.05	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
	z43	10^4	5.10^4	0.36	10^3	10^4	0.01	10^3	10^3	10^4	0.01	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
	z58-2	5.10^3	5.10^4	0.40	10^3	10^3	0.01	10	10^2	10^2	0.30	1.01	10^{-2}	10^2	0.01	3×3	500	0.5
	z59-2	10^3	5.10^5	0.40	10^3	10^3	0.01	10^4	10^3	10^3	0.01	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
	z65	10^5	10^5	0.26	10^3	10^3	0.01	5.10^3	10^3	10^3	0.01	1.01	10^2	10^{-2}	0.01	3×3	1000	0.5
	z80-2	10^3	10^5	0.35	5.10^3	5.10^3	0.01	10^3	5.10^3	10^3	0.16	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
	z82-2	5.10^4	5.10^4	0.35	10^3	10^4	0.01	10^4	10^3	10^3	0.51	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
	z90	10^3	5.10^4	0.01	10^3	10^4	0.01	5.10^3	10^3	5.10^3	0.24	1.01	10^2	10^{-2}	0.01	3×3	1000	0.5
	z92	5.10^3	5.10^4	0.45	10^3	5.10^4	0.01	10^{-1}	10	10^3	0.16	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
MS-TEx1	z64	10^4	10^5	0.45	10^3	10^4	0.01	10^4	10^3	10^3	0.01	1.01	10^{-2}	10^2	0.01	3×3	500	0.5
	z37	10^4	10^5	0.45	5.10^5	5.10^5	0.15	10^4	5.10^3	10^4	0.15	1.01	10^{-2}	10^2	0.01	3×3	100	0.5
	z35	5.10^3	10^4	0.27	10^4	10^5	0.01	10^3	10^3	10^3	0.10	1.01	10^{-2}	10^2	0.01	3×3	100	0.5
	z80	10^4	10^5	0.35	10^3	5.10^3	0.10	10^4	10^4	10^4	0.0	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
	z38	10^3	10^5	0.35	10^5	10^5	0.10	5.10^3	10^3	5.10^3	0.15	1.01	10^{-2}	10^2	0.01	3×3	500	0.5
	z68	10^5	10^5	0.5	10^3	5.10^3	0.10	10^3	10^4	10^3	0.15	1.01	10^{-2}	10^2	0.01	3×3	100	0.5
	z76	10^3	10^5	0.45	10^3	10^5	0.50	$10^4 3$	10^4	10^3	0.10	1.01	10^2	10^{-2}	0.01	3×3	500	0.5
	z58	10^4	10^5	0.35	10^3	10^5	0.58	10	10^2	10^2	0.30	1.01	10^{-2}	10^2	0.01	3×3	500	0.5
	z100	10^4	10^5	0.20	10^3	5.10^5	0.20	10	10	10	0.35	1.01	10^{-2}	10^2	0.01	3×3	500	0.5
	z70	10^3	10^5	0.35	10^3	5.10^3	0.20	10	10^2	10^2	0.30	1.01	10^{-2}	10^2	0.01	3×3	500	0.5
	z82	10^4	10^4	0.35	10^3	5.10^3	0.20	10	10^2	10^2	0.30	1.01	10^{-2}	10^2	0.01	3×3	200	0.5

APPENDIX II

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

This Appendix is adapted from the Supplementary Material of our paper of Chapter 4. It provides additional details and results to the main paper that were omitted there due to space restrictions. All citations refer here to the References section of the main document.

