Décomposition autonome et interprétable d'images multispectrales de documents par apprentissage contraint

par

Kilian DECLERCQ

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE
M. Sc. A.

MONTRÉAL, LE 18 NOVEMBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

Professeur Mohamed Cheriet, directeur de mémoire
Département de génie des systèmes, École de technologie supérieure

Professeur Christian Desrosiers, président du jury
Département de génie logiciel et des TI, École de technologie supérieure

Professeur Abdessamad Ben Hamza, examinateur externe
Département de génie des systèmes d'information, Université Concordia

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 17 NOVEMBRE 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à exprimer toute ma reconnaissance à mon directeur de mémoire, le Professeur Cheriet, pour son soutien précieux et son accompagnement tout au long de ce parcours. Ses conseils éclairés et sa confiance en mon travail m'ont permis de traverser les moments de doute et de difficulté avec sérénité. Grâce à lui, j'ai appris à voir chaque obstacle comme une opportunité d'apprentissage et produire un travail dont je peux être fier.

À toute ma famille, à mes parents et à mon frère, malgré les kilomètres qui nous séparent, je vous remercie du fond du cœur pour votre amour inconditionnel et votre soutien indéfectible. Votre présence, même à distance, a été une source précieuse de motivation. Grâce à vous, j'ai eu la chance de vivre cette expérience unique et de mener à bien ce travail.

À Mae, ma copine, merci d'avoir partagé cette expérience à mes côtés, dans un pays nouveau, avec ses défis et ses découvertes. Ton amour, ton soutien et ta présence au quotidien ont été ma plus grande source de motivation. Ces deux années au Canada n'aurait pas eu la même saveur sans toi à mes côtés. Merci pour chaque moment partagé, chaque rire et chaque encouragement.

Un merci tout particulier à mes amis et à toutes les personnes rencontrées ici à Montréal. Vous avez transformé cette expérience en une belle aventure. Grâce à vous, ces deux ans ont été rythmés de fous rires, de nombreuses découvertes, de soirées mémorables et de discussions enrichissantes. Votre amitié a rendu chaque jour plus agréable, surtout pour survivre à l'hiver.

À mes amis restés en France, même si la distance et le temps ont parfois espacé nos échanges, sachez que je ne vous oublie pas. C'est aussi grâce à vous que je suis devenu la personne et l'ingénieur que je suis aujourd'hui (ou en tout cas très bientôt j'espère).

Enfin, je tiens à remercier l'ÉTS et le NSERC qui ont financé ce projet et permis sa réalisation. Un grand merci à tous les collègues rencontrés au laboratoire Synchromedia, pour leur collaboration, leurs conseils et les moments partagés. Une mention spéciale à Rayane, avec qui j'ai pu vivre les hauts et les bas de cette aventure. Merci pour ces discussions sans fin, ces pauses café salvatrices (sans oublier Mae et Lisa) et cette collaboration qui a rendu chaque défi plus abordable.

Décomposition autonome et interprétable d'images multispectrales de documents par apprentissage contraint

Kilian DECLERCQ

RÉSUMÉ

Les archives numérisées recourent de plus en plus à l'imagerie multispectrale (MS) pour révéler des contenus faibles (encres délavées, palimpsestes, annotations, etc.) et séparer le texte du fond. Or, la décomposition spectrale reste difficile : les approches classiques (*e.g.*, PCA, GMM ou NMF avec rang fixe) exigent des réglages ad hoc, des pré/post-traitements lourds et se généralisent mal à la diversité des supports, des encres et des conditions d'acquisition spectrales.

Pour répondre à ces défis, dans un premier temps, nous introduisons un cadre d'apprentissage bout en bout pour la décomposition multispectrale qui combine un auto-encodeur convolutionnel, couplé à une tête de démélange contrainte (non-négativité, interprétabilité, orthogonalité), enrichie de priors de mise en page (bloc d'attention), afin de préserver la structure des glyphes tout en modélisant le contexte spectro-spatial. Cette approche hybride intègre les principes de la NMF dans une architecture d'auto-encodeur, exploitant ainsi les avantages complémentaires des deux approches. Dans un deuxième temps, face au problème ouvert qu'est le choix manuel du rang, nous proposons un mécanisme pour sa sélection automatique via un élagage (pruning) guidé par longueur de description minimale (MDL), appris conjointement. Les composantes peu informatives sont alors progressivement supprimées pour minimiser simultanément l'erreur de reconstruction et la complexité du modèle. Enfin, dans un troisième temps, nous montrons que ce cadre, nommé **PRISM**, s'applique aux différentes configurations d'images MS, que ce soit pour les cas sur-déterminés (*i.e.*, plus de bandes que de sources) ou sous-déterminés (*i.e.*, moins de bandes, *e.g.* RVB), et se généralise au-delà des documents multispectraux.

Évalué sur MSBin et MStex, deux ensembles de documents variés (*e.g.*, lettres, formulaires, manuscrits) de différentes périodes et états, PRISM améliore de manière constante la séparation encre/fond de +29.5 points F-score contre la binarisation de Howe et dépasse ACE v2 de +1.32 points (état-de-l'art). De plus, pour décomposition d'images MS non-supervisée, PRISM reste jusqu'à 7.4× plus rapide que VBONMF, la meilleure approche NMF concurrente. Des tests sur des scènes hyperspectrales de référence, Jasper Ridge et Urban, ainsi que sur des images RVB, confirment une bonne transférabilité au-delà du domaine documentaire. Des études d'ablation valident l'apport du pruning MDL et des différents priors. Ces résultats indiquent qu'associer contraintes physiques et contexte spatial permet des décompositions interprétables et adaptatives, utiles pour la transcription et la restauration. Le code, les poids et les hyperparamètres de PRISM sont disponibles sur Github et accompagnent le mémoire, dont les contributions ont été intégrées dans une publication acceptée au workshop VisionDocs de la conférence ICCV 2025.

Mots-clés: Factorisation matricielle non-négative, apprentissage automatique interprétable, élagage de réseaux neuronaux, imagerie multispectrale et hyperspectrale, apprentissage non supervisé, documents historiques.

Autonomous and interpretable decomposition of multispectral document images through constrained learning

Kilian DECLERCQ

ABSTRACT

Digitized archives are increasingly using multispectral (MS) imaging to reveal weak content (faded inks, palimpsests, annotations, etc.) and separate text from background. However, spectral decomposition remains difficult : conventional approaches (*e.g.*, fixed-rank NMF, PCA or GMM) require ad hoc settings, cumbersome pre/post-processing and generalize poorly to the diversity of substrates, inks and spectral acquisition conditions.

To address these challenges, we first introduce an end-to-end learning framework for multispectral decomposition that combines a convolutional auto-encoder, coupled with a constrained unmixing head (non-negativity, interpretability, orthogonality), enriched with layout priors (attention block), to preserve glyph structure while modeling the spectro-spatial context. This hybrid approach integrates NMF principles into an auto-encoder architecture, exploiting the complementary advantages of both approaches. Secondly, in response to the open problem of manual rank selection, we propose a mechanism for its automatic selection via pruning guided by minimum description length (MDL), learned jointly. Uninformative components are then progressively removed to simultaneously minimize reconstruction error and model complexity. Finally, in a third step, we show that this framework, named **PRISM**, holds for different MS image configurations, for both overdetermined (*i.e.*, more bands than sources) and underdetermined (*i.e.*, fewer bands, *e.g.* RGB) cases, and generalizes beyond multispectral documents.

Evaluated on MSBin and MStex, two varied document datasets (*e.g.*, letters, forms, manuscripts) from different periods and states, PRISM consistently improves ink/background separation by +29.5 F-score points against Howe's binarization and outperforms ACE v2 by +1.32 points (state-of-the-art). Furthermore, for unsupervised MS image decomposition, PRISM remains up to 7.4× faster than VBONMF, the best competing NMF approach. Tests on reference hyperspectral scenes, Jasper Ridge and Urban, as well as on RGB images, confirm good transferability beyond the documentary domain. Ablation studies validate the contribution of the MDL pruning and the various priors. These results show that combining physical constraints and spatial context enables interpretable and adaptive decompositions, useful for transcription and restoration. PRISM code, weights and hyperparameters are available on Github and accompany this thesis, whose contributions have been integrated into an ICCV 2025 VisionDocs workshop publication.

Keywords: Nonnegative Matrix Factorization, interpretable machine learning, neural network pruning, Multispectral and Hyperspectral imaging, unsupervised representation learning, historical documents.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

CNN	Convolutional Neural Network - Réseau neuronal convolutif
DIBCO	Document Image Binarization Contest
GT	Ground Truth - Vérité terrain
HYDICE	Hyperspectral Digital Imagery Collection Experiment
HS	Hyperspectral
ICA	Independent Component Analysis - Analyse en composantes indépendantes
MSBin	MultiSpectral Binarization Dataset
MDL	Minimum Description Length - Longueur de description minimale (LDM)
MLP	Multilayer Perceptron - Perceptron Multicouche (PMC)
MS	Multispectral
MStex	MultiSpectral Text Extraction contest
NMF	Factorisation Matricielle Non-négative
PCA	Principal Component Analysis - Analyse en composantes principales (ACP)
PRISM	Pruning for Rank-adaptive Interpretable Segmentation Model - Modèle de segmentation interprétable avec adaptation du rang par élagage
RVB	Rouge-Vert-Bleu
SAS	Séparation Aveugle de Source
VCA	Vertex Component Analysis - Analyse des composantes de sommets
ViT	Vision Transformer - Architecture de transformer pour la Vision

LISTE DES SYMBOLES ET UNITÉS DE MESURE

$\lvert \cdot \rvert$	Valeur absolue
$\lVert \cdot \rVert$	Norme de Frobenius
$\odot$	Produit matriciel de Hadamard
$\langle \cdot, \cdot \rangle$	Produit scalaire de Frobenius
ϵ	Constante de stabilité numérique
dB	Décibels
DRD	Distance-Reciprocal Distortion - Distorsion de Distance Réciproque
FM	F-mesure
MSE	Mean Square Error - Erreur quadratique moyenne
nm	Nanomètre
NRM	Negative Rate Metric - Mesure de Rapport Negatif
PSNR	Peak Signal to Noise Ratio - Rapport du Pic du Signal sur Bruit
RMSE	Root Mean Square Error - Racine de l'erreur quadratique moyenne
SAD	Spectral Angle Disance - Distance d'angle spectral

INTRODUCTION

0.1 Contexte et motivation

La lumière pourrait-elle ressusciter des connaissances réduites en cendres ? Des rouleaux carbonisés par l'éruption du Vésuve aux équations d'Archimède effacées sous des textes religieux, la pensée humaine s'est souvent trouvée prisonnière de supports devenus illisibles. Les progrès en imagerie multibandes transforment aujourd'hui ce rapport aux documents historiques, concrétisant cette intuition du grec Anaxagore : « Le visible ouvre nos regards sur l'invisible ». L'analyse des documents historiques joue un rôle crucial dans la préservation et la compréhension de notre héritage culturel. Cependant, sa nature complexe présente des défis uniques pour les techniques traditionnelles d'analyse d'images. Avec le temps, le papier se dégrade, les encres s'estompent et de multiples couches de texte ou d'annotations peuvent se superposer, créant un mélange complexe de matériaux, aussi appelés **sources**. Les méthodes d'imagerie traditionnelles peinent souvent à les démêler, limitant notre capacité à interpréter et à préserver intégralement ces documents. L'imagerie multispectrale (MS) s'est alors imposée comme un outil puissant pour cette tâche, capturant des informations au-delà du spectre électromagnétique visible.

En capturant des informations sur une plus large gamme du spectre électromagnétique, notamment dans l'ultraviolet et l'infrarouge, elle révèle des empreintes spectrales uniques à chaque matériau ; reflétant leur façon spécifique d'absorber, de réfléchir ou d'émettre la lumière à différentes longueurs d'onde, permettant ainsi de révéler du texte caché, de distinguer différentes encres et de fournir des informations sur la composition et l'histoire du document. Un exemple notable de son utilisation et de son efficacité est le Palimpseste illustré sur la Figure 0.1. Une analyse multispectrale de ce livre a permis de révéler sept traités du célèbre mathématicien grec Archimède, dont deux uniques et originaux, qui avaient été grattés et recouverts par des textes religieux au XIIe siècle (Easton, Knox & Christens-Barry, 2003).

Figure 0.1 Application de l'imagerie multispectrale au Palimpseste d'Archimède. Les images en niveaux de gris dévoilent le texte original, tandis que les images Rouge-Vert-Bleu (RVB) montrent le document tel qu'observé à l'œil nu. Images tirées de The Archimedes Palimpsest (Toth, 2004)

L'imagerie MS a aussi plus récemment été utilisée dans le cadre du challenge Vesuvius, une compétition internationale lancée en 2023. Son but étant de déchiffrer les papyrus carbonisés d'Herculanum, une bibliothèque antique ensevelie par l'éruption du Vésuve en 79 apr. J.-C. Ces rouleaux, trop fragiles pour être déroulés physiquement, sont étudiés grâce à des scans 3D et l'intelligence artificielle. L'intérêt majeur réside dans la possibilité unique de révéler des textes antiques perdus depuis près de 2000 ans. Cela pourrait potentiellement enrichir le corpus actuel d'informations textuelles de l'Antiquité européenne dans une proportion significative, estimée à environ 20 %, selon la quantité recouvrable des manuscrits encore présents.

Ainsi, l'approche MS se révèle particulièrement prometteuse pour l'étude et la récupération d'informations textuelles enfouies dans les documents anciens, là où les méthodes classiques basées sur l'imagerie RVB atteignent leurs limites. Cela est bien illustré par la difficulté à distinguer les écritures superposées sur la Figure 0.2, et ce même pour un œil humain averti.

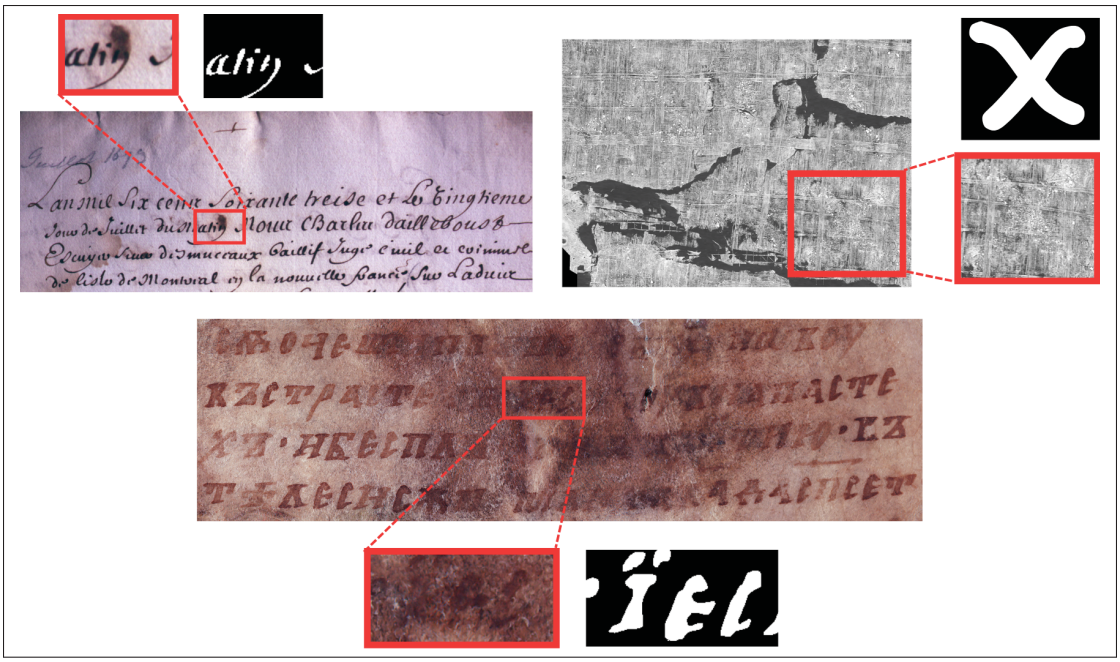

Figure 0.2 Exemple de dégradation du texte sur des documents historiques. En noir et blanc est représenté le texte extrait manuellement

Cependant, bien que capable de capturer une richesse de données spectrales, cette même abondance soulève un défi majeur en termes de traitement et d'interprétation. En effet, ces images **multibandes** de haute dimension, tant en résolution spatiale qu'en nombre de canaux, offrent une mine d'informations dont une grande partie peut s'avérer redondante ou non pertinente pour l'analyse et le déchiffrement. Le défi consiste alors à réduire efficacement cette dimensionnalité tout en préservant un contenu **interprétable**, c.-à-d. compréhensible et pertinent, et ce, parfois sans connaissance préalable de la nature de l'information recherchée. Aussi, contrairement aux caméras RVB omniprésentes, l'acquisition d'images multispectrales demeure une activité très spécialisée, limitant considérablement la disponibilité d'images annotées. Cette rareté de données étiquetées représente un défi majeur pour l'entraînement de modèles d'apprentissage automatique, où la quantité et la qualité des données sont des facteurs déterminants pour la performance, demandant le développement de modèles **non-supervisés**.

0.2 Problématique et questions de recherche

Un cadre particulièrement intéressant pour relever les défis évoqués précédemment est offert par la séparation aveugle des sources (SAS). Le problème de la SAS peut être défini comme le problème d'estimation de sources à partir d'un mélange donné sans connaître ni la fonction de mélange, ni les sources latentes, ni-même leur nombre. Dans le contexte des documents historiques, il est raisonnable d'admettre que l'image finale résulte de l'addition de différentes couches visuelles, chacune venant se superposer à la précédente : l'arrière-plan, le support (*e.g.*, le papier ou le parchemin), les textes originaux, les annotations ajoutées, ainsi que les diverses dégradations liées au vieillissement. Par définition, le cadre de la SAS est non supervisé : il repose uniquement sur les données observées, sans nécessiter de connaissances préalables ou d'annotations. Cette propriété en fait donc une approche adaptée aux documents historiques. Différentes variantes de factorisation matricielle non négative (NMF) ont déjà démontré son efficacité, permettant d'exploiter les propriétés spectrales des images multispectrales de documents (Rahiche, Bakhta & Cheriet, 2019; Rahiche & Cheriet, 2021, 2022).

Cependant, les sources observées interagissent souvent de manière subtile et présentent des signatures spectrales parfois proches, compliquant leur séparation pour les méthodes classique de SAS. Cette difficulté est amplifiée par la grande variabilité des images MS, qui peuvent avoir des caractéristiques variables selon les conditions d'acquisition et les données observées. Cela inclut le nombre de bandes, les longueurs d'ondes utilisées, ou encore l'état de dégradation des images. De plus, ces méthodes traitent souvent chaque pixel indépendamment, ce qui pose problème avec les images de tailles croissantes. L'approche à développer devra donc être robuste en terme de coût de calcul, capable d'intégrer la structure spatiale des documents et de s'adapter à la grande variabilité des configurations rencontrées dans les images multispectrales. Le défi est d'atteindre cette adaptabilité tout en conservant la nature non supervisée de l'approche et en garantissant l'interprétabilité des composantes extraites.

La problématique de recherche devient : **Comment développer un modèle robuste capable de capturer les interactions complexes entre des sources superposées, afin de réaliser une décomposition non supervisée et interprétable des images multibandes de documents ?**

Pour répondre à cette problématique, ce mémoire s'articulera autour de trois questions de recherche qui abordent chacune un défi spécifique de la décomposition non supervisée d'images multispectrales :

QR1. Comment surmonter les limites des approches existantes face aux sources complexes, afin de proposer une décomposition interprétable des images multispectrales de document ?