1. Projected Gradient Descent algorithm

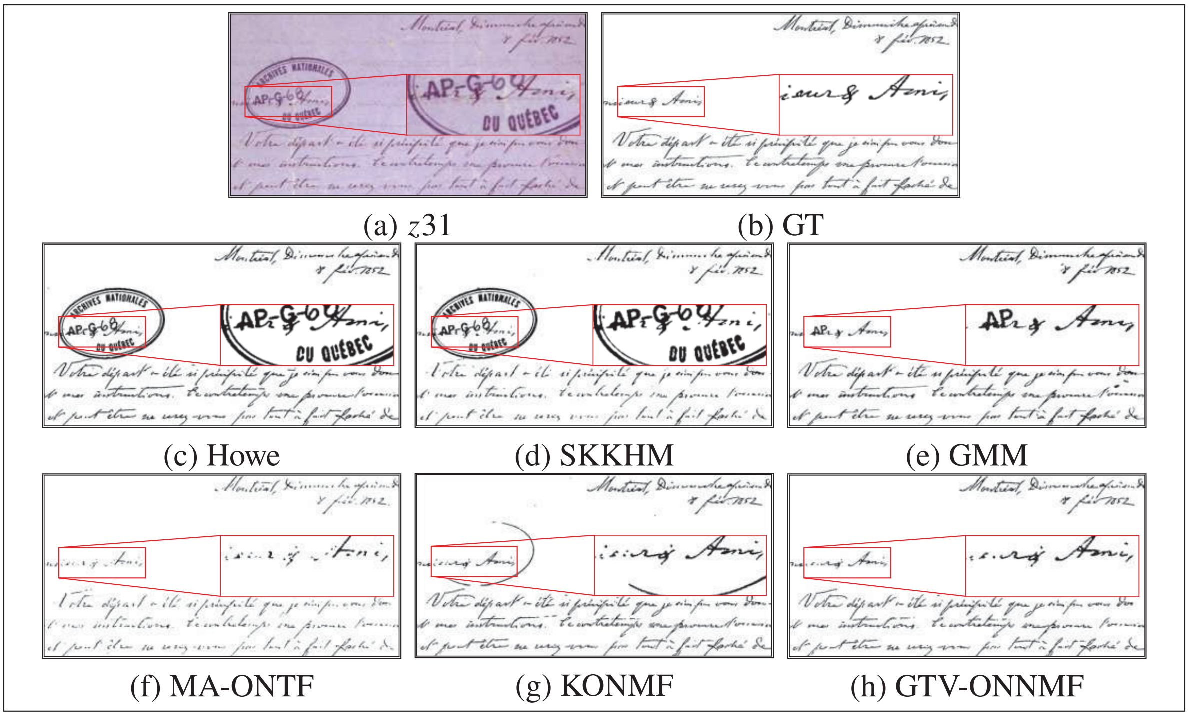
The PGD algorithm adopted to update the variable **M** of the is given by (adapted from (Lin, 2007)):

Algorithm-A II-1 Projected gradient algorithm

```
Input: $\mathbf{M}, \mathbf{X}, \mathbf{J}, \mathbf{B}, \mathbf{R}, maxiter$
Initialization: $0 < \beta < 1, 0 < \epsilon < 1$
for $t = 1$ to maxiter do
    Calculate $\nabla_M \mathcal{L} = \frac{1}{\sigma^2}(\mathbf{M} \odot (\mathbf{JB}) - \mathbf{XB}) + 4\mathbf{M}(\mathrm{Diag}(\mathbf{R}\mathbb{1}) - \mathbf{R})$.
    if $\alpha_t$ satifies condition (4.14) then
        $\alpha_t \leftarrow \alpha_t / \beta$
    else
        $\alpha_t \leftarrow \alpha_t \times \beta$
    end if
    $\mathbf{M}^{t+1} = Proj[\mathbf{M}^t - \alpha_t \nabla_M \mathcal{L}(\mathbf{M}^t)]$
end for
return M
```

2. Binarization results of a sample from the MSTEx-2 dataset

A visual comparison between the results of different methods for the binarization of the MSTEx-2 dataset is illustrated in Fig. II-1. It can be seen that our GTV-ONNMF model provides a result (h) that is comparable to the GT image (b).



(a) $z31$

(b) GT

(c) Howe

(d) SKKHM

(e) GMM

(f) MA-ONTF

(g) KONMF

(h) GTV-ONNMF

Figure-A II-1 Qualitative comparison of the binarization outputs of different approaches on the sample $z31$ from the MSTEx-1 dataset

APPENDIX III

DESCRIPTION OF EVALUATION METRICS USED IN CHAPTER 5

1. PSNR metric

It is worth noting that the definitions and formulas of FM, DRD, and NRM metrics are already given in Appendix I. Therefore, to avoid redundancy, we describe here only the PSNR metric.

The Peak Signal to Noise Ratio (PSNR) is defined as the ratio between the maximum possible value of a signal (here a pixel image) and the power of the corrupting noise:

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right), \tag{A III-1}$$

where MAX_I is the maximum possible pixel value of an image I, and

$$MSE = \frac{\sum\limits_{m,n}^{M,N} (I_1(m,n) - I_2(m,n))^2}{M \times N}$$

is the mean square error between two images I_1 and I_2.