QR2. Comment éliminer la nécessité de spécifier manuellement le nombre de sources à extraire pour réaliser une décomposition réellement non supervisée ?

QR3. Comment s'assurer de la robustesse du modèle face à la variabilité des configurations d'images multibandes (*e.g.*, nombre de bandes variable, cas sur/sous-déterminés) ?

Ces trois questions ciblent donc les limites méthodologiques de l'état de l'art, où les approches classiques requièrent des réglages ad hoc qui limitent à la fois leur automatisation et leur généralisation face à la diversité des données multispectrales rencontrées dans l'analyse de documents historiques. Pour cadrer cette recherche, le terrain d'étude considéré reste celui des documents anciens, qui cristallise l'ensemble de ces défis. La question de la généralisation sera toutefois examinée empiriquement à travers l'applicabilité du modèle à d'autres types d'images multibandes (*e.g.*, HS ou RVB) de domaine présentant des défis similaires.

0.3 Objectifs et organisation du mémoire

Pour relever ce défi majeur, cette thèse s'articulera autour de ces trois axes de recherche principaux, afin de développer une solution robuste et novatrice :

SO1. **Explorer la synergie entre la Factorisation Matricielle Non-négative (NMF) et les Auto-encodeurs (AE) pour une décomposition améliorée**. Bien que la NMF soit une technique non-supervisée éprouvée pour la séparation aveugle de sources (SAS), elle montre ses limites face à des sources complexes et fortement enchevêtrées. Parallèlement, l'utilisation d'architectures d'auto-encodeurs pour l'interprétabilité des cartes de caractéristiques est un domaine de recherche actif et prometteur. Nous chercherons donc à conceptualiser la NMF traditionnelle au sein d'une architecture d'auto-encodeur, combinant ainsi les forces des deux approches pour obtenir une décomposition plus fine et sémantiquement riche. L'étude portera également sur l'intégration de techniques de régularisation, telles que l'orthogonalité, au sein de ce cadre hybride afin d'améliorer la parcimonie des matrices et la distinction des composantes extraites, éléments cruciaux pour l'analyse de documents dégradés où la variabilité spectrale et le chevauchement des sources sont courants.

SO2. **Développer un mécanismes de sélection automatique du rang** pour une décomposition véritablement non supervisée. Un obstacle significatif des méthodes de décomposition existantes, y compris la NMF et de nombreux modèles d'AE, est la nécessité de spécifier manuellement le paramètre de rang (c.-à-d., le nombre de sources ou de composantes à extraire). Cette contrainte limite fortement l'autonomie et l'applicabilité des modèles, en particulier pour les documents historiques où le nombre exact d'encres, de pigments ou de couches de dégradation est inconnu a priori. Cette partie de la recherche se concentrera sur la conception et l'évaluation d'une stratégie permettant au modèle de déterminer de manière autonome le nombre optimal de sources pertinentes directement à partir des données, en s'inspirant par exemple de principes issus de la théorie de l'information comme la Longueur de Description Minimale (MDL) ou de techniques d'élagage (ou pruning) interprétables.

SO3. **Étendre la généralisation du modèle à différents types d'images multibandes**, qu'elles présentent des caractéristiques spectrales et spatiales distinctes. Cela inclut le cas **sur-déterminé**, où le nombre de bandes excède celui des matériaux à séparer, notamment avec

les images hyperspectrales (HS) (*i.e.*, images MS avec plus de 12 canaux, *voir* Fig. 0.3) et, à terme, le cas **sous-déterminé**, avec les images RVB standard, où potentiellement plus de matériaux doivent être séparés que de bandes disponibles pour les discriminer. La complexité de la composition matérielle de nombreux documents historiques peut entraîner des problèmes de sous-détermination, où le nombre de sources distinctes excède le nombre de bandes spectrales disponibles. Résoudre ce défi est non seulement crucial pour les images MS mais ouvrirait également la voie à l'application du modèle à des images RVB naturelles, voire à des images monocanales. De plus, la robustesse du modèle face à des variations dans les données d'entrée *(e.g.*, nombre de bandes, longueurs d'onde spécifiques, résolution spatiale) est essentielle, étant donné la diversité des capteurs d'imagerie MS et HS. Cette partie visera donc à évaluer et à améliorer la capacité du modèle à s'adapter à ces différents types de données, en testant notamment sa performance sur des jeux de données de télédétection HS et en envisageant des adaptations pour l'analyse d'images naturelles RVB.

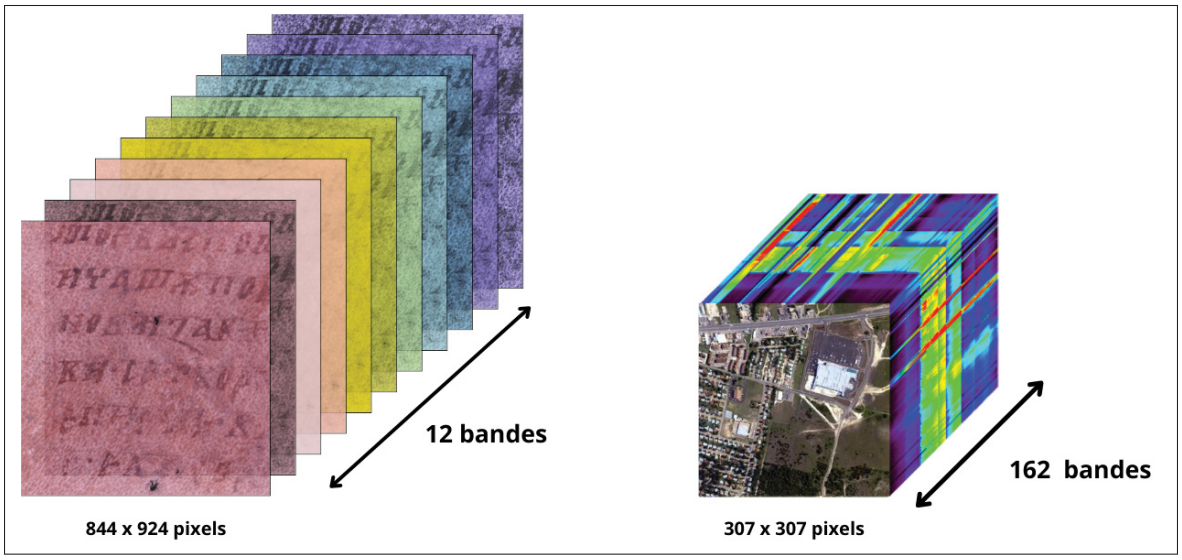

Figure 0.3 Comparaison entre une image multispectrale de document (gauche) et une image satellite hyperspectrale (droite). L'image MS provient de la base de données MSBin, tandis que l'image HS est extraite de la base de données Urban (*voir* section 4.2.1 et 4.3.3). Les couleurs sont utilisées uniquement à des fins de visualisation

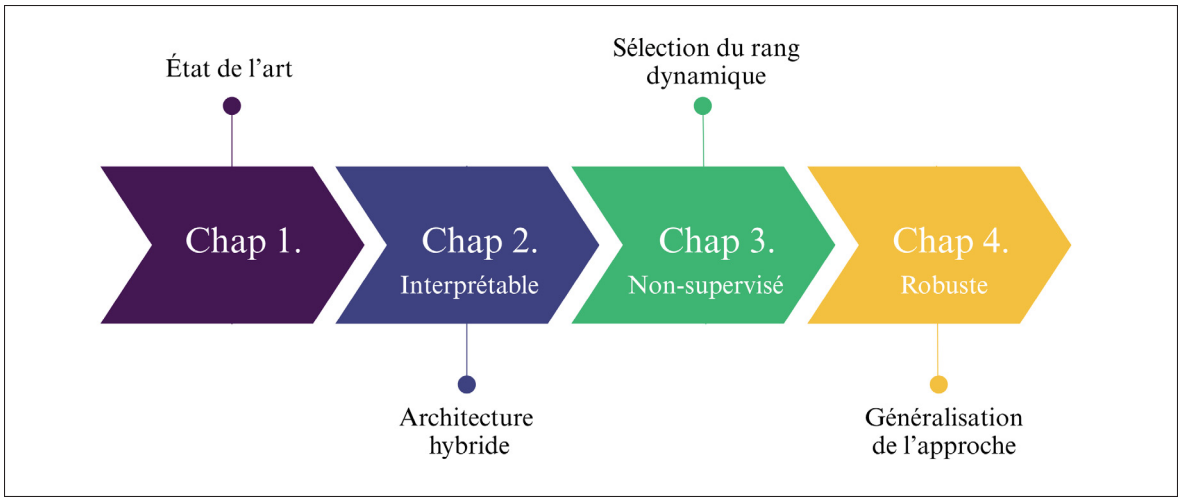

Figure 0.4 Structure du mémoire et axes de recherche

En abordant méthodiquement ces trois sous-objectifs, l'ambition de cette thèse est de proposer un modèle de décomposition d'images multibandes de document qui soit non seulement performant mais aussi complet, non supervisé, et dont les résultats soient directement interprétables, offrant ainsi un outil précieux pour la valorisation du patrimoine documentaire et pour potentiellement d'autres domaines d'application confrontés à des problématiques similaires de mélange de sources. Après un premier chapitre introductif et de revue de littérature globale, ce mémoire explorera les trois sous-problématiques identifiées en trois chapitres distincts correspondants.

CHAPITRE 1

REVUE DE LA LITTERATURE

Ce chapitre établit le **cadre théorique** et les **bases scientifiques** essentielles à la compréhension de la problématique de cette thèse et de l'approche proposée. L'analyse de documents historiques par imagerie multibande vise à résoudre un problème de démélange spectral. L'objectif est de décomposer une image observée, où les signatures spectrales de multiples matériaux comme les encres et le support sont mélangées dans chaque pixel, en ses composantes pures.

Pour un rappel sur les principes physiques de l'acquisition d'images et le concept de signature spectrale, le lecteur est invité à consulter l'Annexe I.

1.1 Analyse des documents historiques par imagerie multibande

L'imagerie multispectrale (MS), développée initialement pour la télédétection, s'est imposée comme une technique essentielle pour l'analyse non-invasive des documents historiques. Un avantage majeur de cette méthode pour le patrimoine culturel est qu'elle ne nécessite pas de prélèvement d'échantillons sur l'objet, préservant ainsi l'intégrité des documents tout en révélant des informations invisibles. Depuis les travaux pionniers du frère Kögel (1920) utilisant la lumière ultraviolette pour la lecture de palimpsestes, l'imagerie MS a permis d'étudier des documents emblématiques tels que les manuscrits de la mer Morte (Shor *et al.*, 2014), le palimpseste d'Archimède (Toth, 2004) ou les lettres de David Livingstone (Knox, Easton Jr, Christens-Barry & Boydston, 2011), permettant de détecter des textes indiscernables à l'œil nu.

1.1.1 Fonctionnement d'une caméra multispectrale

La caméra multispectrale fonctionne selon le même principe fondamental que l'appareil photo conventionnel, avec une différence majeure dans la sélectivité spectrale. Alors que le filtre de Bayer d'un appareil photo RVB utilise trois types de filtres colorés à large bande passante (*voir* Figure I-2) ; la caméra MS emploie entre 4 et 20 filtres passe-bande étroits (10 - 40 nm) qui ne transmettent que des longueurs d'onde spécifiques. Chaque filtre agit comme une

fenêtre spectrale précise, la valeur enregistrée suivant la relation décrite par l'équation A I-1. L'acquisition nécessite des conditions très contrôlées : une chambre noire élimine toute lumière ambiante non désirée, tandis qu'un système d'illumination calibré fournit un éclairage uniforme sur toute la gamme spectrale d'intérêt. L'illumination UV est particulièrement critique, les documents historiques n'émettent pas naturellement ces rayonnements. Une source externe est donc indispensable pour exciter la fluorescence des matériaux et mesurer leur réflectance dans cette région, selon les principes d'interaction lumière-matière (*voir* Figure 1.1 et Équation A I-2). Cette source doit être stable spectralement et spatialement pour garantir des mesures reproductibles et représentatives de la réflectance du document pour chaque longueur d'onde λ.

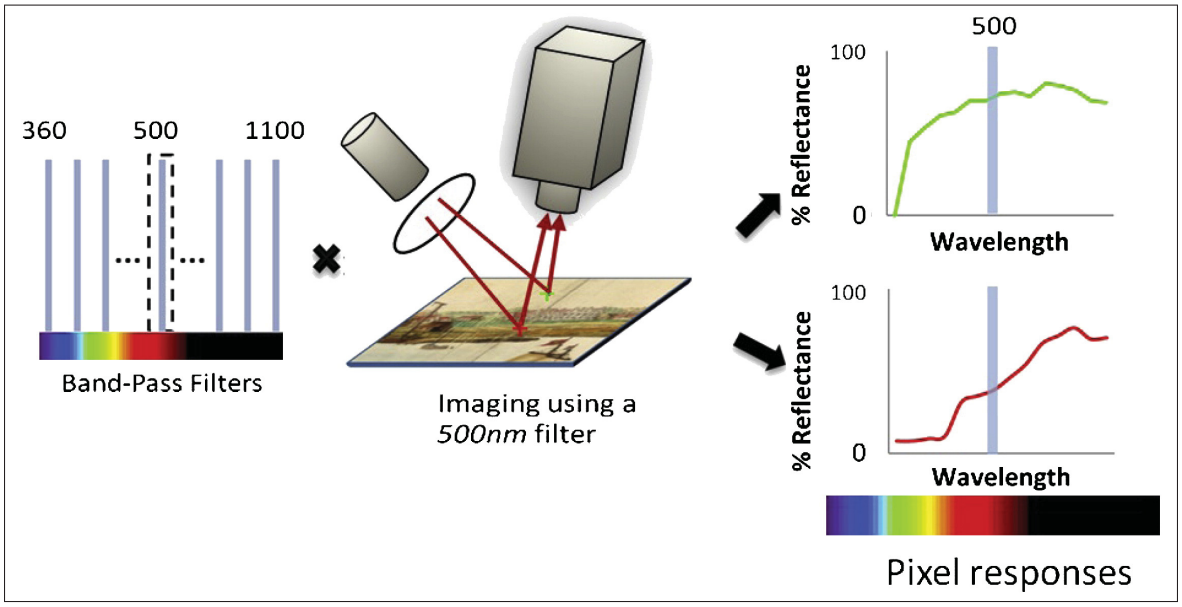

Figure 1.1 Acquisition MS de documents patrimoniaux illustrant la variabilité spectrale des matériaux. Chaque pixel enregistre une signature de réflectance distincte en fonction de la longueur d'onde. Chaque filtre passe-bande étroit permet d'isoler une bande spectrale spécifique. Visualisation tirée de Joo Kim *et al.* (2011)

1.1.2 Mélange de sources : encres, supports, dégradations et annotations

Les documents historiques présentent une complexité visuelle élevée liée à la diversité des matériaux et aux altérations accumulées. Les supports (papier, parchemin, papyrus), issus de procédés artisanaux, sont hétérogènes : la structure des fibres végétales ou animales varie

localement, influençant densité et couleur. Les encres, principalement ferro-galliques, se déclinent en plus de 250 recettes recensées en Europe (Duh, Krstić, Desnica & Fazinić, 2018), modifiées selon les régions et époques par l'ajout d'additifs variés. Cette diversité chimique produit une large palette spectrale, parfois au sein d'un même manuscrit. La dégradation des encres et supports accentue cette complexité : écaillage, perte de couches, corrosion avec halos bruns, infiltrations dans le support, fragilisation et effritement du papier (Melo *et al.*, 2022). L'humidité favorise les moisissures et l'effacement partiel, l'infestation d'insectes crée des perforations, et les dépôts de poussière ou taches de cire masquent le texte. Enfin, les corrections, annotations marginales et ajouts ultérieurs sont fréquemment observées pour les manuscrits les plus âgés. Les parchemins anciens, coûteux à produire en leur temps, étaient souvent effacés et réparés afin de réutiliser le support, créant ce que l'on appelle des palimpsestes. Il est courant que ce genre de document puisse présenter jusqu'à trois couches de texte superposées, chacune écrite avec une encre différente à des époques distinctes, ajoutant leurs signatures optiques au texte original et aux dégradations. Des exemples notables comme le palimpseste syriaque de Galien (Easton, Knox, Christens-Barry & Boydston, 2018) illustrent ces défis, tandis que certaines situations extrêmes, comme les manuscrits carbonisés d'Herculanum, mentionnés dans l'introduction, sont complètement inaccessibles à l'imagerie RVB (Parker *et al.*, 2019).

Tous ces éléments se combinent pour former un «mélange» où chaque pixel de l'image observée contient l'information spectrale de multiples sources superposées. Cette superposition rend souvent le texte totalement illisible à l'œil nu ou par imagerie RVB conventionnelle. Les recherches récentes sur les palimpsestes démontrent que l'imagerie spectrale appliquée à ces manuscrits pose de nombreux défis mais promet également de produire beaucoup de résultats (Rossi, Zoleo, Bertoncello, Meneghetti & Deiana, 2021; Mazzocato, Cimino & Daffara, 2024). L'imagerie MS s'avère alors particulièrement efficace pour restaurer l'information perdue (Leão *et al.*, 2024). Elle permet d'examiner chaque page des documents comme une superposition de couches qui donnent des réponses différentes en fonction des bandes, révélant ainsi leur nature complexe matérielle et historique.

1.1.3 Variabilité spectrale et similarité entre composants

L'imagerie MS des manuscrits souffre cependant d'une variabilité spectrale importante, lié à la fois aux documents observés mais aussi à l'absence de standardisation des caméras utilisées (Jones, Duffy, Gibson & Terras, 2020). D'une part, les signatures d'un même matériau peuvent varier selon les conditions d'acquisition (illumination, sensibilité des capteurs, température) (Hedjam & Cheriet, 2013; Hollaus, Diem, Fiel, Kleber & Sablatnig, 2015a) et selon les systèmes (caméras monochromes, LEDs, logiciels), qui possèdent chacun leur profil spectral propre. D'autre part, les objets similaires en couleur peuvent avoir des réflectances spectrales différentes, rendant complexe l'interprétation automatique des données spectrales. Cette variabilité dépend des objets en eux-mêmes ainsi que des processus de dégradation qui modifient leurs propriétés optiques de manière non uniforme. Deux encres ferro-galliques, bien que chimiquement similaires, peuvent présenter des signatures spectrales significativement différentes en fonction de nombreux paramètres. La concentration en fer, le type de tanins utilisés (*p. ex.*, noix de galle, écorce de chêne, sumac, etc.), les additifs (*p. ex.*, gomme arabique, sulfate de cuivre, etc.) et même les conditions de préparation peuvent influencer la réponse spectrale de l'encre (Teixeira, Nabais, de Freitas, Lopes & Melo, 2021). Aussi, une même encre sur un même matériel peut présenter une signature différente selon son état de dégradation aux différents endroits de l'image. Elle peut présenter des teintes visible allant du noir au brun pâle, ou même des teintes verdâtres dues à la corrosion. Ce vieillissement introduit une dimension temporelle dans la variabilité spectrale, pouvant être utile lorsque exploitée par certaines méthodes de datation (Rahiche, Hedjam, Al-maadeed & Cheriet, 2020; Ursescu, Malutan & Ciovica, 2009).

En revanche, la conséquence de cette haute variabilité est qu'il n'existe pas de méthode universelle pour l'analyse des images MS de documents historiques. La plupart des systèmes de compréhension des documents sont basés sur l'application de techniques de reconnaissance de formes pour les images RVB conventionnelles, qui peuvent avoir du mal face à la complexité des données MS. Les approches doivent être adaptées à chaque manuscrit selon son support, ses encres, ses dégradations et son histoire de conservation (Tonazzini *et al.*, 2019), les algorithmes devant tenir compte des propriétés physiques spécifiques des matériaux.

1.2 Séparation Aveugle des Sources pour la décomposition d'images multibandes

Face à cette complexité, les méthodes traditionnelles de segmentation et de classification atteignent leurs limites. Lorsque peu de bandes spectrales sont disponibles, il est difficile de trouver une stratégie de segmentation efficace pour estimer correctement les différentes classes ou objets des documents historiques. La nature même du problème ; séparer des signaux mélangés sans connaissance préalable ni des sources originales ni du processus de mélange, oriente naturellement vers les techniques de Séparation Aveugle des Sources (SAS). Le cadre théorique offert par la SAS, cherche à résoudre un problème de séparation de signaux où les sources originales ainsi que leur méthode de combinaison sont inconnues, en se basant uniquement sur les signaux mélangés accessibles (Cardoso, 1998). Dans le contexte des documents historiques, chaque pixel de l'image MS peut être considéré comme un mélange de plusieurs sources : premier plan (texte, encre), arrière-plan (parchemin, support), informations de dégradation et potentielles couches d'écriture superposées. Les recherches récentes démontrent l'efficacité de cette approche pour traiter les documents multibandes (Rahiche *et al.*, 2019).

Le caractère «aveugle» des techniques de SAS est un atout majeur lorsqu'il s'agit d'analyser des documents historiques. En effet, l'information a priori disponible sur la composition exacte des matériaux utilisés ou sur l'étendue et la nature des dégradations est souvent partielle, voire totalement absente avant une analyse approfondie. Les méthodes qui reposeraient sur des bibliothèques spectrales de référence pour chaque matériau pourraient se heurter aux difficultés liées à la variabilité discutées plus tôt. La SAS, en revanche, exploite principalement des hypothèses statistiques sur les propriétés des sources (*p. ex.*, leur indépendance statistique ou leur non-négativité) pour les estimer directement à partir des données observées (Naik, Wang *et al.*, 2014). Cette capacité à laisser parler les données rend la SAS particulièrement adaptée à l'exploration et à la découverte de la composition matérielle des documents anciens, où la part d'inconnu est souvent importante (Giacometti *et al.*, 2017). Par conséquent, la SAS ouvre la voie à une caractérisation matérielle non invasive plus objective et détaillée, ce qui est essentiel non seulement pour la lecture et la compréhension des textes, mais aussi pour les problématiques de conservation, d'authentification et de datation des documents.

1.2.1 Principes fondamentaux et modèle mathématique

Au cœur de la Séparation Aveugle des Sources se trouve le problème de la récupération d'un ensemble de signaux sources originaux, notés $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_r(t)]^T$, à partir d'un ensemble d'observations mélangées, notées $\mathbf{y}(t) = [y_1(t), y_2(t), \ldots, y_m(t)]^T$ (Cardoso, 1998). Dans le contexte spécifique des images multibandes de documents historiques, l'indice t peut être assimilé à un index représentant la position spatiale d'un pixel, **r** est le nombre de sources que l'on cherche à identifier, et **m** est le nombre de bandes spectrales acquises. Ainsi, chaque composante $y_i(t)$ du vecteur d'observation $\mathbf{y}(t)$ représente l'intensité (ou la réflectance) du pixel t dans la i-ème bande spectrale.

Le modèle le plus couramment utilisé pour décrire la relation entre les sources et les observations en SAS est le modèle de mélange linéaire (Linear Mixture Model, LMM). Ce modèle postule que chaque signal observé $y_i(t)$ est une combinaison linéaire des signaux sources, affectée par un bruit additif. Mathématiquement, cela s'exprime sous forme vectorielle comme :

$$\mathbf{y}(t) = \mathbf{E}\mathbf{a}(t) + \mathbf{n}(t) \tag{1.1}$$

où :

- $\mathbf{y}(t) \in \mathbb{R}^m$ est le vecteur des **m** observations au pixel t (spectre du pixel t) ;
- $\mathbf{a}(t) \in \mathbb{R}^r$ est le vecteur des **r** abondances au pixel t. Dans le contexte de l'imagerie multispectrale, ces abondances représentent les proportions de chaque matériau pur présent au pixel t ;
- $\mathbf{E} \in \mathbb{R}^{m \times r}$ est la matrice de mélange, supposée inconnue en SAS. Ses colonnes contiennent les signatures spectrales des r sources pures, et ses coefficients e_{ij} représentent la contribution de la j-ème source dans la i-ème bande spectrale ;
- $\mathbf{n}(t) \in \mathbb{R}^m$ est un vecteur de bruit additif, généralement supposé de moyenne nulle et statistiquement indépendant des sources.

Dans la pratique de l'imagerie multispectrale, on travaille souvent avec la factorisation matricielle globale :

$$\mathbf{Y} \approx \mathbf{EA} \tag{1.2}$$

où $\mathbf{Y} \in \mathbb{R}^{m \times n}$ est la matrice des données observées (chaque colonne étant le spectre d'un pixel), $\mathbf{E} \in \mathbb{R}^{m \times r}$ contient les signatures spectrales des sources pures, et $\mathbf{A} \in \mathbb{R}^{r \times n}$ contient les abondances correspondantes pour les n pixels de l'image.

Pour que ce problème ; qui sera admis comme sur-déterminé dans le cadre des images MS (*i.e.*, plus de bandes disponibles que de matériaux observés), admette une solution unique, certaines hypothèses fondamentales peuvent être formulées concernant les signaux sources :

1. **Indépendance statistique mutuelle des signaux sources** : Les composantes $s_j(t)$ du vecteur source $\mathbf{s}(t)$ sont supposées être statistiquement indépendantes les unes des autres. C'est l'hypothèse la plus cruciale et distinctive de nombreuses méthodes de SAS, notamment l'Analyse en Composantes Indépendantes (ACI). Elle signifie que la connaissance de la valeur d'une source ne fournit aucune information sur la valeur des autres sources.
2. **Non-gaussianité des sources (pour l'ACI)** : Au plus une des sources $s_j(t)$ peut avoir une distribution de probabilité gaussienne. En effet, si plusieurs sources sont gaussiennes et indépendantes, le théorème central limite implique que leur mélange linéaire tend également vers une distribution gaussienne, et la matrice de mélange ne peut alors être identifiée de manière unique au-delà d'une simple décorrélation.
3. **Nombre de sources** : Le nombre de sources r est généralement supposé inférieur ou égal au nombre d'observations m ($r \leq m$) pour le modèle de base.

Il est important de noter que la SAS souffre de certaines **indéterminations fondamentales** :

- **Amplitude des sources** : Il est impossible de déterminer de manière unique l'amplitude des sources estimées. Si $\mathbf{a}(t)$ est une solution, alors $\mathbf{Da}(t)$ est aussi une solution, où $\mathbf{D}$ est une matrice diagonale inversible, et la matrice de mélange devient $\mathbf{ED}^{-1}$.

- **Ordre des sources** : L'ordre dans lequel les sources sont estimées est arbitraire. Si $\mathbf{P}$ est une matrice de permutation, alors $\mathbf{Pa}(t)$ est aussi une solution valide, avec une matrice de mélange $\mathbf{EP}^{-1}$.
- **Polarité des sources** : Le signe des sources peut être inversé, absorbé par la matrice de mélange.

Dans le contexte de l'analyse d'images multibandes de documents, ces indéterminations sont souvent acceptables. L'échelle absolue des signatures spectrales est moins importante que leur forme relative, et l'ordre des matériaux identifiés n'affecte pas leur caractérisation. La forme des spectres et leur distribution spatiale relative (cartes d'abondance) sont les informations primordiales recherchées.

1.2.2 Techniques algébriques traditionnelles associées à la SAS et au regroupement

Le modèle de mélange linéaire $\mathbf{y}(t) = \mathbf{Ea}(t) + \mathbf{n}(t)$ est une simplification de la réalité physique complexe des interactions lumière-matière au sein d'un document historique. Des phénomènes tels que la diffusion multiple de la lumière dans les couches d'encre ou de pigment, les interactions chimiques entre l'encre et le support, ou la pénétration de l'encre dans les fibres du papier peuvent introduire des non-linéarités. Néanmoins, pour de nombreuses applications en imagerie multispectrale de documents, où l'objectif principal est de discriminer différents matériaux, d'améliorer la lisibilité d'un texte effacé ou de cartographier des dégradations, le modèle linéaire fournit une approximation souvent suffisante et mathématiquement traitable. Avant d'aborder la Factorisation Matricielle Non-négative (NMF), il est utile d'évoquer plusieurs méthodes classiques souvent employées pour le traitement des images MS de documents. Certaines, comme l'Analyse en Composantes Principales (ACP) et l'Analyse en Composantes Indépendantes (ACI), sont directement associées à la SAS (Yu, Hu & Xu, 2013). D'autres, comme k-moyennes et GMM, relèvent du regroupement non supervisé et sont plus souvent utilisées pour la segmentation.

- **L'Analyse en Composantes Principales (ACP)** est souvent employée en imagerie MS de documents pour la réduction de dimensionnalité, le débruitage et l'amélioration du contraste (Rodarmel & Shan, 2002; Kaarna, Zemcik, Kalviainen & Parkkinen, 2002). Des études de cas spécifiques ont démontré son utilité pour révéler des détails cachés dans des œuvres d'art et manuscrits anciens (Knox *et al.*, 2011; Leão *et al.*, 2024; López-Baldomero, Buzzelli, Moronta-Montero, Martínez-Domingo & Valero, 2025). Cependant, l'ACP n'est pas une véritable technique de séparation de sources : elle se limite à une décorrélation statistique et ses composantes correspondent rarement aux spectres physiques des matériaux (Tonazzini *et al.*, 2019; Jones *et al.*, 2020).

- **L'Analyse en Composantes Indépendantes (ACI)** représente une évolution conceptuelle en cherchant l'indépendance statistique des composantes (Choi, Cichocki, Park & Lee, 2005; Hyvarinen, 1999). Elle a montré son potentiel pour la séparation des couches dans les palimpsestes et l'amélioration de la lisibilité des textes effacés (Davies & Zawacki, 2019; Salerno, Tonazzini & Bedini, 2007; Tonazzini, Bedini & Salerno, 2004). Néanmoins, l'hypothèse d'indépendance est souvent compromise, avec par exemple des mécanismes de dégradation qui affectent simultanément plusieurs composants, ou la dépendance statistique inhérentes à la décomposition des données MS en cartes d'abondances [1]. De plus, les composantes indépendantes estimées peuvent contenir des valeurs négatives, ce qui est physiquement incohérent pour des grandeurs comme la réflectance spectrale ou les concentrations de matériaux. Cette limitation fondamentale compromet l'interprétation directe des résultats comme des spectres de matériaux purs ou des cartes d'abondance.

- **Les K-moyennes**, méthodes de regroupement, permettent une segmentation rapide des régions spectralement distinctes (MacQueen, 1967). Des variantes adaptées aux documents ont étés développées comme les K-moyennes sphériques utilisant la dissimilarité cosinus (Hornik, Feinerer, Kober & Buchta, 2012) et SKKHM intégrant l'information spatiale (Li, Mitianoudis & Stathaki, 2007). Cependant, ces variantes conservent le même inconvénient

[1] La somme des fractions d'abondance étant constante, celles-ci présentent forcément une dépendance statistique (Nascimento & Dias, 2005).

que les K-moyennes standard : les centroïdes résultants sont des moyennes mathématiques et ne garantissent pas la correspondance avec des signatures spectrales pures.

- **Le Modèle de Mélange Gaussien (GMM)** offre une segmentation probabiliste plus flexible pour le regroupement non supervisé (Reynolds *et al.*, 2009). L'idée fondamentale est de supposer que les données observées proviennent d'un mélange de K distributions de probabilité Gaussiennes multidimensionnelles. Le GMM a été appliqué aux images multispectrales de documents, comme le montrent les travaux de Hollaus, Diem & Sablatnig (2018). Bien que son application soit prometteuse, ces travaux soulignent une limitation importante : la sensibilité des GMM aux variations des données brutes. La méthode nécessite un pré-traitement crucial et met en évidence le manque d'informations spatiales nécessaire pour obtenir une segmentation plus robuste.

En résumé, ces méthodes ont permis des avancées notables en imagerie MS de documents, mais elles partagent un défaut majeur : l'absence de contrainte de **non-négativité**, qui limite leur interprétation physique. Cette faiblesse motive l'adoption d'approches alternatives comme la **Factorisation Matricielle Non-négative**, qui intègre naturellement cette contrainte et fournit une interprétation directe en termes de spectres de matériaux et de cartes d'abondance.

1.2.3 La Factorisation Matricielle Non-négative (NMF) pour la séparation aveugle de sources

La **Factorisation Matricielle Non-négative (NMF)** a connu un essor considérable au sein de diverses communautés scientifiques, de l'analyse informatique au décodage génétique (Gillis, 2020). Initialement développée par Paatero & Tapper (1994), elle fut popularisée par Lee et Seung à travers une série d'articles à la fin des années 90 (Lee & Seung, 1999, 2000). La NMF repose sur une observation fondamentale : dans la nature, la plupart des structures complexes résultent de l'addition de sous-structures ou de composants plus simples. Que ce soit en architecture (bâtiments comme une somme de matériaux), en chimie (molécules comme une combinaison d'atomes) ou en imagerie (pixels comme une somme d'ondes électromagnétiques), ce principe

additif prévaut. La NMF postule qu'en décomposant une structure complexe en une somme de «briques» constitutives non-négatives, on peut obtenir une représentation plus interprétable. En effet, notamment dans le contexte des images de documents, des composantes négatives manquent de sens physique, que ce soit pour l'intensité de réflectance ou indiquer la présence d'objets. Cette présence est signifiée par un nombre positif, l'absence par zéro (Ngoc-Diep, 2008). L'adéquation particulière de cette méthode pour l'étude des documents historiques multibandes justifie un examen détaillé de ses principes, de ses avantages, de ses applications spécifiques.

1.2.3.1 Formulation Mathématique

La NMF vise à factoriser une matrice de données $\mathbf{Y}$ en deux matrices de rang inférieur, $\mathbf{U}$ et $\mathbf{V}$, dont tous les éléments sont non-négatifs [2] :

$$\mathbf{Y} \approx \mathbf{UV} \tag{1.3}$$

Dans le contexte de l'imagerie multibande de documents historiques, $\mathbf{Y}$ est une matrice $\mathbb{R}_+^{b\times n}$, où b est le nombre de bandes spectrales et n le nombre de pixels (après réorganisation de l'image 2D en une structure 1D de pixels). La NMF cherche à estimer $\mathbf{U}$ de dimension $\mathbb{R}_+^{b\times r}$ et $\mathbf{V}$ de dimension $\mathbb{R}_+^{r\times n}$, avec r (le rang) représentant le nombre de sources recherchées. Typiquement, r est choisi de manière à ce que $r \ll \min(m, n)$, ce qui implique que la NMF réalise également une réduction de dimensionnalité. Cette factorisation obtenue constitue une approximation, la résolution exacte du problème ayant été prouvée NP-difficile par Vavasis (2010), c'est-à-dire qu'il n'existe aucun d'algorithme déterministe capable de garantir une solution optimale en temps polynomial par rapport à la taille des données, d'où la nécessité de relaxer le problème. Un développement mathématique de la NMF, avec son algorithme, les fonctions de coût utilisées ainsi que les méthodes d'optimisation associées, est proposé en Annexe II.

[2] Les notations peuvent varier selon les publications mais restent équivalentes : $\mathbf{Y} \approx \mathbf{UV}^T$, $Y \approx \mathbf{MA}$, $Y \approx \mathbf{AX}$, $\mathbf{V} \approx \mathbf{WH}$, $\mathbf{A} \approx \mathbf{XY}^T$.

1.2.3.2 Interprétation des facteurs dans l'analyse d'images MS de documents

L'un des principaux avantages de la NMF pour l'analyse des documents historiques multibandes réside dans l'interprétabilité physique des matrices de factorisation $\mathbf{U}$ et $\mathbf{V}$:

- **Matrice des composantes** (Endmembers en anglais) : Les r colonnes de la matrice $\mathbf{U} \in \mathbb{R}^{b \times r}$ peuvent être interprétées comme les signatures spectrales des r composantes pures présentes dans l'image. Chaque colonne $\mathbf{u}_j$ constitue un vecteur de b valeurs non-négatives représentant la réflectance du j-ème matériau pur (*p. ex.*, encre spécifique, support, ou dégradation etc.).
- **Matrice d'abondance** : Les n colonnes de la matrice $\mathbf{V} \in \mathbb{R}^{r \times n}$ contiennent les coefficients d'abondance correspondants à chaque source pour chacun des n pixels. La j-ème ligne de $\mathbf{V}$, notée $\mathbf{v}_j^T$, peut être réorganisée spatialement pour former une carte d'abondance illustrant la distribution spatiale du j-ème matériau à travers l'image. Chaque coefficient $\mathbf{v}_{ji}$ indique la proportion (contribution) de la j-ème signature spectrale au spectre observé du i-ème pixel.

Ces éléments sont illustrés sur la Figure 1.2, qui schématise la décomposition d'une image MS synthétique en ses signatures spectrales pures et ses cartes d'abondance correspondantes. Bien que la NMF soit fondamentalement une méthode de séparation de sources non-supervisée, son application aux documents historiques nécessite une approche semi-supervisée, avec une interprétation guidée par l'expertise. L'identification des colonnes de $\mathbf{U}$ comme matériaux spécifiques requiert généralement une comparaison avec des spectres de référence (si disponibles), une analyse contextuelle des cartes d'abondance, ou une validation par des techniques analytiques complémentaires. Des variantes de la NMF permettent alors d'intégrer explicitement des connaissances a priori dans le processus de factorisation, grâce à l'implémentation de différentes contraintes, améliorant la décomposition en fonction des besoins.