BIBLIOGRAPHY

Absil, P.-A., Mahony, R. & Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Aldrovandi, A., Bertani, D., Cetica, M., Matteini, M., Moles, A., Poggi, P. & Tiano, P. (1988). Multispectral image processing of paintings. *Studies in Conservation*, 33(3), 154–159.

An, S., Yun, J.-M. & Choi, S. (2011). Multiple kernel nonnegative matrix factorization. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1976–1979.

Arai, K. (2008). Nonlinear mixture model of mixed pixels in remote sensing satellite images based on Monte Carlo simulation. *Advances in Space Research*, 41(11), 1715–1723.

Balas, C., Papadakis, V., Papadakis, N., Papadakis, A., Vazgiouraki, E. & Themelis, G. (2003). A novel hyper-spectral imaging apparatus for the non-destructive analysis of objects of artistic and historic value. *Journal of Cultural Heritage*, 4, 330–337.

Baronti, S., Casini, A., Lotti, F. & Porcinai, S. (1998). Multispectral imaging system for the mapping of pigments in works of art by use of principal-component analysis. *Applied optics*, 37(8), 1299–1309.

Bateson, C. A., Asner, G. P. & Wessman, C. A. (2000). Endmember bundles: A new approach to incorporating endmember variability into spectral mixture analysis. *IEEE transactions on geoscience and remote sensing*, 38(2), 1083–1094.

Berger, P., Hannak, G. & Matz, G. (2017). Graph signal recovery via primal-dual algorithms for total variation minimization. *IEEE Journal of Selected Topics in Signal Processing*, 11(6), 842–855.

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3), 334–334.

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P. & Chanussot, J. (2012). Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2), 354–379.

Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N. & Chanussot, J. (2013). Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and remote sensing magazine*, 1(2), 6–36.

Boslaugh, S. & Watters, P. A. (2008). *Statistics in a nutshell: A desktop quick reference*. O'Reilly Media, Inc.

Boukouvalas, Z., Levin-Schwartz, Y. & Adalı, T. (2017). Enhancing ICA performance by exploiting sparsity: Application to FMRI analysis. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 2532–2536.

Boutsidis, C. & Gallopoulos, E. (2008). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, 41(4), 1350–1362.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1–122.

Broadwater, J. & Banerjee, A. (2009). A comparison of kernel functions for intimate mixture models. *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pp. 1–4.

Brouwer, T. & Lió, P. (2017). Prior and Likelihood Choices for Bayesian Matrix Factorisation on Small Datasets. *stat*, 1050, 1.

Brouwer, T., Frellsen, J. & Lio, P. (2016). Fast Bayesian non-negative matrix factorisation and tri-factorisation. *Advances in Approximate Bayesian Inference, NeurIPS 2016 workshop, 30th Conference on Neural Information Processing Systems (NeurIPS 2016), December 9, 2016, Barcelona, Spain.*, 1–17.

Brouwer, T., Frellsen, J. & Lió, P. (2017). Comparative study of inference methods for bayesian nonnegative matrix factorisation. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 513–529.

Bruna, J., Sprechmann, P. & LeCun, Y. (2015). Source separation with scattering non-negative matrix factorization. *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 1876–1880.

Buciu, I., Nikolaidis, N. & Pitas, I. (2008). Nonnegative matrix factorization in polynomial feature space. *IEEE Transactions on Neural Networks*, 19(6), 1090–1100.

Burmester, A., Cupitt, J., Derrien, H., Dessipris, N., Hamber, A., Martinez, K., Müller, M. & Saunders, D. (1992). The examination of paintings by digital image analysis. *Proc. 3rd Int. Conf. Non-Destructive Testing, Microanalytical Methods and Environmental Evaluation for Study and Conservation of Works of Art*, pp. 201–214.

Cai, D., He, X., Han, J. & Huang, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8), 1548–1560.

Calvo-Zaragoza, J. & Gallego, A.-J. (2019). A selectional auto-encoder approach for document image binarization. *Pattern Recognition*, 86, 37–47.

Charlier, B., Feydy, J., Glaunès, J., Collin, F.-D. & Durif, G. (2021). Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74), 1–6.

Cheriet, M. & Moghaddam, R. F. (2008). Processing of low quality document images: issues and directions. *2008 16th European Signal Processing Conference*, pp. 1–5.

Chikuse, Y. (1990). Distributions of orientations on Stiefel manifolds. *Journal of multivariate analysis*, 33(2), 247–264.

Choi, S. (2008). Algorithms for orthogonal nonnegative matrix factorization. *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)*, pp. 1828–1832.

Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S.-i. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory MultiWay Data Analysis and Blind Source Separation*. John Wiley & Sons.

Comon, P. & Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Cosentino, A. (2014). Identification of pigments by multispectral imaging; a flowchart method. *Heritage Science*, 2(1), 8.

Diem, M., Hollaus, F. & Sablatnig, R. (2016). MSIO: MultiSpectral Document Image BinarizatIOn. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pp. 84–89.

Ding, C., Li, T., Peng, W. & Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 126–135.

Dobigeon, N. & Tourneret, J.-Y. (2010). Bayesian orthogonal component analysis for sparse representation. *IEEE Transactions on Signal Processing*, 58(5), 2675–2685.

Donoho, D. & Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? *Proceedings of the 16th International Conference on Neural Information Processing Systems*, 16, 1141–1148.

Easton, R. L. & Kelbe, D. (2014). Statistical processing of spectral imagery to recover writings from erased or damaged manuscripts. *manuscript cultures*, 7, 35–46.

Easton, R. L., Christens-Barry, W. A. & Knox, K. T. (2011). Spectral image processing and analysis of the Archimedes Palimpsest. *Signal Processing Conference, 2011 19th European*, pp. 1440–1444.

Eckstein, J. & Yao, W. (2015). Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pac. J. Optim.*, 11(4), 619–644.

Edelman, G., Gaston, E., Van Leeuwen, T., Cullen, P. & Aalders, M. (2012). Hyperspectral imaging for non-contact analysis of forensic traces. *Forensic science international*, 223(1-3), 28–39.

Feydy, J., Glaunès, J., Charlier, B. & Bronstein, M. (2020). Fast geometric learning with symbolic matrices. *Thirty-fourth Conference on Neural Information Processing Systems*.

Fischer, C. & Kakoulli, I. (2006). Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications. *Studies in Conservation*, 51(sup1), 3–16.

Garini, Y., Young, I. T. & McNamara, G. (2006). Spectral imaging: principles and applications. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 69(8), 735–747.

George, S. & Hardeberg, J. Y. (2015). Ink classification and visualisation of historical manuscripts: Application of hyperspectral imaging. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1131–1135.

Giacometti, A., Campagnolo, A., MacDonald, L., Mahony, S., Robson, S., Weyrich, T., Terras, M. & Gibson, A. (2017). The value of critical destruction: Evaluating multispectral image processing methods for the analysis of primary historical texts. *Digital Scholarship in the Humanities*, 32(1), 101–122.

Gillis, N. (2012). Sparse and unique nonnegative matrix factorization through data preprocessing. *The Journal of Machine Learning Research*, 13(1), 3349–3386.

Gillis, N. (2014). The Why and How of Nonnegative Matrix Factorization. *stat*, 1050, 7.

Gobinet, C., Vrabie, V., Tfayli, A., Piot, O., Huez, R. & Manfait, M. (2007). Pre-processing and source separation methods for Raman spectra analysis of biomedical samples. *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6207–6210.

Goetz, A. F., Vane, G., Solomon, J. E. & Rock, B. N. (1985). Imaging spectrometry for earth remote sensing. *science*, 228(4704), 1147–1153.

Grabowski, B., Masarczyk, W., Głomb, P. & Mendys, A. (2018). Automatic pigment identification from hyperspectral data. *Journal of Cultural Heritage*, 31, 1–12.

Gupta, A. & Nagar, D. (1999). Matrix variate distributions. PMS series. Addison-Wesley Longman, Limited Reading.

Gutierrez-Navarro, O., Campos-Delgado, D. U., Arce-Santana, E., Mendez, M. & Jo, J. A. (2013). Blind Decomposition of Multi-spectral Fluorescence Lifetime Imaging Microscopy Data: Further Validation. *Procedia Technology*, 7, 118–125.

Halimi, A., Altmann, Y., Dobigeon, N. & Tourneret, J.-Y. (2011). Nonlinear unmixing of hyperspectral images using a generalized bilinear model. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11), 4153–4162.

Hanif, M., Tonazzini, A., Savino, P., Salerno, E. & Tsagkatakis, G. (2018). Document Bleed-Through Removal Using Sparse Image Inpainting. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 281–286.

Hedjam, R. (2013). *Visual image processing in various representation spaces for documentary preservation*. (Ph.D. thesis, École de technologie supérieure).

Hedjam, R. & Cheriet, M. (2013a). Ground-truth estimation in multispectral representation space: Application to degraded document image binarization. *2013 12th International Conference on Document Analysis and Recognition*, pp. 190–194.