1.2.3.3 Contraintes et variantes de la NMF

Dans le contexte du démélange spectral, deux contraintes fondamentales sont généralement appliquées aux cartes d'abondance : la contrainte de non-négativité (Abundance Non-negativity

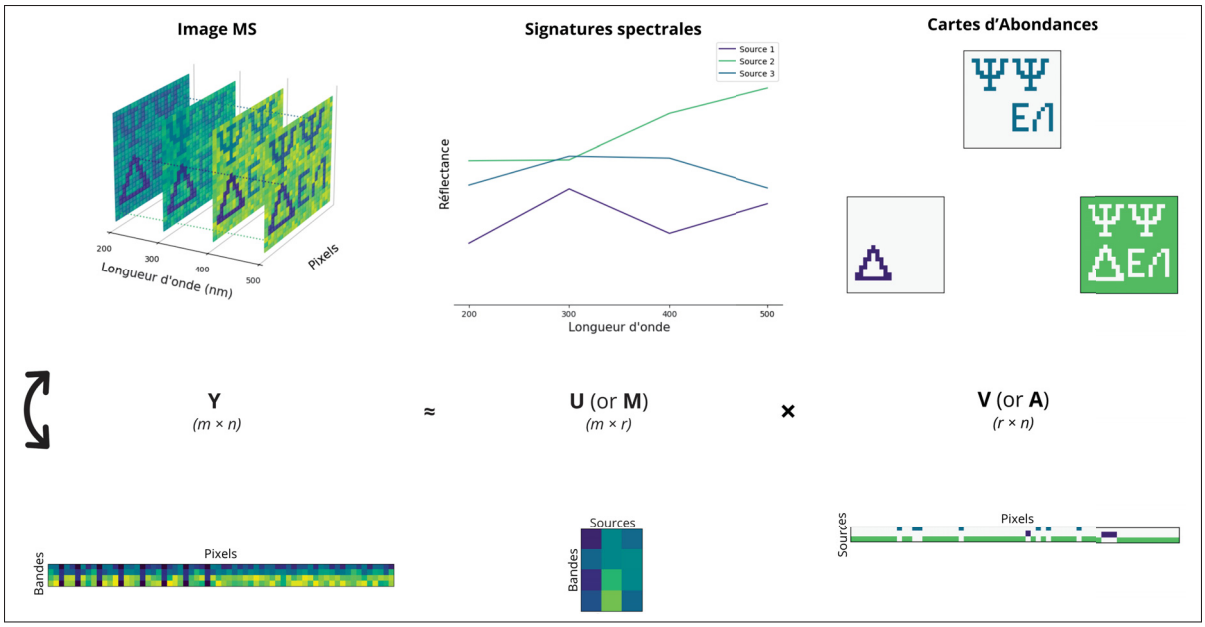

Figure 1.2 Illustration schématique de la décomposition par NMF d'une image multispectrale synthétique. L'image originale est constituée de trois matériaux (encre violette, encre bleue et support) réagissant de manière unique aux différentes longueurs d'onde. La factorisation permet d'extraire les signatures spectrales pures **U** et les cartes d'abondance spatiales correspondantes **V**, représentés sous leur forme matricielle. Les données présentées sont entièrement simulées à des fins de visualisation, les couleurs étant utilisées uniquement pour faciliter l'interprétation visuelle

Constraint - ANC), ainsi que la contrainte de somme à l'unité (Abundance Sum-to-one Constraint - ASC), imposant que $\sum_i V_{ij} = 1$ pour chaque pixel j (*i.e.*, somme à 100%). Ce cadre général de NMF contrainte permet alors d'intégrer des connaissances a priori (*p. ex.*, colonnes de **U** partiellement fixées lorsque des signatures spectrales sont connues), en gardant une interprétation cohérente des abondances comme contributions relatives des matériaux. Au-delà de ces contraintes de base, diverses variantes de la NMF ont été développées pour répondre à différents besoins spécifiques (*voir* Cichocki, Zdunek, Phan & Amari (2009) ou Wang & Zhang (2012) pour une revue développée). Parmi les principales variantes développées, on peut citer :

- **NMF Parcimonieuse (Sparse NMF) :** Impose des contraintes de parcimonie (*p. ex.*, régularisation L_0, L_1, $L_{1/2}$ ou log-norme) sur les matrices de facteurs **U** et/ou **V** favorisant des

représentations localisées et basées sur les parties, améliorant potentiellement l'interprétabilité et la robustesse au bruit (Peharz & Pernkopf, 2012; Le Roux, Weninger & Hershey, 2015).

- **NMF Orthogonale (ONMF) :** Impose des contraintes d'orthogonalité sur les matrices de facteurs, telles que $\mathbf{U^T U} = \mathbf{I}$ (colonnes de $\mathbf{U}$ orthogonales) et/ou $\mathbf{V V^T} = \mathbf{I}$ (lignes de $\mathbf{V}$ orthogonales). Cette orthogonalité garantit que les composantes extraites sont maximalement indépendantes et non-redondantes, ce qui améliore l'identification de matériaux distincts. Dans le contexte des documents historiques, l'ONMF est prouvée efficace pour séparer des pigments aux signatures spectrales proches mais chimiquement différents (Choi, 2008; Yoo & Choi, 2010a; Rahiche & Cheriet, 2020).
- **NMF Lisse (Smooth NMF) :** Impose des contraintes de régularité spatiale sur les cartes d'abondance $\mathbf{V}$, particulièrement utile pour les matériaux variant de manière graduelle. Cette approche améliore la cohérence spatiale et réduit l'impact du bruit (Salehani & Gazor, 2017).
- **NMF Régularisée par Graphe (GNMF) :** Incorpore la structure géométrique locale des données via un terme de régularisation basé sur un graphe de similarité, préservant les relations de voisinage dans l'espace de faible dimension défini par U ou $\mathbf{V}$ (Yi *et al.*, 2019; Rahiche & Cheriet, 2020; Rhodes, Jiang & Jiang, 2025).
- **Tri-NMF** : Décompose la matrice de données $\mathbf{Y}$ en trois facteurs, souvent non négatifs, par exemple $\mathbf{Y} \approx \mathbf{U S V^T}$. Dans ce cadre, $\mathbf{S}$ peut être une matrice centrale de taille variable, et des contraintes supplémentaires (comme l'orthogonalité sur $\mathbf{U}$ et $\mathbf{V}$, ou la parcimonie) peuvent être imposées. Cette formulation peut offrir plus de flexibilité et est utilisée pour relaxer problème d'optimisation (Ding, Li, Peng & Park, 2006; Ding, Li & Jordan, 2008).

1.2.3.4 Applications de la NMF pour l'analyse des documents

La NMF est une technique non supervisé employée pour l'analyse d'images de documents, notamment avec des données MS ou HS (Pauca, Piper & Plemmons, 2006; Soukup & Bajla, 2008; Rahiche *et al.*, 2019). L'utilisation de contraintes assure que les caractéristiques extraites sont additives et physiquement interprétables, ce qui est fondamental pour les données d'imagerie

dont les valeurs sont intrinsèquement non négatives (Magkanas, Bagán, Sistach & García, 2021). Les applications principales de la NMF dans l'analyse de documents historiques incluent :

- **Analyse et discrimination d'encres et pigments :** La NMF permet d'identifier et de discriminer les signatures spectrales de divers matériaux d'écriture, tels que les encres et les pigments, même lorsque ceux-ci sont visuellement indiscernables (López-Baldomero *et al.*, 2023; Lyu *et al.*, 2020). Cette capacité est cruciale pour l'authentification de documents, l'étude des techniques artistiques et potentiellement la datation (Magkanas *et al.*, 2021). Pour les données MS/HS, des variantes orthogonales (ONMF) sont souvent employées pour améliorer la distinction des composantes spectrales extraites (Rahiche *et al.*, 2019, 2020).
- **Détection de falsifications :** La technique est appliquée à la détection de falsifications, comme l'identification d'ajouts d'encre. La NMF peut révéler des composantes distinctes présentant des incohérences spectrales indiquant ainsi des altérations. Des modèles spécifiques, tels que la NMF orthogonale régularisée par graphe, ont été proposés pour la détection de discordance d'encres dans les images MS de documents (Rahiche & Cheriet, 2020).
- **Restauration et analyse de documents :** Dans le domaine de la restauration et de l'analyse, la NMF contribue à la séparation virtuelle des couches d'une image permettant la suppression de l'effet de transparence parasite du texte verso (Merrikh-Bayat, Babaie-Zadeh & Jutten, 2010) ou la récupération de textes effacés ou superposés (Wang & Zhang, 2011; Phon-Amnuaisuk, 2013). Elle est également utilisée pour l'identification de pigments dans les manuscrits enluminés et les œuvres d'art, aidant à caractériser les matériaux et les techniques de création (Magkanas *et al.*, 2021; Lyu *et al.*, 2020).
- **Extraction de contenu et modélisation thématique :** En plus de son application dans l'analyse d'images, la NMF est avant tout une technique reconnue pour la modélisation thématique et l'analyse sémantique latente de contenu textuel (Kulkarni, Madurwar, Narlawar, Pandya & Gawande, 2023). Dans ces applications, la matrice $\mathbf{U}$ peut représenter des thèmes (ensembles de mots) et $\mathbf{V}$ leur distribution ou prévalence au sein d'un corpus de documents.
- **Segmentation et binarisation :** La NMF est employée pour la segmentation d'images de documents, permettant de séparer le texte du fond, les illustrations ou différentes régions matérielles (Mazack, 2009). Des cadres spécifiques, comme MSdB-NMF (MultiSpectral

Document image Binarization via NMF) (Salehani, Arabnejad, Rahiche, Bakhta & Cheriet, 2020) ou des approches basées sur l'ONMF, ont été développés pour la binarisation et la décomposition de documents multispectraux (Rahiche *et al.*, 2019).

Cependant, l'efficacité de la NMF est conditionnée par la résolution de plusieurs défis méthodologiques. Le problème d'optimisation en NMF, non convexe, peut converger vers des minima locaux, et la solution obtenue est sensible aux valeurs initiales des matrices **U** et **V** (Magkanas *et al.*, 2021). De plus, la NMF standard repose sur l'hypothèse d'un mélange linéaire additif des composantes spectrales. Cependant, les interactions lumière-matière dans les documents, telles que celles impliquant des couches d'encre épaisses ou des phénomènes de diffusion multiple, peuvent introduire des non-linéarités significatives. Ces effets peuvent invalider le modèle linéaire, un problème parfois désigné comme le «problème du mélange linéaire» (López-Baldomero *et al.*, 2025). Aussi, la signature spectrale d'un même matériau (encre, pigment) peut varier au sein d'un document en raison de facteurs tels que les différences de concentration, l'état de dégradation, ou les interactions avec le support. Cette variabilité peut conduire à des erreurs de modélisation ou à l'identification erronée de multiples sources pour un seul matériau. Les données MS sont d'ailleurs souvent affectées par le bruit, ce qui peut fausser les composantes extraites par la NMF. Un autre défi est le passage à l'échelle ; avec des images MS volumineuses, présentant une haute résolution spatiale et spectrale, des contraintes augmentant la complexité algorithmique et le besoin de détermination du nombre de composantes, l'automatisation de la décomposition par NMF devient particulièrement critique, certaines méthodes demandant le réglage de paramètres pour chaque image analysée (Rahiche *et al.*, 2019; Rahiche & Cheriet, 2022). Par ailleurs, l'absence de critères objectifs universels pour évaluer la qualité d'une décomposition complique l'établissement de protocoles standardisés pour le traitement en masse de collections documentaires. Ces limitations, partagées à divers degrés par l'ensemble des **méthodes algébriques traditionnelles** (*voir* Tableau 1.1 pour une synthèse comparative), soulignent la nécessité d'explorer des approches alternatives.

Tableau 1.1 Synthèse comparative des méthodes algébriques traditionnelles pour l'analyse de documents historiques multibandes

Caractéristique	ACP	ACI	K-moyennes	GMM	NMF
Objectif Principal	Maximisation de la variance, décorrélation	Maximisation de l'indépendance statistique	Regrouper les données en K groupes distincts, basés sur la proximité	Modéliser les données comme un mélange de K distributions Gaussiennes	Factorisation en matrices non-négatives, reconstruction additive
Application aux images MS de documents	Réduction de dimension, amélioration de contraste, pré-traitement	Séparation de couches, discrimination d'encres	Segmentation rapide en régions spectralement distinctes	Segmentation probabiliste ("douce")	Démélange spectral (encres, support, dégradations)
Statistiques Utilisées	Second ordre (covariance)	Ordres supérieurs (cumulants)	Distance (Euclidienne, Cosinus), Moyenne	Probabilités, EM, Moyenne et Covariance	Basé sur la fonction de coût
Hypothèses *a priori*	Aucune	Non-gaussiennes, indépendance statistique	K prédéfini, clusters sphériques	Données suivent un mélange de K Gaussiennes	Données Non-négatives, K sources distinctes
Avantages	Simple, rapide, bien établie	Meilleure décomposition que l'ACP	Simple, rapide	Assignation flexible	Interprétation physique directe, adaptée aux images MS
Inconvénients	Ne sépare pas les sources, interprétation difficile	Hypothèse d'indépendance forte, non-négativité non garantie	Sensible à l'initialisation, optimums locaux	Sensible aux variations, K à définir	Sensible à l'initialisation, K à définir
Contexte spatial	X	X	$\approx$ (SKKHM)	X	X
Non-négativité	X	X	X	X	V (pour E et A)
Interprétabilité Physique	Limitée (combinaisons linéaires, valeurs négatives)	Modérée (plus proche des sources, valeurs négatives)	Limitée (centroïdes = moyennes pas de spectres purs garantis)	Limitée (paramètres Gaussiens donc pas de spectres purs garantis)	Élevée (signatures spectrales et abondances non-négatives)

Les méthodes **d'apprentissage profond** émergent alors comme une voie prometteuse, offrant la capacité d'apprendre automatiquement des représentations non-linéaires complexes et de s'adapter aux spécificités de donnée hautement variables (Squires, 2019; Yu *et al.*, 2024).

1.3 Apprentissage profond pour l'extraction de composantes et l'analyse d'images

L'apprentissage profond (Deep Learning, DL) représente une branche de l'intelligence artificielle (IA) qui a transformé de nombreux domaines, notamment celui de l'analyse d'images. Sa capacité à modéliser des abstractions de haut niveau dans les données en utilisant des architectures composées de multiples couches de traitement non linéaire en fait un outil puissant pour l'extraction de composantes complexes.

Cette section introductive vise à définir l'apprentissage profond, à le situer par rapport à l'apprentissage automatique classique, et à introduire certains des concepts fondamentaux.

1.3.1 Définition et positionnement par rapport aux méthodes algébriques traditionnelles

L'apprentissage profond est une branche de l'apprentissage automatique (Machine Learning, ML) qui repose sur l'entraînement de réseaux de neurones artificiels à partir de grandes quantités de données. Contrairement aux **méthodes algébriques traditionnelles**, les méthodes d'apprentissage profond apprennent des représentations de données complexes de manière adaptative, sans hypothèses fortes sur la structure sous-jacente. Les modèles d'apprentissage profond, sont conçus pour apprendre des caractéristiques complexes de manière hiérarchique et autonome, directement à partir des données brutes. Leur structure algorithmique en couches permet un traitement progressif de l'information, où chaque couche affine les résultats de la précédente, facilitant l'identification de motifs complexes souvent sans étiquetage préalable exhaustif. Dans le contexte de l'analyse d'images, cette approche se distingue par l'extraction automatique des caractéristiques. Contrairement aux méthodes traditionnelles qui exigent une sélection manuelle des caractéristiques pertinentes (couleur, contours, texture) par un expert, l'apprentissage profond adopte un concept *d'apprentissage de bout en bout*. L'algorithme identifie automatiquement les caractéristiques les plus saillantes pour chaque classe en analysant des exemples, rendant cette approche particulièrement adaptée à la gestion d'une grande variabilité des objets et des conditions d'acquisition. Il est important de noter que les modèles d'apprentissage profond sont généralement plus complexes et de plus grande taille que leurs homologues classiques. Ils requièrent par conséquent des volumes de données plus importants pour l'entraînement et une puissance de calcul considérable, les unités de traitement graphique (GPU) étant souvent privilégiées pour leur capacité à effectuer des calculs parallèles massifs.

1.3.2 Réseaux de Perceptron Multicouche (MLP) : La base

Le **Perceptron Multicouche** (Multilayer Perceptron - MLP) est l'une des architectures fondamentales des réseaux de neurones profonds. Un MLP est structuré en plusieurs couches de neurones interconnectés :

1. **La couche d'entrée** reçoit les données initiales. Chaque neurone de cette couche correspond à une caractéristique d'entrée (*p. ex.*, valeurs d'un pixel si une image est aplatie en vecteur).
2. Les **couches cachées** se situent entre la couche d'entrée et la couche de sortie. Elles effectuent des transformations successives sur les données. Dans un MLP, chaque neurone d'une couche est typiquement connecté à tous les neurones de la couche précédente (on parle de couches «entièrement connectées» ou fully connected) et transmet sa sortie à la couche suivante. Le nombre de couches cachées et le nombre de neurones dans chaque couche sont des hyperparamètres importants du modèle, définis lors de sa conception.
3. **La couche de sortie** produit la prédiction finale du réseau. Le nombre de neurones dans cette couche dépend de la nature de la tâche : par exemple, un seul neurone peut être utilisé pour une classification binaire (avec une fonction d'activation sigmoïde), tandis que plusieurs neurones (avec softmax) sont nécessaires pour une classification multi-classes.

Mathématiquement, la **propagation avant** dans un MLP peut être décrite par les équations suivantes. Pour une couche l avec $n^{(l)}$ neurones, la sortie est calculée comme :

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \tag{1.4}$$

$$\mathbf{a}^{(l)} = f(\mathbf{z}^{(l)}) \tag{1.5}$$

où $\mathbf{W}^{(l)} \in \mathbb{R}^{n^{(l)} \times n^{(l-1)}}$ est la matrice des poids de la couche l, $\mathbf{b}^{(l)} \in \mathbb{R}^{n^{(l)}}$ est le vecteur de biais, $\mathbf{a}^{(l-1)}$ est la sortie de la couche précédente, et $f(\cdot)$ est la fonction d'activation (*p. ex.*, ReLU, sigmoïde, tanh, etc.). L'entraînement du réseau s'effectue par **rétropropagation des gradients**, qui minimise une fonction de coût $\mathcal{L}$ en ajustant les paramètres. La nature «entièrement connectée» des MLP, bien que simple à conceptualiser, est à la fois une force et sa faiblesse majeure. Elle offre la capacité de modéliser des interactions complexes si le réseau est

suffisamment profond, mais elle conduit aussi à une explosion du nombre de paramètres, en particulier pour les entrées de haute dimension comme les images. Ce choix architectural est à l'origine de ses limitations dans les tâches de vision. En effet, si chaque neurone d'une couche est connecté à ceux de la suivante, le nombre de poids est le produit du nombre de neurones dans ces couches adjacentes. Pour une image de taille modeste (*p. ex.*, 224×224 pixels soit 50176 valeurs d'entrée), la première couche cachée, même si comportant un nombre raisonnable de neurones, aura un nombre massif de poids à apprendre, entraînant des coûts de calcul et des besoins en mémoire élevés. La structure du MLP est donc mal adaptée aux données d'image brutes de haute dimension. De plus, avec l'aplatissement en vecteurs unidimensionnels, la structure spatiale des images est perdue, poussant vers le développement d'architectures spécialisées comme les CNN.

1.3.3 Réseaux de neurones convolutifs (CNN) : Le contexte spatial

Les **Réseaux de neurones convolutifs (CNN)** exploitent la structure spatiale des images grâce à la convolution, qui constitue l'opération mathématique centrale de leur architecture. Pour des images multibandes, cette opération s'étend sur tous les canaux. Mathématiquement, pour une image d'entrée $X \in \mathbb{R}^{H \times W \times C}$ et un ensemble de filtres (ou kernel) $K \in \mathbb{R}^{F_H \times F_W}$, l'opération de convolution 2D discrète est définie par :

$$(X * K^k)_{i,j} = \sum_{c=0}^{C-1} \sum_{m=0}^{F_H-1} \sum_{n=0}^{F_W-1} X_{i+m,j+n,c} \cdot K^k_{m,n,c} + b^k \tag{1.6}$$

où C est le nombre de canaux d'entrée, b est le terme de biais, et où $k \in \{1, ..., N_f\}$ est l'indice du filtre produisant N_f cartes de caractéristiques en sortie et où (i, j) représente la position dans ces cartes. La taille de ces cartes dépend de deux hyperparamètres :

- **Le Stride** s : le pas entre chaque déplacement du filtre.
- **Le Padding** p : le nombre de valeurs ajoutées autour de l'image pour le calcul des bordures.