Hedjam, R. & Cheriet, M. (2013b). Historical document image restoration using multispectral imaging system. *Pattern Recognition*, 46(8), 2297–2312.

Hedjam, R., Nafchi, H. Z., Moghaddam, R. F., Kalacska, M. & Cheriet, M. (2015). ICDAR 2015 contest on MultiSpectral text extraction (MS-TEx 2015). *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 1181–1185.

Heinz, D. C. et al. (2001). Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE transactions on geoscience and remote sensing*, 39(3), 529–545.

Heylen, R., Burazerovic, D. & Scheunders, P. (2010). Non-linear spectral unmixing by geodesic simplex volume maximization. *IEEE Journal of Selected Topics in Signal Processing*, 5(3), 534–542.

Hinrich, J. L. & Mørup, M. (2018). Probabilistic sparse non-negative matrix factorization. *International Conference on Latent Variable Analysis and Signal Separation*, pp. 488–498.

Hinrich, J. L. & Mørup, M. (2018). Probabilistic sparse non-negative matrix factorization. *International Conference on Latent Variable Analysis and Signal Separation*, pp. 488–498.

Hoff, P. D. (2007). Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478), 674–685.

Hollaus, F., Gau, M. & Sablatnig, R. (2012). Multispectral image acquisition of ancient manuscripts. *Euro-Mediterranean Conference*, pp. 30–39.

Hollaus, F., Diem, M. & Sablatnig, R. (2015). Binarization of multispectral document images. *International Conference on Computer Analysis of Images and Patterns*, pp. 109–120.

Hollaus, F., Diem, M. & Sablatnig, R. (2018a, 08). MultiSpectral Image Binarization using GMMs. *ICFHR 2018, The 16th International Conference on Frontiers in Handwriting Recognition*, pp. 570-575. doi: 10.1109/ICFHR-2018.2018.00105.

Hollaus, F., Diem, M. & Sablatnig, R. (2018b). MultiSpectral Image Binarization using GMMs. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 570–575.

Hollaus, F., Brenner, S. & Sablatnig, R. (2019). CNN based binarization of multispectral document images. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 533–538.

Howe, N. R. (2013). Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(3), 247–258.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov), 1457–1469.

Huang, K., Sidiropoulos, N. D. & Swami, A. (2013). Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1), 211–224.

Hütter, J.-C. & Rigollet, P. (2016). Optimal rates for total variation denoising. *Conference on Learning Theory*, pp. 1115–1146.

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5), 411–430.

Iordache, M.-D., Bioucas-Dias, J. M. & Plaza, A. (2011). Sparse unmixing of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(6), 2014–2039.

Jain, A. K. & Yu, B. (1998). Document representation and its application to page decomposition. *IEEE Transactions on pattern analysis and machine intelligence*, 20(3), 294–308.

Jauch, M., Hoff, P. D. & Dunson, D. B. (2021). Monte Carlo simulation on the Stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics*, 1–10.

Karoui, M. S., Deville, Y., Hosseini, S. & Ouamri, A. (2012). Blind spatial unmixing of multispectral images: New methods combining sparse component analysis, clustering and non-negativity constraints. *Pattern Recognition*, 45(12), 4263–4278.

Kendall, A. & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems*, 30, 5574–5584.

Khan, Z., Shafait, F. & Mian, A. (2013). Hyperspectral imaging for ink mismatch detection. *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pp. 877–881.

Khan, Z., Shafait, F. & Mian, A. (2015). Automatic ink mismatch detection for forensic document analysis. *Pattern Recognition*, 48(11), 3615–3626.

Khatri, C. & Mardia, K. V. (1977). The von Mises–Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 95–106.

Kopriva, I. & Cichocki, A. (2009a). Blind decomposition of low-dimensional multi-spectral image by sparse component analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(11), 590–597.

Kopriva, I. & Cichocki, A. (2009b). Blind multispectral image decomposition by 3D nonnegative tensor factorization. *Optics letters*, 34(14), 2210–2212.

Kuang, D., Ding, C. & Park, H. (2012). Symmetric nonnegative matrix factorization for graph clustering. *Proceedings of the 2012 SIAM international conference on data mining*, pp. 106–117.

Kwok, J.-Y. & Tsang, I.-H. (2004). The pre-image problem in kernel methods. *IEEE transactions on neural networks*, 15(6), 1517–1525.

Lai, R. & Osher, S. (2014). A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2), 431–449.