La dimension de l'image de sortie, dépendante de s et p, est alors calculée comme :

$$H_{out} = \left\lfloor \frac{H + 2p - F_H}{s} \right\rfloor + 1 \quad , \quad W_{out} = \left\lfloor \frac{W + 2p - F_W}{s} \right\rfloor + 1 \tag{1.7}$$

Une autre composante importante associé aux CNN sont les **couches de pooling**, qui réduisent la dimensionnalité spatiale et contribuent à l'invariance translationnelle (Ng *et al.*, 2014). Le **partage de poids** permet de réduire drastiquement le nombre de paramètres, en utilisant les mêmes valeurs de filtre pour toute l'image. Contrairement à un MLP où chaque connexion a un poids unique, un CNN avec N_f filtres de taille $F \times F$ sur C canaux d'entrée n'a que $N_f \times (F \times F \times C + 1)$ paramètres, indépendamment de la taille de l'image d'entrée. Cette propriété permet aux CNN de traiter des images de différentes tailles avec le même modèle. Aussi, **l'invariance translationnelle**, obtenue par la nature glissante de la convolution, permet de reconnaître les objets indépendamment de leur position exacte dans l'image. L'architecture CNN construit une **hiérarchie de caractéristiques** : les premières couches détectent des éléments simples (*p. ex.*, contours, textures), grâce à des champs récepteurs locaux de petite taille. Les couches intermédiaires assemblent ces éléments en formes plus complexes avec des champs récepteurs plus larges, et les couches profondes reconnaissent des objets entiers ou des concepts abstraits. Cette progression peut être formalisée par la croissance du champ récepteur effectif :

$$RF_l = RF_{l-1} + (K_l - 1) \times \prod_{i=1}^{l-1} S_i \tag{1.8}$$

où RF_l est le champ récepteur à la couche l, K_l la taille du noyau et S_i le stride de la couche i. La Figure 1.3 illustre un CNN simple suivi de couches de MLP pour une tâche de classification de chiffres manuscrits. Cette approche hiérarchique s'inspire du système visuel biologique et a prouvé son efficacité pour diverses tâches de vision, en étant à la base d'architectures compétitives plus complexes comme les modèles LeNet, AlexNet, VGG, GoogLeNet ou encore le ResNet (Bhatt *et al.* (2021) propose une revue complète de l'évolution de ces architectures). Hollaus, Brenner & Sablatnig (2019) propose d'utiliser un ResNet supervisé pour la binarisation d'images de documents, les résultats confirmant l'efficacité des modèles CNN pour cette tâche.

1.3.4 Auto-encodeurs : Apprentissage non-supervisé de représentations

Les **Auto-encodeurs (AE)** constituent une famille d'architectures de réseaux de neurones principalement utilisée pour l'apprentissage non supervisé de représentations de données. A

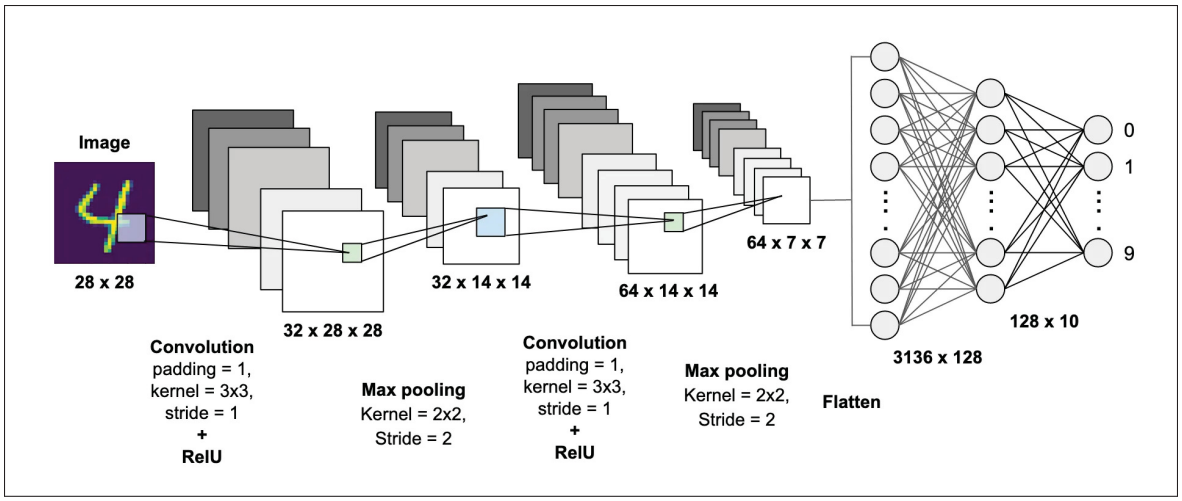

Figure 1.3 Exemple d'application d'un CNN pour la tâche de classification de chiffres manuscrits. Figure tirée de Patel (2019)

l'instar de la NMF, leur objectif est d'apprendre des représentations vectorielles (ou embeddings) compactes des données, souvent dans un espace de dimensionnalité réduite, sans supervision. Le principe de base d'un autoencodeur s'articule autour de deux composantes principales :

1. **L'encodeur** a pour rôle de transformer les données d'entrée en une représentation de plus faible dimensionnalité. L'encodeur apprend à extraire les informations les plus saillantes et pertinentes de l'entrée pour former cette représentation, aussi appelée espace latent.
2. **Le décodeur** prend ensuite cette représentation latente en entrée et tente de reconstruire les données d'origine aussi fidèlement que possible.

L'ensemble du réseau est entraîné en minimisant une fonction de reconstruction, qui mesure la différence entre les données d'entrée originales et les données reconstruites par le décodeur. Le goulot d'étranglement contraint la représentation latente à avoir une dimensionnalité inférieure à celle de l'entrée, forçant le réseau à apprendre une représentation compressée qui capture l'essence des données, plutôt que de simplement apprendre une fonction identité.

Les **Auto-encodeurs Convolutifs (CAE)** adaptent l'architecture des auto-encodeurs en intégrant des couches convolutives. L'encodeur d'un CAE agit comme un CNN classique, extrayant une hiérarchie de caractéristiques spatiales de l'image d'entrée. Le décodeur effectue l'opération

inverse pour reconstruire l'image à partir de la représentation latente. Cette architecture est particulièrement adaptée pour les images de document car elle préserve l'information spatiale et apprend des filtres qui capturent des motifs locaux (Calvo-Zaragoza & Gallego, 2019).

Les **Auto-encodeurs Variationnels (VAE)** sont une extension des auto-encodeurs qui appartiennent à la catégorie des modèles génératifs. De manière analogue à la NMF qui possède une variante basée sur la divergence KL, les auto-encodeurs ont leur extension probabiliste sous la forme des VAE. Leur particularité est de représenter explicitement l'espace latent comme une distribution de probabilité $\mathcal{N}(\mu, \sigma^2 I)$, permettant la génération de données synthéthiques réalistes. La Figure 1.4 illustre une piste d'application de VAE pour la SAS, visant à distinguer des textes superposés. Bien que prometteuse, cette approche présente des limitations notables, puisque le modèle ne fonctionne que sur des images monobandes de faible résolution et que son entraînement requiert un vaste corpus de données déjà séparées (Neri, Badeau & Depalle, 2021).

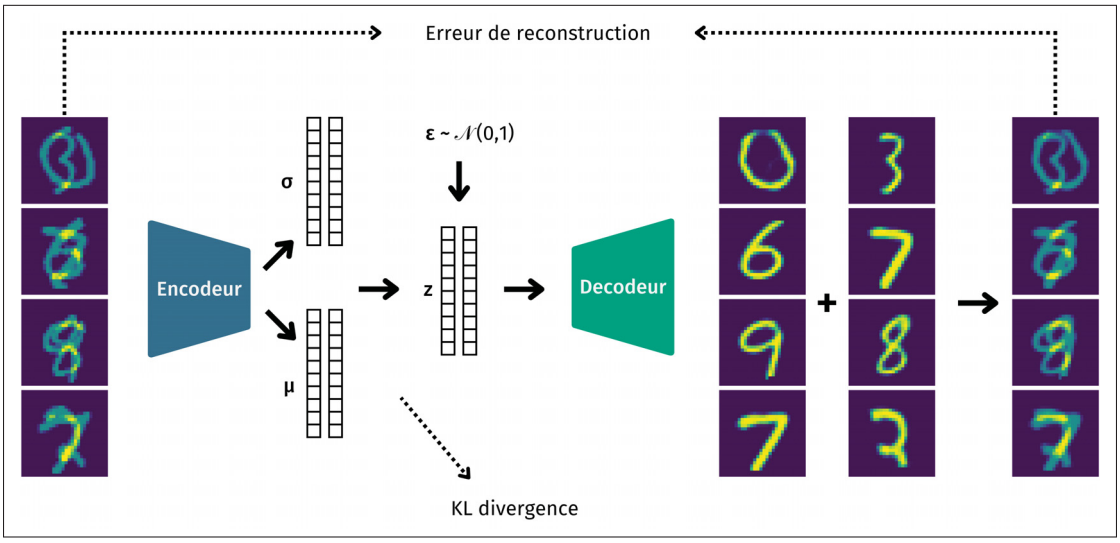

Figure 1.4 Exemple d'application d'un Autoencodeur Variationnel (VAE) pour la tâche de séparation de chiffres manuscrits. Réimplémentation de Neri *et al.* (2021)

1.3.5 Transformers en vision (ViT) : Le contexte global

Initialement conçus pour le traitement automatique du langage naturel (Natural Langage Processing - NLP), les **Transformers** ont récemment révolutionné la vision par ordinateur après

leur introduction par Vaswani *et al.* (2017). Le cœur des modèles Transformer est le mécanisme **d'auto-attention**. Ce mécanisme permet au modèle de pondérer l'importance de différentes parties de la séquence d'entrée les unes par rapport aux autres, afin de calculer une représentation de chaque élément de la séquence qui tienne compte du contexte global. Contrairement aux CNN qui construisent un contexte global progressivement, les transformers ont la capacité de capturer des relations et dépendances de longue portée entre tous les éléments dès les premières couches du réseau. Chaque élément peut «regarder» tous les autres afin de trouver les plus pertinents pour sa propre représentation. Cependant, l'attention possède une haute complexité computationnelle, quadratique par rapport à la longueur de la séquence d'entrée. Pour adapter cette architecture aux images, le **Vision Transformer (ViT)** divise alors l'image en patchs de taille fixe, evitant ainsi de traiter chaque pixels. Les ViT ont prouvé leur capacité à extraire une compréhension sémantique des images à partir de grands ensembles de données, apportant les meilleurs performance en segmentation d'images. Contrairement aux CNN qui extraient des relations locales, les ViT parviennent à regrouper des objets par leur sens ou contexte, permettant une meilleure compréhension globale d'une scène. Des variantes comme le **Swin Transformer** adaptent le mécanisme d'attention pour améliorer l'efficacité computationnelle, permettant de traiter des images plus grandes tout en conservant une compréhension sémantique globale. Cependant, le découpage en patchs inhérent aux ViT sacrifie la résolution spatiale fine, les rendant moins adaptés pour l'analyse de documents haute résolution nécessitant une segmentation précise au pixel près. Cela explique l'échec des methodes de fondations pour la segmentation et l'extraction de texte manuscrit (*voir* Figure 1.5).

1.3.6 Applications pratiques aux images de documents

Les architectures d'apprentissage profond trouvent des applications spécifiques dans l'analyse de documents. Les CNN excellent dans la détection et la reconnaissance de texte (OCR), la segmentation de mise en page, et l'extraction de tableaux. Les AE sont utilisés pour le débruitage de documents anciens ou de mauvaise qualité, la compression de documents, l'extraction de caractéristiques ou la séparation de textes superposés (*voir* Figure 1.4). Les ViT avec leur capacité

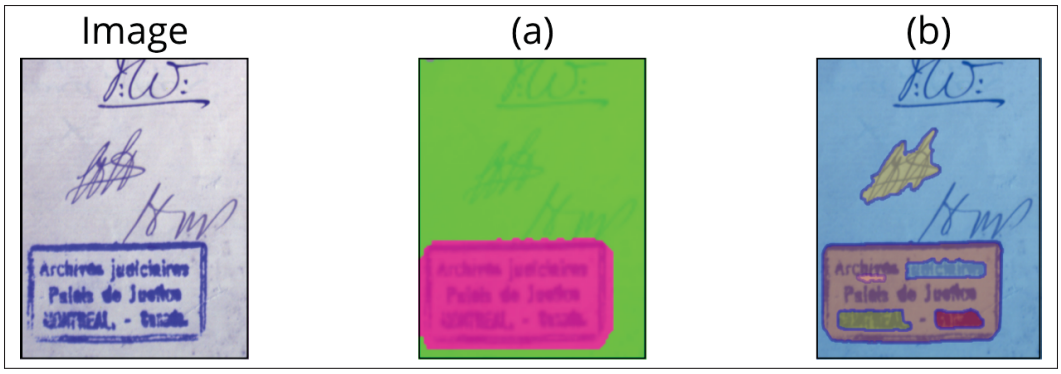

Figure 1.5 Exemple d'application de deux ViT reconnus, pour la tâche de segmentation automatique d'une image de texte, avec (a) Dinov2 (Oquab *et al.*, 2024) et (b) Segment Anything Model (SAM) (Kirillov *et al.*, 2023)

à capturer le contexte global, montrent des performances prometteuses pour la compréhension de la structure de documents et l'analyse de leur contenu, comme démontré sur la tâche de DocVQA (Mathew, Karatzas & Jawahar, 2021) mais restent limités pour une segmentation fine.

1.3.7 Apprentissage supervisé contre non-supervisé

La distinction entre apprentissage supervisé et non-supervisé constitue alors un choix fondamental dans la conception de modèles pour l'analyse d'images. Le premier, qui s'appuie sur des paires entrée-sortie annotées, permet d'obtenir des performances remarquables lorsque les données étiquetées sont abondantes et représentatives. Dans le contexte de l'analyse documentaire, cette approche excelle pour des tâches bien définies comme la segmentation sémantique ou la classification de régions. Cependant, la création d'annotations précises pour les images MS représente un défi majeur : elle requiert une expertise spécialisée, s'avère chronophage et coûteuse, et peut introduire des biais liés à l'interprétation subjective des experts. À l'inverse, l'apprentissage non-supervisé, comme utilisé par les auto-encodeurs, explore la structure intrinsèque des données pour automatiquement découvrir des motifs latents et des décompositions naturelles, le rendant particulièrement adapté aux contextes où celles-ci sont rares ou inexistantes.

1.4 Identification des lacunes et justification d'une approche hybride

La revue des méthodes existantes révèle un paysage contrasté où chaque approche présente des forces et limitations complémentaires :

Les méthodes **algébriques traditionnelles**, notamment la NMF, ont démontré leur efficacité pour la décomposition non supervisée d'images MS. Leur principal atout réside dans l'interprétabilité directe des composantes extraites, chaque facteur pouvant être associé à un matériau ou une caractéristique physique spécifique, essentielle pour l'analyse de documents patrimoniaux. Cependant, elles souffrent de limitations face à la complexité des données réelles : nécessité de définir le rang de factorisation a priori, difficulté à capturer les relations spatiales complexes, et capacité limitée à modéliser des phénomènes non linéaires.

À l'inverse, les modèles **d'apprentissage profond**, notamment les auto-encodeurs convolutifs (CAE), excellent dans l'extraction automatique de caractéristiques complexes et la modélisation de relations non linéaires. Leur capacité à apprendre des représentations hiérarchiques directement à partir des données leur confère une puissance de modélisation supérieure. Néanmoins, cette puissance s'accompagne d'un manque d'interprétabilité : les représentations latentes apprises sont souvent opaques et difficiles à relier aux propriétés physiques des matériaux analysés.

Face à ces constats, une **approche hybride** combinant l'interprétabilité de la factorisation matricielle avec la puissance d'extraction des réseaux de neurones émerge comme une solution naturelle. En intégrant les contraintes de non-négativité garantes **d'interprétabilité physique** avec les capacités de **modélisation non linéaire et spatiale** des CAE, il devient possible de concevoir un modèle préservant la transparence des résultats tout en bénéficiant de la richesse représentationnelle de l'apprentissage profond. Cette synergie permettrait de répondre aux exigences spécifiques de l'analyse non supervisée de documents MS : résultats interprétables pour les experts, robustesse aux données bruitées, et fonctionnement avec des ensembles de données limités. L'originalité de cette direction réside dans la fusion de deux paradigmes traditionnellement distincts, ouvrant la voie au développement présenté dans le chapitre suivant.

CHAPITRE 2

APPROCHE HYBRIDE AUTOENCODEUR-FACTORISATION NMF

L'une des révolutions apportées par l'**apprentissage profond** est sa capacité à apprendre une hiérarchie de caractéristiques. Plutôt que d'analyser les données sur un seul niveau, un modèle profond extrait des concepts de complexité croissante à travers ses différentes couches. Les premières couches apprennent des caractéristiques simples et locales (des contours, des textures, des gradients de couleur), tandis que les couches suivantes les combinent pour former des motifs plus complexes (des parties d'objets), jusqu'à la reconnaissance d'objets entiers dans les couches finales. Cette **approche hiérarchique** est directement inspirée du fonctionnement du cortex visuel humain et permet de modéliser des **relations non-linéaires et complexes**, bien au-delà des capacités d'un modèle *plat* comme la NMF classique.

Face à ce constat, l'idée de rendre la NMF multicouche a émergé suivant deux paradigmes à distinguer, tout deux pourtant désignés sous le terme «**NMF profonde**» :

1. **La NMF profonde par factorisation en cascade :** Cette première approche est une extension directe du modèle NMF en une structure multi-couches. La matrice obtenue à une couche k sert de matrice d'entrée pour la factorisation à la couche $k+1$, instaurant ainsi une décomposition en cascade. Le principe repose sur une factorisation hiérarchique où chaque niveau de décomposition affine les représentations du niveau précédent, dans le but d'obtenir une hiérarchie de facteurs où les parties sont combinées pour former des touts. La matrice décomposée peut être la matrice d'abondance, comme $\mathbf{Y} \approx \mathbf{E}_1\mathbf{E}_2 \ldots \mathbf{E}_k\mathbf{A}_k$ où $\mathbf{A}_{i-1} \approx \mathbf{E}_i\mathbf{A}_i$; la matrice des signatures spectrales, comme $\mathbf{Y} \approx \mathbf{E}_k\mathbf{A}_k \ldots \mathbf{A}_2\mathbf{A}_1$ où $\mathbf{E}_{i-1} \approx \mathbf{E}_i\mathbf{A}_i$; ou les deux. Toutefois, bien que hiérarchique, ce modèle demeure fondamentalement une technique de factorisation matricielle. Il nécessite un entraînement en deux phases, d'abord un pré-entraînement de chaque de factorisation, suivi d'une étape de «fine-tuning» apportant une cohérence globale. Il n'intègre pas nativement de mécanismes pour l'apprentissage de dépendances spatiales et est dépourvu de la flexibilité architecturale des réseaux de neurones modernes, ce qui le rend moins optimal pour le traitement direct de données d'images.

2. **L'intégration de contraintes de type NMF au sein d'architectures neuronales profondes :** Cette seconde approche, plus flexible, consiste à s'appuyer sur des architectures neuronales connexes de l'état de l'art, telles que les auto-encodeurs, et d'y intégrer les principes de la NMF non comme un bloc architectural, mais comme un ensemble de contraintes de régularisation. L'objectif n'est donc pas d'empiler des modules de factorisation, mais de contraindre un réseau puissant à reproduire des représentations interprétables similaires à la NMF. Cette contrainte se matérialise souvent par l'imposition de la non-négativité sur les poids du modèle, sur les cartes d'activation ou sur les fonctions d'activation de l'espace latent. Le modèle est alors optimisé de bout en bout (end-to-end) par rétropropagation des gradients en respectant ces contraintes. Cette méthodologie présente des avantages clés : elle permet d'exploiter les opérateurs convolutifs pour prendre en compte les dépendances spatiales et offre la flexibilité d'intégrer ces contraintes dans des architectures modernes robustes et éprouvées, bénéficiant de leur meilleure capacité de généralisation.

Une revue détaillée des différents types de NMF ***profondes*** est proposé par Chen, Zeng & Pan (2022). Face à notre problématique ; la décomposition d'images multi-bandes de documents historiques, le second paradigme s'avère le plus pertinent. En effet, cette tâche requiert un modèle capable de généraliser et de s'adapter à la grande variabilité inhérente à ces données. Notre objectif est de capitaliser sur la puissance et la flexibilité des réseaux de neurones convolutifs, tout en leur injectant l'interprétabilité physique de la NMF au moyen de contraintes ciblées.

2.1 Fondements et synergies entre NMF et réseaux de neurones

Les mathématiques de l'auto-encodeur, lorsque certaines contraintes sont imposées, ont des caractéristiques similaires à celles de la NMF. En considérant une matrice d'entrée $\mathbf{Y} \in \mathbb{R}^{b \times n}$ avec b dimensions et n pixels, un auto-encodeur peut traiter un point de données $\mathbf{y}_i \in \mathbb{R}^{b \times 1}$ en cherchant à reconstruire $\hat{\mathbf{y}}_i \in \mathbb{R}^{b \times 1}$. Si l'on définit un nombre de neurones r dans l'espace latent comme $r < b$, le modèle est forcé d'apprendre une représentation simplifiée de la donnée. Il se sert de celle-ci pour reconstruire une prédiction en sortie, permettant ainsi d'évaluer sa qualité.

Le nombre de poids appris par une couche de neurones connectés est défini comme le produit du nombre de neurones dans la couche précédente par le nombre dans la couche suivante. Pour le décodeur, il s'agit donc du produit du nombre de neurones dans l'espace latent, r, par le nombre neurones dans la couche de sorties, b. On peut alors noter ces poids comme une matrice $\mathbf{W}_d^T \in \mathbb{R}^{r \times b}$, de taille similaire à la matrice des composantes $\mathbf{M} \in \mathbb{R}^{r \times b}$ dans la NMF. D'une manière similaire, la matrice de poids reliant l'entrée à la couche cachée est $\mathbf{W}_e \in \mathbb{R}^{r \times b}$ où r est le nombre de neurones dans la couche cachée. Nous pouvons alors exprimer la sortie de la couche cachée $h_i = \sigma_e(\mathbf{W}_e \mathbf{y}_i)$, où σ_e est une fonction non-linéaire qui opère élément par élément et $h_i \in \mathbb{R}^{r \times 1}$. La Figure 2.1 illustre un modèle simple de type auto-encodeur.

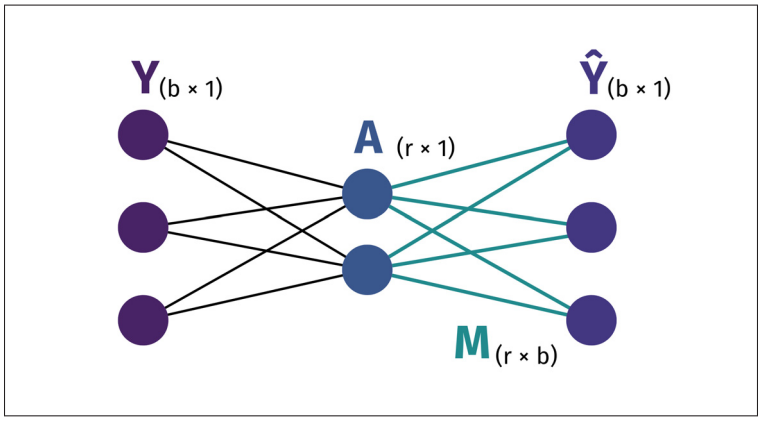

Figure 2.1 La NMF vue comme un réseau neuronal

Il est alors possible de traiter tous les pixels d'une image par ce réseau de neurones pour obtenir n vecteurs $\mathbf{h}_i$. On peut alors construire $\mathbf{H}^T \in \mathbb{R}^{n \times r}$, de taille similaire aux cartes d'abondances $\mathbf{A} \in \mathbb{R}^{n \times r}$. La couche finale produit alors la sortie $\hat{\mathbf{Y}} = \sigma_d(\mathbf{W}_d \mathbf{H})$. Si la fonction d'activation σ_d est l'identité, et que $\mathbf{W}_d$ et $\mathbf{H}$ sont tous deux non-négatifs, nous avons un auto-encodeur qui peut être interprété comme effectuant une NMF. Pour qu'un auto-encodeur de base effectue une NMF standard, les contraintes suivantes doivent être respectées (Squires, 2019) :

- La fonction d'activation σ_e doit produire une sortie non-négative.
- La fonction d'activation σ_d doit être l'identité.
- Les poids de $\mathbf{W}_d$ doivent être non-négatifs.
- Le nombre d'unités cachées dans la couche latente doit être le même que le rang désiré.

Contrairement à la NMF classique où l'optimisation est formulée à partir de la mesure de reconstruction ; soit la divergence Kullback-Leibler, soit l'erreur quadratique moyenne, cette formulation permet l'utilisation de différentes fonctions de coût et l'utilisation d'optimiseur robuste pour les réseaux de neurones, comme Adam (Kingma & Ba, 2017).

2.2 État de l'art des approches combinant NMF et réseaux de neurones

La fusion de la NMF avec des architectures d'apprentissage profond spécifiques a donné naissance à une nouvelle génération de modèles hybrides, capitalisant sur les forces respectives de chaque approche. Les réseaux de neurones convolutifs (CNN), récurrents (RNN) et les mécanismes d'attention ont été intégrés de manière innovante pour aborder des problèmes complexes dans divers domaines. Cependant les progrès rapides et itératifs des architectures neuronales de type auto-encodeurs, notamment appliqués au domaine du démélange HS, ont créé une scission entre la NMF classique et ces architectures semblables à des NMF «profondes». La Figure 2.2 illustre cette scission entre des papiers appartenant à ces deux communautés.

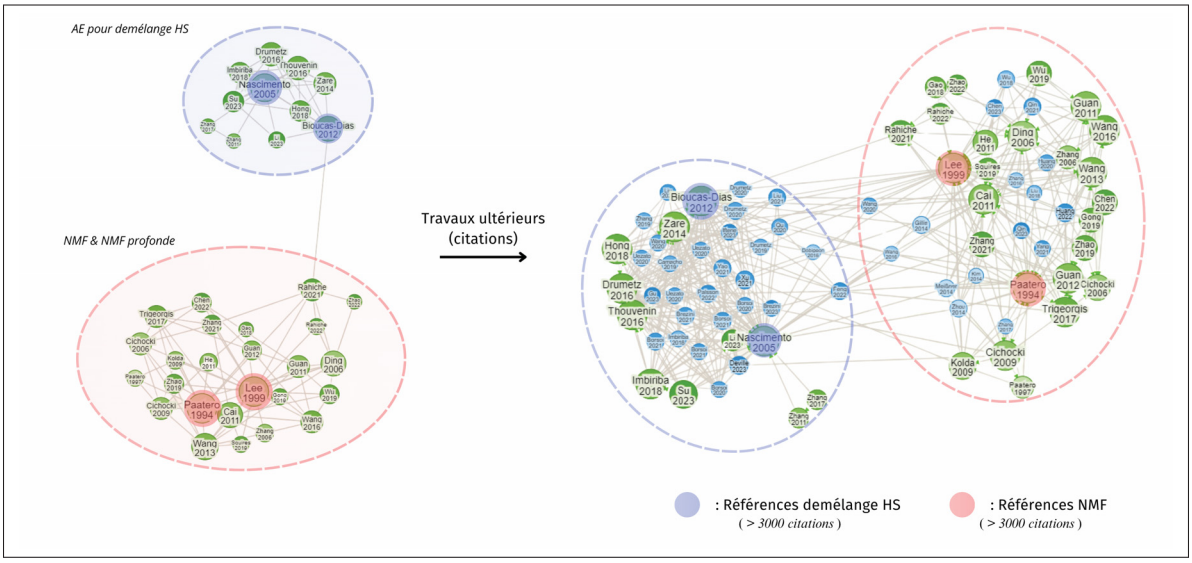

Figure 2.2 Visualisation de la scission entre la recherche sur le démélange HS et sur la NMF. Chaque noeud représente un papier tandis que chaque lien représente une citation. Les travaux des deux communautés aboutissent au développement de modèle AE non-négatifs bien qu'ayant eu un développement distinct avec peu d'interactions. Visualisation réalisée à l'aide de l'outil ResearchRabbit

Cette scission a engendré une redondance de certains développements, créant deux «communautés» de recherches avec peu de lien entre elles, mais qui ont abouti toutes les deux au développement de modèles de type AE non-négatifs. C'est notamment le cas des travaux de Ngoc-Diep (2008) et de Squires (2019) pour la NMF ; ou des travaux de Palsson, Sveinsson & Ulfarsson (2022) pour le démélange HS. Ce dernier papier propose notamment une critique comparative de différents modèles AE développés pour le démélange HS et note que ces modèles HS performent implicitement une NMF, sans toutefois faire de rapprochement entre les deux communautés. La suite de cet état-de-l'art présentera donc différents modèles NMF hybrides, ainsi que différents modèles AE non-négatifs, qui seront comparés et catégorisés de manière interchangeable.

Le traitement des données multispectrales peut être réalisé selon plusieurs stratégies, qui se distinguent principalement par la manière dont elles exploitent l'information spatiale et spectrale contenue dans le cube de données. Trois grandes familles d'approches se dégagent, comme illustré à la Figure 2.3, traitant les images par pixels, par bandes ou directement comme un cube.

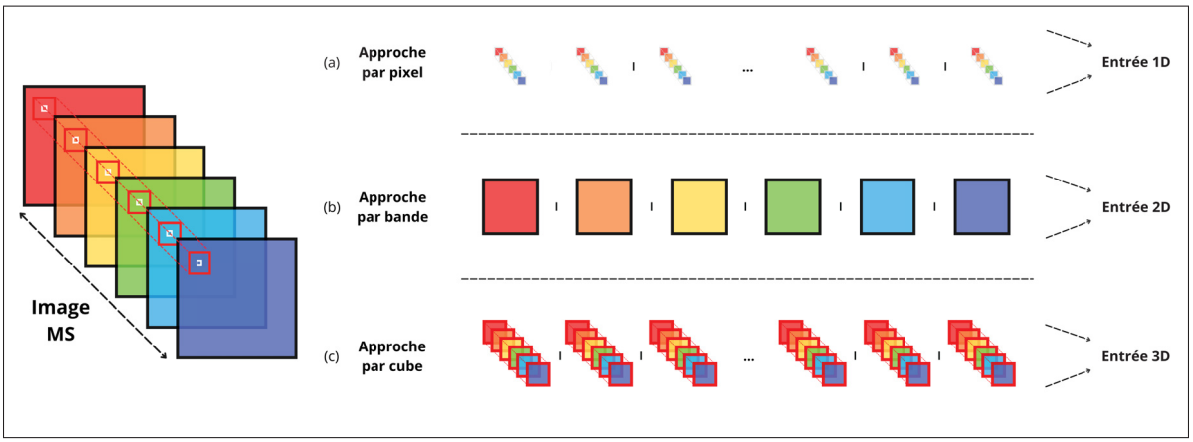

Figure 2.3 Schématisation de trois approches pour le traitement d'un cube MS : (a) approche traitant chaque pixel comme un vecteur spectral, (b) approche traitant chaque bande spectrale distinctement, et (c) approche traitant une image ou des patchs de l'image directement comme un cube spatio-spectral

2.2.1 L'approche par pixel

C'est l'approche la plus classique et la plus intuitive. Elle considère chaque pixel de manière isolée et l'analyse repose uniquement sur son vecteur spectral, qui représente l'intensité lumineuse à travers les différentes bandes. Chaque pixel est ainsi traité comme une signature indépendante, ignorant toute information sur son contexte spatial (ses pixels voisins). Les auto-encodeurs utilisés dans ce cadre sont typiquement des MLP, dont l'origine remonte aux travaux fondateurs de Rumelhart, Hinton, Williams *et al.* (1985) et Bourlard & Kamp (1988). L'idée est de forcer le réseau à apprendre une représentation compressée de la signature spectrale d'un pixel.

Dans le contexte du démélange HS, cette représentation latente est conçue pour correspondre aux abondances des matériaux purs. Les premières applications de réseaux de neurones pour le démélange par pixel, comme celles de Licciardi & Del Frate (2011), utilisaient des auto-encodeurs non linéaires pour la réduction de dimensionnalité. Plus récemment, des modèles d'apprentissage profond ont été proposés. Par exemple, Palsson, Sigurdsson, Sveinsson & Ulfarsson (2018) ont développé un autoencodeur profond où le décodeur est contraint pour représenter le modèle de mélange linéaire, et dont les poids correspondent aux endmembers. De même, les travaux de Su *et al.* (2019) proposent un réseau d'auto-encodeurs profonds (DAEN) en deux étapes : des auto-encodeurs empilés pour initialiser les signatures spectrales, suivis d'un autoencodeur variationnel (VAE) pour estimer simultanément les signatures et les abondances, tout en garantissant les contraintes de non-négativité (ANC) et de somme à l'unité (ASC).

L'avantage de cette approche est sa simplicité et son coût de calcul relativement faible. Cependant, son incapacité à exploiter la corrélation spatiale la rend sensible au bruit et moins performante dans les scènes à forte variabilité spatiale.

2.2.2 L'approche par bande

Cette méthode traite les données en considérant chaque bande spectrale comme une image 2D indépendante. Le traitement est donc appliqué sur chaque «canal», ce qui permet d'utiliser des algorithmes de traitement d'images 2D traditionnels. Une stratégie proposée par Zhou,

Hang, Liu & Yuan (2019) pour la classification, consiste à traiter les canaux spectraux comme une séquence temporelle à l'aide d'un réseau LSTM (Long Short-Term Memory) pour modéliser les dépendances spectrales, ajoutant une dimension spatiale aux encodeurs par pixels. Dans le domaine de la binarisation de documents, Calvo-Zaragoza & Gallego (2019) utilise un autoencodeur convolutif 2D pour transformer une image en une carte de probabilité d'appartenance au premier plan ou à l'arrière-plan.

L'avantage principal de cette approche est de pouvoir capitaliser sur des architectures 2D très matures et performantes. Cependant, en traitant chaque bande séparément ou séquentiellement, elle risque de ne pas capturer efficacement les corrélations subtiles et complexes qui existent entre les différentes bandes spectrales, qui sont pourtant au cœur de l'analyse multibande.

2.2.3 L'approche par cube

Plus récente et souvent plus performante, cette approche exploite simultanément les informations spatiale et spectrale. Elle analyse directement des sous-volumes 3D (appelés *patchs*) extraits du cube de données. Cette stratégie permet de prendre en compte à la fois la signature spectrale d'un pixel et les caractéristiques spatiales de son voisinage.

Cette approche est la plus naturelle pour les auto-encodeurs convolutifs. Ces modèles utilisent des filtres de convolution capables d'apprendre des caractéristiques directement depuis le cube de données, capturant ainsi les dépendances locales à la fois dans le domaine spatial et spectral. L'évolution des travaux de Palsson, Ulfarsson & Sveinsson (2021) est assez emblématiques de cette catégorie. Ils proposent un autoencodeur convolutif spécifiquement conçu pour le démélange hyperspectral. Le réseau apprend à extraire les signatures spectrales et les cartes d'abondances en analysant des patchs 3D, ce qui le rend beaucoup plus robuste au bruit et améliore significativement la qualité du démélange par rapport aux approches pixel par pixel. Le modèle encode un patch 3D en une série de cartes d'abondances 2D, qu'il décode pour reconstruire le patch original. L'article de synthèse de Palsson *et al.* (2022) analyse et compare d'ailleurs plusieurs de ces modèles, soulignant la supériorité des approches spectro-spatiales.

Cette méthode offre le meilleur compromis en exploitant toute la richesse de l'information disponible, tout en gardant un coût de calcul bas grâce à l'utilisation de CNN. En revanche, un de ses points faible réside dans la limitation du champ récepteur effectif du CNN. En effet, pour que les cartes d'abondances conservent la même résolution spatiale que l'image d'entrée, ces réseaux évitent les couches de *pooling*, ce qui restreint leur analyse à un contexte purement local.

2.2.4 Analyse de l'état de l'art et perspectives

L'analyse des trois grandes familles d'approches, résumée dans le Tab. 2.1, met en évidence la supériorité des méthodes par cube pour le traitement de données MS. En exploitant à la fois l'information spatiale et spectrale, ces modèles, notamment basés sur des AE convolutifs, offrent les performances les plus robustes, comme souligné par la synthèse de Palsson *et al.* (2022).

Tableau 2.1 Tableau comparatif de différentes approches hybrides factorisation–autoencodeur

Méthode	**Approche**	**Spectral**	**Spatial**	**Non-local**	**Contraintes**	**Coût**	**Rang**
Licciardi *et al.* (2011)	Pixel	+	-	-	/	MSE	-
Liu *et al.* (2017)	Bande	+	+	-	/	Cross-entropy	-
Flenner *et al.* (2017)	Cube	-	+	-	ANC	ACC	-
Palsson *et al.* (2018)	Pixel	+	-	-	ANC, ASC Parcimonie	SAD	-
Palsson *et al.* (2019)	Pixel	+	+	-	ASC, ANC	SAD	-
Squires (2019)	Pixel	+	-	-	ANC	MSE, RAE NRAE	-
Su *et al.* (2019)	Pixel	+	-	-	ASC, ANC	MSE	-
Mei *et al.* (2019)	Cube	+	+	-	/	MSE	-
Debain (2020)	Bande	-	+	-	ANC	$\mathcal{B}$ divergence	-
Palsson *et al.* (2021)	Cube	+	+	-	ASC, ANC	SAD, MSE	-
Zhao *et al.* (2022)	Cube	+	+	-	ASC, ANC	MSE	-
Li *et al.* (2023)	Cube	+	+	-	ASC, ANC	SAD Cross-entropy	-
Alfaro-Meija (2023)	Cube	+	+	-	ASC, ANC	Cross-entropy	-
Su *et al.* (2023)	Cube	+	+	+	ASC, $\mathcal{L}_{SHC}$ Uniformité	SAD	-
Zheng *et al.* (2024)	Cube	+	-	-	ASC, ANC	SAD, Log SAD	-
Su *et al.* (2024)	Cube	+	+	-	ANC, ASC Uniformité	SAD	-

Cependant, une limitation majeure de ces modèles réside dans leur champ récepteur effectif limité, souvent restreint par l'absence de couches de *pooling*, nécessaires pour préserver la résolution spatiale des cartes d'abondance. Cette contrainte devient particulièrement problématique pour des données qui, comme les images MS de documents, possèdent une dimension spatiale importante, significativement plus grande que celle des images HS typiques (voir Fig. 0.3). Une perspective de recherche essentielle est donc de développer des architectures capables d'étendre ce champ récepteur pour capturer des dépendances à plus longue portée, sans pour autant sacrifier la résolution spatiale, par exemple via l'intégration de mécanismes d'attention. L'approche convolutive conserve un avantage fondamental : sa capacité à traiter des images de taille variable, y compris très grandes, sans augmenter le nombre de paramètres du modèle, garantissant ainsi un coût de calcul maîtrisé. Cette scalabilité marque une rupture avec les approches NMF classiques, dont la complexité dépend souvent directement de la taille des données d'entrée. Le traitement de l'image dans son entièreté permet un suivi visuel des cartes d'abondances durant l'entraînement, à la différence des approches par pixels où l'image complète n'est disponible qu'à la toute fin du processus, après que chaque pixel unique soit traité.