Landgrebe, D. (1999). Information extraction principles and methods for multispectral and hyperspectral image data. In *Information Processing For Remote Sensing* (pp. 3–37). World Scientific.

Laurberg, H. (2007). Uniqueness of non-negative matrix factorization. *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pp. 44–48.

Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K. & Jensen, S. H. (2008). Theorems on positive data: On the uniqueness of NMF. *Computational intelligence and neuroscience*, 2008, 764206.

Le, Q., Karpenko, A., Ngiam, J. & Ng, A. (2011). ICA with reconstruction cost for efficient overcomplete feature learning. *Advances in neural information processing systems*, 24, 1017–1025.

Le, V. P., Nayef, N., Visani, M., Ogier, J.-M. & De Tran, C. (2015). Text and non-text segmentation based on connected component features. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1096–1100.

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791.

Legnaioli, S., Grifoni, E., Lorenzetti, G., Marras, L., Pardini, L., Palleschi, V., Salerno, E. & Tonazzini, A. (2013). Enhancement of hidden patterns in paintings using statistical analysis. *Journal of Cultural Heritage*, 14(3), S66–S70.

Lettner, M. & Sablatnig, R. (2009). Spatial and spectral based segmentation of text in multispectral images of ancient documents. *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pp. 813–817.

Lettner, M. & Sablatnig, R. (2010). Higher order MRF for foreground-background separation in multi-spectral images of historical manuscripts. *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 317–324.

Li, B., Zhou, G. & Cichocki, A. (2014). Two efficient algorithms for approximately orthogonal nonnegative matrix factorization. *IEEE Signal Processing Letters*, 22(7), 843–846.

Li, Q., Mitianoudis, N. & Stathaki, T. (2007). Spatial kernel K-harmonic means clustering for multi-spectral image segmentation. *IET Image Processing*, 1(2), 156–167.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10), 2756–2779.

Lin, C.-J. & Jorge, J. (1999). Moré. Newton's method for large-scale bound constrained problems. *SIAM Journal on Optimization*, 9(1100-1127), 10–1137.

Lin, L., Rao, V. & Dunson, D. (2017). Bayesian nonparametric inference on the Stiefel manifold. *Statistica Sinica*, 535–553.

Lu, H., Kot, A. C. & Shi, Y. Q. (2004). Distance-reciprocal distortion measure for binary document images. *IEEE Signal Processing Letters*, 11(2), 228–231.

Lyu, H., Liao, N., Li, H. & Wu, W. (2016). High resolution ultraviolet imaging spectrometer for latent image analysis. *Optics express*, 24(6), 6459–6468.

Melessanaki, K., Papadakis, V., Balas, C. & Anglos, D. (2001). Laser induced breakdown spectroscopy and hyper-spectral imaging analysis of pigments on an illuminated manuscript. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 56(12), 2337–2346.

Minc, H. (1988). *Nonnegative matrices*. Wiley.

Mitianoudis, N. & Papamarkos, N. (2014). Multi-spectral document image binarization using image fusion and background subtraction techniques. *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 5172–5176.

Moghaddam, R. F. & Cheriet, M. (2015). A multiple-expert binarization framework for multispectral images. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 321–325.

Morales, A., Ferrer, M. A., Diaz-Cabrera, M., Carmona, C. & Thomas, G. L. (2014). The use of hyperspectral analysis for ink identification in handwritten documents. *2014 International Carnahan Conference on Security Technology (ICCST)*, pp. 1–5.

Moussaoui, S., Brie, D., Mohammad-Djafari, A. & Carteret, C. (2006). Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. *IEEE transactions on signal processing*, 54(11), 4133–4145.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory.*

Muja, M. & Lowe, D. G. (2014). Scalable nearest neighbor algorithms for high dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11), 2227–2240.

Nafchi, H. Z., Moghaddam, R. F. & Cheriet, M. (2014). Phase-based binarization of ancient document images: Model and applications. *IEEE transactions on image processing*, 23(7), 2916–2930.

Nuzillard, D. & Bijaoui, A. (2000). Blind source separation and analysis of multispectral astronomical images. *Astronomy and Astrophysics Supplement Series*, 147(1), 129–138.

Oliveira, S. A., Seguin, B. & Kaplan, F. (2018). dhSegment: A generic deep-learning approach for document segmentation. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 7–12.

Ouji, A., Leydier, Y. & LeBourgeois, F. (2013). A hierarchical and scalable model for contemporary document image segmentation. *Pattern Analysis and Applications*, 16(4), 679–693.

Paatero, P. & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111–126.

Pal, S., Sengupta, S., Mitra, R. & Banerjee, A. (2020). Conjugate Priors and Posterior Inference for the Matrix Langevin Distribution on the Stiefel Manifold. *Bayesian Analysis*, 15(3), 871–908.

Pan, B., Lai, J. & Chen, W.-S. (2011). Nonlinear nonnegative matrix factorization based on Mercer kernel construction. *Pattern Recognition*, 44(10-11), 2800–2810.

Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D. & Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE transactions on pattern analysis and machine intelligence*, 28(3), 403–415.

Pauca, V. P., Piper, J. & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, 416(1), 29–47.

Pei, X., Wu, T. & Chen, C. (2014). Automated graph regularized projective nonnegative matrix factorization for document clustering. *IEEE transactions on cybernetics*, 44(10), 1821–1831.

Pompili, F., Gillis, N., Absil, P.-A. & Glineur, F. (2014). Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. *Neurocomputing*, 141, 15–25.

Pratikakis, I., Gatos, B. & Ntirogiannis, K. (2012). ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). *2012 international conference on frontiers in handwriting recognition*, pp. 817–822.

Pratikakis, I., Zagoris, K., Karagiannis, X., Tsochatzidis, L., Mondal, T. & Marthot-Santaniello, I. (2019). ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019). *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1547-1556.

Psorakis, I., Roberts, S., Ebden, M. & Sheldon, B. (2011). Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, 83(6), 066-114.

Rahiche, A. & Cheriet, M. (2020). Forgery Detection in Hyperspectral Document Images Using Graph Orthogonal Nonnegative Matrix Factorization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 662–663.

Rahiche, A. & Cheriet, M. (2021a). Blind decomposition of multispectral document images using orthogonal nonnegative matrix factorization. *IEEE Transactions on Image Processing*, 30, 5997–6012.

Rahiche, A. & Cheriet, M. (2021b). Kernel Orthogonal Nonnegative Matrix Factorization: Application to Multispectral Document Image Decomposition. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3275–3279.

Rahiche, A., Hedjam, R., Al-maadeed, S. & Cheriet, M. (2020). Historical documents dating using multispectral imaging and ordinal classification. *Journal of Cultural Heritage*, 45, 71–80.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton university press.

Salakhutdinov, R. & Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. *Proceedings of the 25th international conference on Machine learning*, pp. 880–887.

Salerno, E., Tonazzini, A. & Bedini, L. (2007). Digital image analysis to enhance underwritten text in the Archimedes palimpsest. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4), 79–87.

Sauvola, J. & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition*, 33(2), 225–236.

Schmidt, M. N. & Mohamed, S. (2009). Probabilistic non-negative tensor factorization using markov chain monte carlo. *Signal Processing Conference, 2009 17th European*, pp. 1918–1922.

Schmidt, M. N., Winther, O. & Hansen, L. K. (2009). Bayesian non-negative matrix factorization. *International Conference on Independent Component Analysis and Signal Separation*, pp. 540–547.

Settle, J. & Drake, N. (1993). Linear mixing and the estimation of ground cover proportions. *International Journal of Remote Sensing*, 14(6), 1159–1177.

Shaw, G. A. & Burke, H. K. (2003). Spectral imaging for remote sensing. *Lincoln laboratory journal*, 14(1), 3–28.

Silva, C. S., Pimentel, M. F., Honorato, R. S., Pasquini, C., Prats-Montalbán, J. M. & Ferrer, A. (2014). Near infrared hyperspectral imaging for forensic analysis of document forgery. *Analyst*, 139(20), 5176–5184.

Šmídl, V. & Quinn, A. (2007). On Bayesian principal component analysis. *Computational statistics & data analysis*, 51(9), 4101–4123.

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.

Squires, S., Prügel-Bennett, A. & Niranjan, M. (2017). Rank selection in nonnegative matrix factorization using minimum description length. *Neural computation*, 29(8), 2164–2176.

Tagare, H. D. (2011). Notes on optimization on stiefel manifolds. In *Technical report, Technical report*. Yale University.

Tan, V. Y. & Févotte, C. (2012). Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7), 1592–1605.

Tichỳ, O. & Šmídl, V. (2015). Bayesian Blind Source Separation with Unknown Prior Covariance. *International Conference on Latent Variable Analysis and Signal Separation*, pp. 352–359.