Enfin une observation clef dans cet état-de-l'art, notamment formulée par Squires (2019), est que la factorisation NMF, dans ces architectures, s'opère seulement au sein de l'espace latent et du décodeur. L'encodeur, quant à lui, n'est pas soumis aux mêmes contraintes de non-négativité et peut être conçu de manière aussi complexe que nécessaire. Il peut ainsi profiter de tous les bénéfices des architectures profondes modernes, telles que l'utilisation de couches de normalisation et de fonctions d'activation non linéaires. Ce mode de fonctionnement mène naturellement à des architectures d'auto-encodeurs asymétriques, où un encodeur profond et puissant extrait les caractéristiques, tandis qu'un décodeur plus simple et interprétable assure la reconstruction. Un exemple de ce type est schématisé sur la Figure 2.4.

2.3 Description du modèle hybride proposé

L'architecture du modèle NMF de type autoencodeur proposé se base donc sur une structure encodeur-décodeur asymétrique dotée d'une couche d'attention conçue pour estimer des

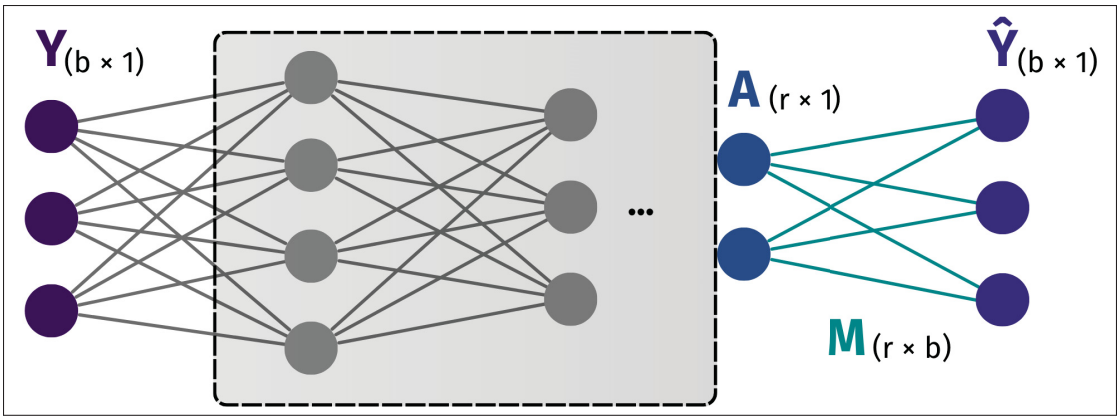

Figure 2.4 Schéma d'un autoencodeur asymétrique : le décodeur est contraint à une opération NMF, laissant l'encodeur (boîte noire) non contraint et modulable

caractéristiques non-locales durant l'entraînement. L'encodeur extrait une représentation riche en caractéristiques du cube MS d'entrée, $\mathbf{Y} \in \mathbb{R}^{h \times w \times b}$, où $h \times w$ est la taille spatiale et b est le nombre de bandes spectrales. Il génère les cartes d'abondance $\mathbf{A} \in \mathbb{R}^{h \times w \times r}$, où chaque bande représente les proportions relatives des signatures spectrales présents dans chaque pixel. Le décodeur reconstruit ensuite l'image d'entrée à partir de ces abondances, en apprenant les signatures spectrales $\mathbf{M} \in \mathbb{R}^{r \times b}$ comme les poids d'une convolution. Un aperçu de cette architecture hybride légère est illustré dans la Figure 2.5, traitant l'exemple d'un document MS.

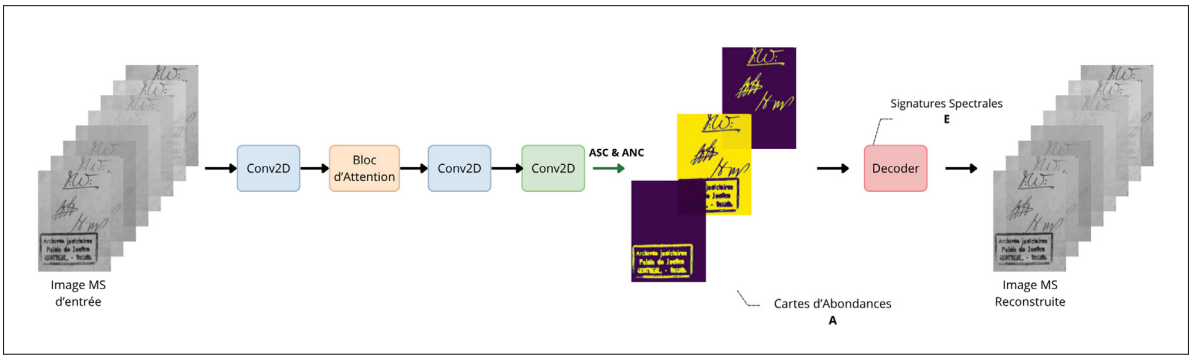

Figure 2.5 Schéma de l'architecture hybride. En bleu des convolutions dites «depthwise», en vert des convolutions dites «pointwise»

2.3.1 Encodeur

L'encodeur se compose de quatre couches séquentielles conçues pour traiter l'ensemble des bandes MS simultanément. Contrairement aux approches existantes Palsson *et al.* (2021) qui adoptent une stratégie basée sur des patchs, notre encodeur opère sur l'image entière, extrayant des relations utiles entre l'ensemble des pixels. Il commence par une couche de convolution 2D employant des noyaux de tailles 3×3. Les cartes de caractéristiques résultantes sont acheminées vers un bloc d'attention, qui affine la représentation des caractéristiques en mettant l'accent sur les régions les plus informatives (voir Section 2.3.1.1). L'encodeur se termine par une seconde couche 2D, similaire à la première, possédant $c = 4 \times b$ noyaux, suivie d'une convolution ponctuelle (*i.e.*, convolution avec un noyau 1×1 agissant dans la profondeur), créant un motif d'expansion-compression qui capture les dépendances spatiales et inter-canaux. La convolution ponctuelle permet de réduire la dimensionnalité en terme de profondeur (*i.e.*, la dimension spectrale), afin de produire les r cartes d'abondances.

2.3.1.1 Bloc d'Attention Non-local

Le cœur du bloc d'attention proposé est adapté du bloc Large Kernel Attention (LKA) (Guo, Lu, Liu, Cheng & Hu, 2023), qui est conçu pour améliorer la représentation des caractéristiques en capturant les relations locales et globales entre les pixels des images MS. Le module commence par traiter la carte de caractéristiques d'entrée via une couche de convolution 2D standard «DW-Conv» avec un noyau de taille 5×5 pour extraire les caractéristiques spatiales locales. Celui-ci est suivi d'une convolution dilatée «DW-D-Conv» avec un noyau de 7×7 et un stride de 3 , qui étend le champ récepteur sans augmenter le nombre de paramètres (*voir* Eq. 1.8), capturant un contexte plus large en traitant chaque canal indépendamment. Une convolution ponctuelle combine ensuite les caractéristiques de tous les canaux, affinant la représentation spectrale. La carte de caractéristiques résultante est multipliée élément par élément avec la carte de caractéristiques d'entrée originale, agissant comme un mécanisme d'attention, mettant en évidence les régions informatives tout en supprimant les caractéristiques moins pertinentes. Enfin, une convolution ponctuelle, avec le même nombre de canaux de sortie que d'entrée, affine

davantage les caractéristiques pour produire la sortie finale. Ce bloc d'attention est illustré sur la Figure 2.6.

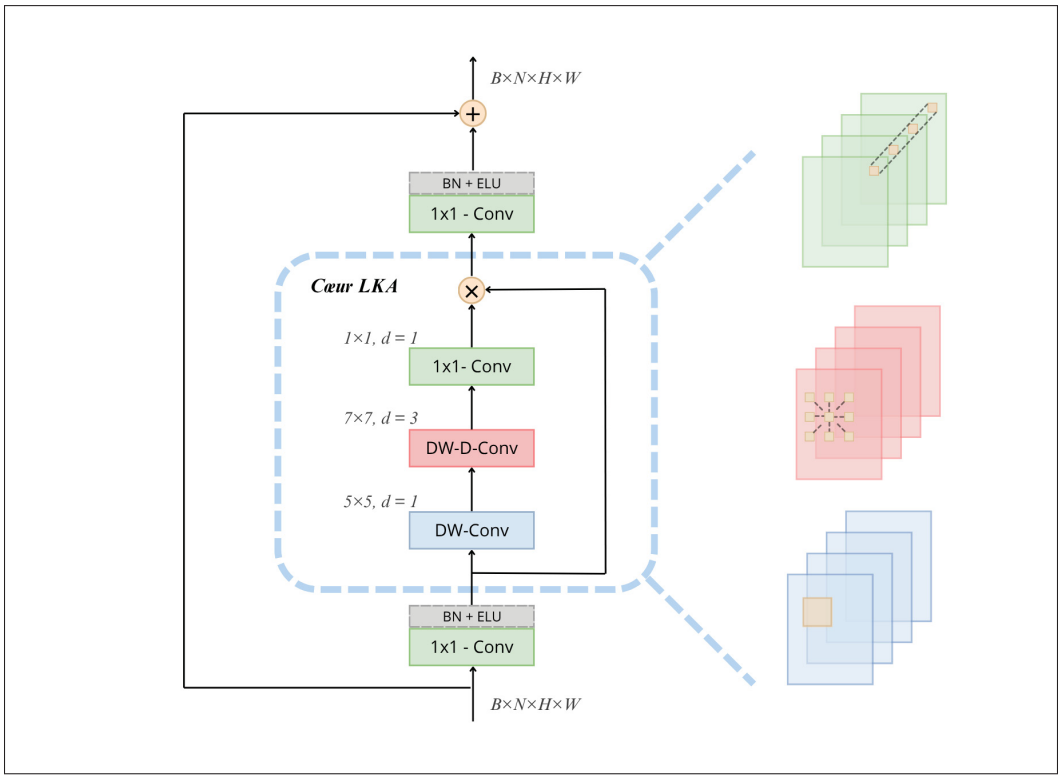

Figure 2.6 Visualisation du bloc d'attention LKA

2.3.2 Decodeur

2.3.2.1 Signatures spectrales comme poids de convolution contrainte

Contrairement aux modèles utilisant des structures encodeur-décodeur symétriques comme Alayrac, Carreira & Zisserman (2019), notre architecture intègre un décodeur léger et non négatif qui préserve l'interprétabilité de ses composantes apprises. Ce décodeur reconstruit les b bandes spectrales d'entrée à partir des r cartes d'abondance $\mathbf{A}$ en utilisant une convolution ponctuelle avec des poids contraint non négatifs et non biaisée. Ces poids non négatifs de taille $r \times b$ constituent la matrice des signatures spectral apprise $\mathbf{M}$.

2.3.2.2 La factorisation tripartite (NMF à 3 facteurs) : flexibilité et potentiel

A la manière de Ding *et al.* (2008) pour la NMF ou de Su *et al.* (2023) pour le démélange HS, notre architecture intègre une matrice d'interaction **S**. Le rôle de cette matrice est d'émuler une tri-factorisation NMF comme $\mathbf{Y} \approx \mathbf{MSA}$. Pour cela, les r cartes d'abondance produites par l'encodeur sont d'abord modulées par cette **S**, qui est une matrice non négative de taille $r \times r$ initialisée comme la matrice identité (*i.e.*, I_r). L'introduction de cette matrice poursuit un double objectif :

1. **Modéliser la variabilité spectrale :** Elle permet de capturer des interactions complexes et non linéaires entre les différentes composantes. Cela est crucial pour adresser des phénomènes comme la variabilité de la signature spectrale d'un même matériau en fonction de son état (*p. ex.*, son niveau d'humidité ou son état de dégradation).
2. **Relâcher les contraintes d'optimisation :** La matrice **S** découple l'échelle des signatures spectrales **M** de celle des cartes d'abondance **A**. Ce découplage est essentiel dans des cas où l'intensité physique d'un signal ne correspond pas à sa proportion relative. Par exemple, dans un document, le texte a souvent une réflectance bien plus faible que celle du papier. Sans la matrice **S**, la contrainte de somme à l'unité forcerait la carte d'abondance du texte à être très activée. En pondérant cette abondance, **S** permet au décodeur d'estimer correctement la signature spectrale du texte par rétropropagation, sans que les poids de cette signature ne soient artificiellement tirés vers zéro.

La conception asymétrique optimisée du décodeur réduit considérablement les exigences de calcul et le nombre total de paramètres ; par exemple, les paramètres totaux du décodeur totalisent seulement $r(r + b)$ paramètres pour une image quelque soit sa résolution spatiale. Enfin, les deux matrices **S** et **M** sont toutes deux contraintes à être non négatives et sont mises à jour par rétropropagation.

2.3.3 Ajout de contraintes (non-négativité, somme à l'unité, orthogonalité)

Pour assurer le respect des contraintes physiques des abondances, plusieurs approches peuvent être envisagées au niveau du décodeur, telles que l'utilisation de fonctions d'activation comme ReLU, la valeur absolue, ou encore la fonction Softmax (Palsson *et al.*, 2021). Parmi celles-ci, la fonction Softmax se distingue comme la solution la plus adaptée. En effet, elle est différentiable et garantit de manière élégante le respect à la fois de la contrainte de non-négativité (ANC) ainsi que celle de la somme à l'unité (ASC). Nous utilisons une fonction Softmax dotée d'un paramètre de température τ, ce qui offre une flexibilité supplémentaire pour contrôler la distribution des coefficients d'abondance. L'équation s'écrit comme suit pour chaque pixel i et chaque abondance k parmi r cartes, où a_i est le logit d'entrée :

$$\text{Softmax}(a_i) = \frac{\exp(a_i/\tau)}{\sum_{k=1}^{r} \exp(a_k/\tau)} \tag{2.1}$$

L'ajustement de l'hyperparamètre τ permet de moduler la dispersion des cartes d'abondance. Une température élevée conduit à des abondances plus douces et réparties, adaptées au démélange de spectres où les matériaux sont intrinsèquement mélangés. À l'inverse, une température faible pousse les probabilités vers des distributions quasi-binaires, ce qui est particulièrement pertinent pour des tâches comme la segmentation de documents textuels qui requièrent des pixels purs.

Enfin, pour améliorer la séparabilité des matériaux extraits, une contrainte d'orthogonalité est imposée sur la matrice des abondances $\mathbf{A}$. Cette contrainte a démontré son efficacité pour accentuer la distinction entre les différentes composantes, notamment dans le contexte de l'analyse de documents (Rahiche *et al.*, 2019; Rahiche & Cheriet, 2022). La régularisation par l'orthogonalité est formulée comme suit :

$$\mathcal{L}_{orth} = \|\mathbf{A}\mathbf{A}^T - \mathbf{I}_r\|_1 \tag{2.2}$$

où $\|\cdot\|_1$ désigne la norme L_1, $\mathbf{I}_r$ est la matrice identité de taille $r \times r$, et r est le nombre de cartes d'abondances r. Cette pénalité minimise la corrélation linéaire entre les cartes d'abondances estimées, décourageant ainsi la représentation d'un même matériau sur plusieurs composantes.

2.3.4 Fonction de coût globale

Contrairement à la factorisation en matrices non-négatives (NMF) classique, où l'optimisation est généralement contrainte à une mesure de reconstruction spécifique, notre approche basée sur un réseau de neurones offre une plus grande flexibilité. La fonction de coût totale, que le modèle cherche à minimiser, est une combinaison linéaire de trois termes principaux. Chacun de ces termes vise un objectif spécifique, permettant d'équilibrer la qualité de la reconstruction spatiale, la fidélité spectrale et la décorrélation des composantes extraites. La fonction de coût globale $\mathcal{L}_{Total}$ est définie comme suit :

$$\mathcal{L}_{Total} = \mathcal{L}_{SAD} + \lambda_1 \mathcal{L}_{MSE} + \lambda_{orth} \mathcal{L}_{orth}, \tag{2.3}$$

où λ_1 et λ_{orth} sont des hyperparamètres de pénalisation ajustés empiriquement (voir la section 2.4.4 pour les valeurs détaillées).

Le premier terme, l'erreur quadratique moyenne (*Mean Square Error*, MSE), évalue la qualité de la reconstruction pixel par pixel. Il mesure la différence entre l'image multispectrale d'entrée $\mathbf{Y}$ et l'image reconstruite $\hat{\mathbf{Y}}$:

$$\mathcal{L}_{MSE} = \frac{1}{hwb} \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{l=1}^{b} (\mathbf{Y}_{ijl} - \hat{\mathbf{Y}}_{ijl})^2, \tag{2.4}$$

où h, w, b sont les dimensions (hauteur, largeur, nombre de bandes) de l'image.

Pour compléter la MSE, nous utilisons la distance angulaire spectrale (*Spectral Angle Distance*, SAD), qui est équivalente à la similarité cosinus. L'avantage principal de cette métrique est son insensibilité aux variations d'illumination, car elle se concentre sur la forme de la signature spectrale plutôt que sur son intensité absolue. Suivant l'approche de Palsson *et al.* (2022), ce

terme mesure la similarité spectrale entre le vecteur spectral d'un pixel d'entrée et celui du pixel de sortie correspondant :

$$\mathcal{L}_{SAD}(\mathbf{Y}_{ij.}, \hat{\mathbf{Y}}_{ij.}) = \text{acos}\left(\frac{\langle \mathbf{Y}_{ij.}, \hat{\mathbf{Y}}_{ij.} \rangle}{\|\mathbf{Y}_{ij.}\| \|\hat{\mathbf{Y}}_{ij.}\|}\right), \tag{2.5}$$

où $\mathbf{Y}_{ij.}$ et $\hat{\mathbf{Y}}_{ij.}$ sont respectivement les vecteurs spectraux d'entrée et de sortie pour le pixel à la position (i, j).

Enfin, pour garantir l'indépendance des cartes d'abondance extraites, nous intégrons le terme de régularisation par l'orthogonalité $\mathcal{L}_{orth}$, tel que défini dans l'équation (2.2). Comme dit précédemment, cette contrainte pénalise la corrélation linéaire entre les différentes cartes d'abondance, favorisant ainsi une décomposition où chaque composante est distincte et représente un élément unique.