Tichỳ, O., Bódiová, L. & Šmídl, V. (2019). Bayesian Non-Negative Matrix Factorization With Adaptive Sparsity and Smoothness Prior. *IEEE Signal Processing Letters*, 26(3), 510–514.

Tolić, D., Antulov-Fantulin, N. & Kopriva, I. (2018). A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering. *Pattern Recognition*, 82, 40–55.

Tonazzini, A., Bedini, L. & Salerno, E. (2004a). Independent component analysis for document restoration. *Document Analysis and Recognition*, 7(1), 17–27.

Tonazzini, A., Salerno, E., Mochi, M. & Bedini, L. (2004b). Blind source separation techniques for detecting hidden texts and textures in document images. *International Conference Image Analysis and Recognition*, pp. 241–248.

Tonazzini, A., Salerno, E. & Bedini, L. (2007). Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1), 17–25.

Tonazzini, A., Bianco, G. & Salerno, E. (2009a). Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality. *2009 10th International Conference on Document Analysis and Recognition*, pp. 546–550.

Tonazzini, A., Gerace, I. & Martinelli, F. (2009b). Multichannel blind separation and deconvolution of images for document analysis. *IEEE Transactions on Image Processing*, 19(4), 912–925.

Tonazzini, A., Salerno, E., Abdel-Salam, Z. A., Harith, M. A., Marras, L., Botto, A., Campanella, B., Legnaioli, S., Pagnotta, S., Poggialini, F. et al. (2019). Analytical and mathematical methods for revealing hidden details in ancient manuscripts and paintings: A review. *Journal of advanced research*, 17, 31–42.

Toque, J. A., Sakatoku, Y. & Ide-Ektessabi, A. (2009). Pigment identification by analytical imaging using multispectral images. *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2861–2864.

Townsend, J., Koep, N. & Weichwald, S. (2016). Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *The Journal of Machine Learning Research*, 17(1), 4755–4759.

Van der Meer, F. & De Jong, S. (2000). Improving the results of spectral unmixing of Landsat Thematic Mapper imagery by enhancing the orthogonality of end-members. *International Journal of Remote Sensing*, 21(15), 2781–2797.

Van der Meer, F. D. & Jia, X. (2012). Collinearity and orthogonality of endmembers in linear spectral unmixing. *International Journal of Applied Earth Observation and Geoinformation*, 18, 491–503.

Vavasis, S. A. (2010). On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3), 1364–1377.

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3), 1066–1074.

Wang, Y.-X. & Zhang, Y.-J. (2012). Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6), 1336–1353.

Yang, Z. & Oja, E. (2010). Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(5), 734–749.

Yang, Z., Zhou, G., Xie, S., Ding, S., Yang, J.-M. & Zhang, J. (2010). Blind spectral unmixing based on sparse nonnegative matrix factorization. *IEEE Transactions on Image Processing*, 20(4), 1112–1125.

Yokota, T., Zdunek, R., Cichocki, A. & Yamashita, Y. (2015). Smooth nonnegative matrix and tensor factorizations for robust multi-way data analysis. *Signal Processing*, 113, 234–249.

Yoo, J. & Choi, S. (2008). Orthogonal nonnegative matrix factorization: Multiplicative updates on Stiefel manifolds. *International conference on intelligent data engineering and automated learning*, pp. 140–147.

Yoo, J. & Choi, S. (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information processing & management*, 46(5), 559–570.

Young, D. P. & Ferryman, J. M. (2005). Pets metrics: On-line performance evaluation service. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 317–324.

Zafeiriou, S. & Petrou, M. (2009). Nonlinear non-negative component analysis algorithms. *IEEE Transactions on Image Processing*, 19(4), 1050–1066.

Zhang, D. & Liu, W.-q. (2009). An efficient nonnegative matrix factorization approach in flexible kernel space. *Proceedings of the 21st international joint conference on Artifical intelligence*, pp. 1345–1350.

Zhang, D., Zhou, Z.-H. & Chen, S. (2006). Non-negative matrix factorization on kernels. *Pacific Rim International Conference on Artificial Intelligence*, pp. 404–412.

Zhang, Q., Wang, H., Plemmons, R. J. & Pauca, V. P. (2008). Tensor methods for hyperspectral data analysis: a space object material identification study. *JOSA A*, 25(12), 3001–3012.

Zhu, F. & Honeine, P. (2016). Biobjective nonnegative matrix factorization: Linear versus kernel-based models. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7), 4012–4022.