Cette fonction de coût multi-termes permet ainsi une optimisation équilibrée qui prend en compte différent facteurs :

- $\mathcal{L}_{MSE}$ force le modèle à reconstruire l'intensité exacte de chaque pixel.
- $\mathcal{L}_{SAD}$ force le modèle à préserver la fidélité spectrale (la nature du matériau).
- $\mathcal{L}_{orth}$ pousse le modèle à chercher une indépendance spatiale des différentes composantes.

2.3.5 Stratégie d'apprentissage et d'optimisation

Une force majeure de l'architecture proposée, par rapport aux approches algorithmiques traditionnelles, est sa capacité à tirer parti d'optimiseurs avancés issus de l'apprentissage profond. En conséquence, l'optimisation du modèle est réalisée à l'aide de l'optimiseur Adam (Kingma & Ba, 2017), une méthode de descente de gradient stochastique particulièrement efficace pour l'entraînement des réseaux de neurones. Adam se distingue en calculant des taux d'apprentissage adaptatifs pour chaque paramètre, en se basant sur des estimations des moments du premier et du second ordre des gradients. Cette approche flexible est bien adaptée à notre fonction de coût composite, permettant un ajustement robuste et efficace des poids du réseau.

La procédure d'entraînement est, elle, régulée par un mécanisme d'arrêt anticipé (*early stopping*) afin de prévenir le surapprentissage et de conserver le modèle le plus performant. Plus spécifiquement, l'entraînement est configuré pour un maximum de nombre d'époques. L'arrêt anticipé est déclenché si aucune amélioration de la fonction de coût n'est observée après une période d'époques consécutives, ou *patience*. Cela signifie que le modèle ne parvient plus à trouver une représentation des abondances plus orthogonale tout en conservant une bonne reconstruction du cube d'origine. Ce plateau dans l'optimisation est conceptuellement similaire aux critères de convergence qui gouvernent l'arrêt des algorithmes itératifs de la NMF classique, lorsque les mises à jour n'apportent plus de gain significatif ou qu'une tolérance seuil est atteinte.

2.4 Validation de l'architecture hybride

Afin de valider les bénéfices de l'architecture neuronale hybride proposée, une série d'ablations et d'expériences ont été menées pour la comparer à différentes implémentations de la NMF. Cette section s'articule autour de quatre axes principaux : le temps de calcul, la qualité de la décomposition, l'apport du contexte spatial et l'influence de l'orthogonalité et de la température.

2.4.1 Temps de calcul

Afin d'évaluer l'efficacité de notre approche, nous avons mesuré son temps d'exécution sur des cubes de données de taille croissante et l'avons comparé à plusieurs variantes de la NMF. Les méthodes de référence incluent : **EM-ONMF** (Pompili, Gillis, Absil & Glineur, 2014), **ONMF** (Yoo & Choi, 2010b), et **ONPMF** (Pompili *et al.*, 2014), trois NMF orthogonales conçues pour le clustering, **psNMF** (Hinrich & Mørup, 2018), un modèle de NMF probabiliste parcimonieux, **MA-ONMF** (Rahiche *et al.*, 2019), un modèle NMF imposant l'orthogonalité par optimisation riemannienne sur variété de Stiefel ; et **VBONMF** (Rahiche & Cheriet, 2022), une approche bayésienne avec contrainte d'orthogonalité, intégrant la détermination automatique de la pertinence (ARD). Enfin nous avons aussi comparé à une **NMF classique** par mises à jour multiplicatives, avec l'implémentation optimisée proposée par la librairie Scikit-learn (Pedregosa *et al.*, 2011), servant de référence pour une NMF sans contrainte.

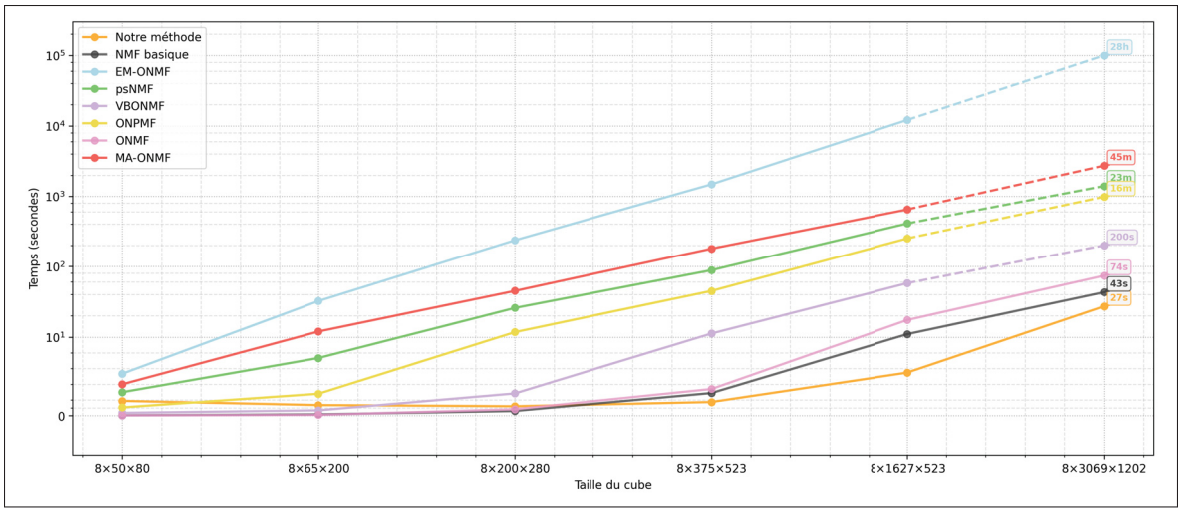

Figure 2.7 Comparaison de l'évolution du temps de calcul pour 100 itérations en fonction de la taille du cube de données pour plusieurs algorithmes de NMF de l'état de l'art et notre méthode. L'axe du temps suit une échelle logarithmique

La Figure 2.7 illustre les résultats de cette analyse comparative. Les méthodes ont été évaluées sur 100 itérations, et ont été lancées dix fois pour chaque volume afin de garantir la reproductibilité des résultats. Les résultats représentés en pointillés correspondent aux valeurs extrapolées, obtenues par interpolation log-log, afin d'estimer les gains de temps pour des volumes de données plus importants. On observe que les variantes de la NMF intégrant des contraintes, qu'elles soient orthogonales comme EM-ONMF, parcimonieuses comme psNMF ou bayésiennes comme VBONMF, présentent des temps de calcul significativement plus élevés que la NMF basique, avec une croissance souvent supérieure à linéaire. Pour le plus grand cube de données, de dimensions $8 \times 3069 \times 1202$, notre méthode ne requiert que 37 secondes. En comparaison, la NMF orthogonale EM-ONMF est la moins performante, requérant environ 99788s estimées, soit près de 28h. Ces résultats soulignent l'avantage computationnel significatif de la méthode proposée, la rendant particulièrement adaptée au traitement de données MS/HS volumineuses où la rapidité d'exécution est un critère essentiel.

Une régression par loi de puissance a permis d'estimer la complexité empirique de chaque méthode en fonction du volume de données. EM-ONMF présente une croissance clairement sur-linéaire, avec $O(n^{1.43})$. Les méthodes psNMF, VBONMF et MAONMF suivent une croissance

proche de la linéarité, avec des exposants respectifs de $(n^{0.92})$,$(n^{0.97})$ et $(n^{1.04})$. La NMF basique affiche aussi une croissance linéaire, $O(n^{1.00})$, mais reste néanmoins nettement plus rapide pour chaque volume étudié, avec un temps de calcule de 43s pour le plus grand cube.

En comparaison, notre méthode présente une croissance nettement plus lente, $O(n^{0.38}) < O(\sqrt{n})$. Cette complexité sous-linéaire s'explique par l'implémentation optimisée des CNN, qui exploite efficacement l'accélération GPU pour garder des temps de calculs faibles, tout en permettant l'intégration de contraintes similaires à celles des méthodes plus coûteuses. Le parallélisme du GPU est moins bien exploité sur les petites données, avec un coût légèrement plus élevé pour le plus petit cube, de l'ordre de 1.85s. Ces coûts fixes sont néanmoins largement amortis sur les gros volumes, où notre méthode devient alors particulièrement rapide et extensible. L'approche proposée constitue ainsi un compromis optimal entre expressivité, intégration de contraintes et performance de calcul, la rendant particulièrement adaptée à l'analyse de grands volumes de données des images MS (*p. ex.*, le cube moyen pour le jeu de donnée MStex de document historique est de l'ordre de $8 \times 1627 \times 523$).

2.4.2 Qualité de la décomposition

Afin d'évaluer la pertinence de notre approche, nous avons mené une étude qualitative sur la qualité de la décomposition obtenue. L'objectif est de vérifier visuellement la capacité des différentes méthodes de factorisation de matrices non-négatives (NMF) à séparer distinctement les différents matériaux constituant une image de document ancien. Une décomposition de haute qualité doit isoler chaque composant, tel que le texte, l'arrière-plan ou le support papier, dans une matrice d'abondance distincte, sans chevauchement ni résidu d'information provenant des autres composantes. Cette analyse visuelle permet de juger de l'efficacité des contraintes de régularisation, notamment l'orthogonalité, reconnue pour éliminer la redondance et produire une séparation physiquement interprétable.

La Figure 2.8 présente une comparaison visuelle des résultats de décomposition de différentes méthodes présentées précédemment. L'analyse met en évidence des différences significatives

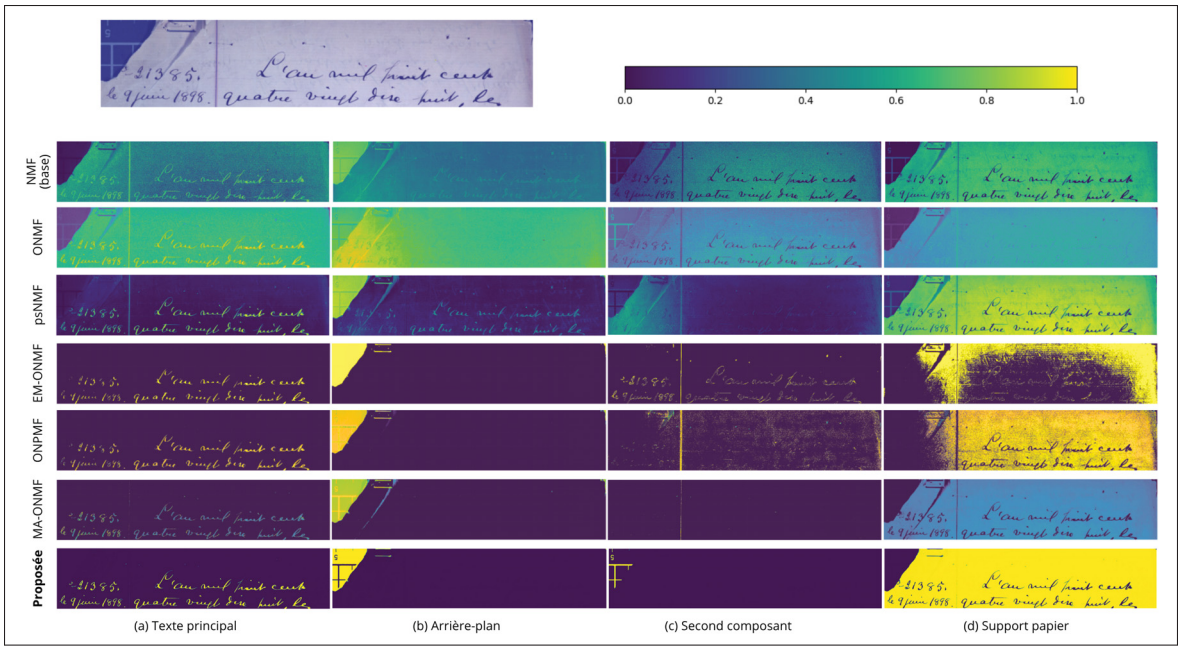

Figure 2.8 Comparaison qualitative de la décomposition obtenue par différentes méthodes NMF sur une image de document ancien. Chaque ligne représente une méthode et chaque colonne une composante extraite

dans la qualité de la séparation. La méthode NMF de base produit une décomposition plus simple où les composantes sont fortement mélangées. C'est aussi le cas pour la méthode probabiliste psNMF ainsi que pour la méthode ONMF. A l'exception de cette dernière, on peut observer que l'introduction de contraintes d'orthogonalité améliore sensiblement la discrimination entre les différents matériaux. Les trois méthodes, ONPMF, EM-ONMF et surtout MA-ONMF, permettent une interprétation directe des composantes extraites, malgré certaines régions incertaines.

La méthode proposée suit aussi cette observation, en fournissant une décomposition nette et interprétable. Elle parvient à isoler chaque matériau dans une composante distincte, où le texte est clairement extrait, le papier rendu homogène, et l'arrière plan décomposé en deux composants clairs et distincts. Cette performance est le fruit de la synergie entre la régularisation orthogonale et l'utilisation de la fonction softmax avec température, qui force le modèle à effectuer un «choix franc» lors de l'assignation des pixels. La décomposition résultante est physiquement interptétable. Cette méthode est notamment la seule à séparer clairement le texte du tapis de

découpe visible en arrière plan (*i.e.*, composants (b) et (c)). A l'inverse de toutes les méthodes NMF présentées, l'approche hybride est la seule à prendre en compte les relations spatiales entre les pixels afin de produire les cartes d'abondances. Ce contexte spatial permet d'apporter une cohérence locale entre les zones identifiée, son apport est donc détaillé dans la section suivante.

2.4.3 Apport du contexte spatial

Afin de quantifier rigoureusement l'impact du module d'attention à grand noyau (LKA) sur la portée spatiale de notre encodeur convolutionnel, un protocole expérimental a été mis en place pour comparer un modèle de référence à une variante intégrant un bloc LKA. Dans un premier temps, le champ récepteur théorique (TRF) peut être déterminé en appliquant l'équation (1.8). Pour le modèle de référence, la succession des couches convolutionnelles aboutit à un TRF de 5×5, tandis que pour le modèle expérimental, l'insertion du bloc LKA , qui contient notamment une convolution dilatée avec un noyau effectif de taille 19 (K=7,D=3), étend le TRF à 27×27.

Le champ récepteur effectif (ERF) est ensuite mesuré empiriquement en moyennant sur 150 échantillons la norme L1 du gradient du neurone de sortie central rétropropagé jusqu'à l'entrée. Pour garantir une initialisation fonctionnelle des poids, les deux architectures ont été pré-entraînées sur une tâche d'auto-reconstruction. Une cartographie comparative des topologies de l'ERF est alors obtenue pour les deux configurations : avec ou sans bloc d'attention LKA.

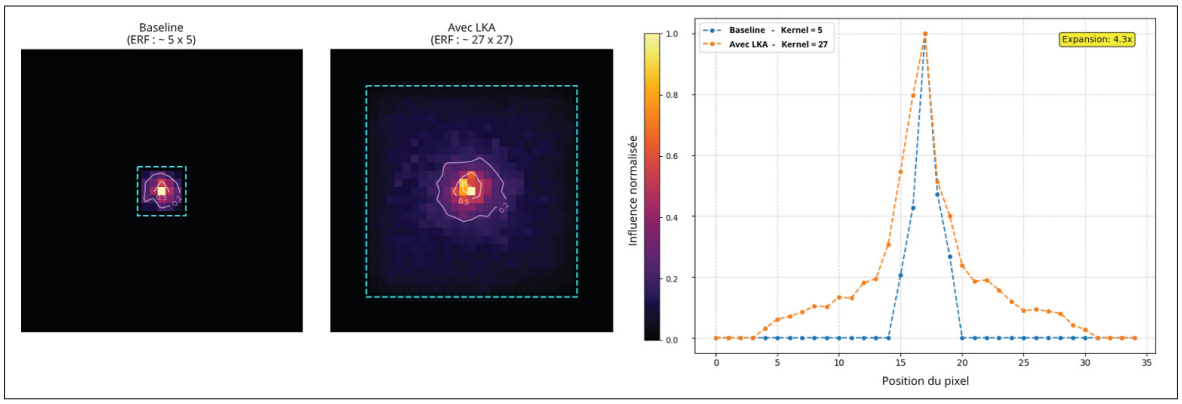

Figure 2.9 Visualisation comparative des champs récepteurs effectifs pour le modèle sans LKA (gauche) et avec LKA (centre). Le graphique (droite) illustre une coupe permettant du champ réceptif, permettant de mieux quantifier cette augmentation

La Figure 2.9 montre des différences très nettes entre les deux modèles. Le modèle de contrôle sans LKA a un champ récepteur effectif très concentré, ce qui correspond bien à son petit champ théorique de 5×5 et confirme qu'il est limité à un contexte local. À l'inverse, l'ajout du module LKA agrandit fortement l'ERF pour atteindre **27×27 pixels**, en accord avec le calcul théorique. Le graphique de coupe confirme cette observation : le modèle de base présente un pic très étroit, tandis que le modèle LKA affiche une courbe beaucoup plus étalée, montrant que les pixels lointains peuvent influer sur la sortie du modèle. Ces résultats expérimentaux illustrent l'expansion spatiale prédite par les calculs théoriques, agrandissant le champ récepteur d'un facteur 4.3 permettant d'agréger l'information contextuelle sur une plus grande échelle, et donc une meilleur compréhension globale de l'image. Cette analyse de l'ERF s'appuie sur un pré-entraînement sur du bruit aléatoire afin d'éviter tout biais. Elle n'évalue pas ce que le modèle apprend, mais illustre plutôt sa capacité à chercher de l'information spatiale.

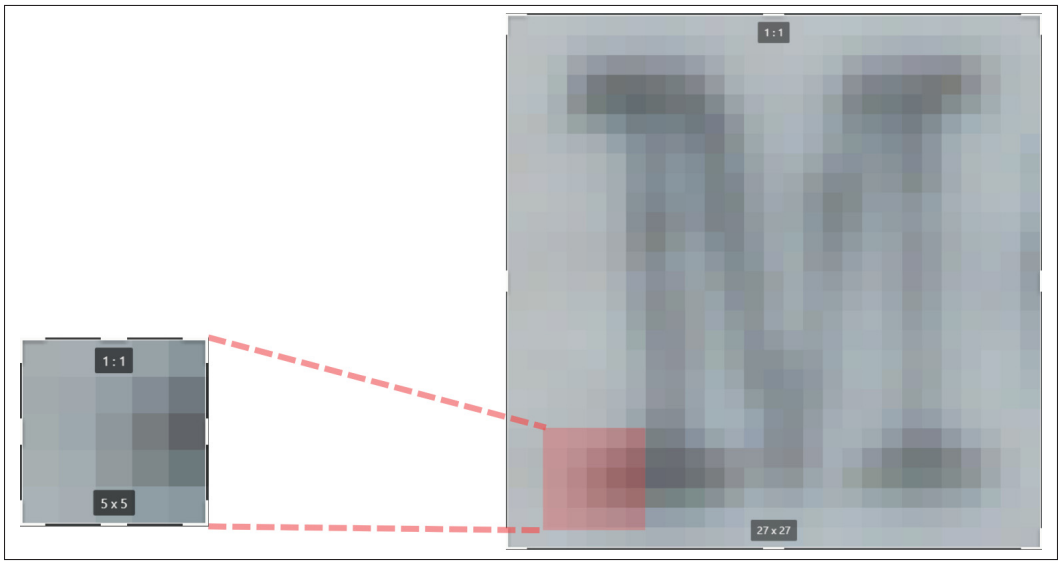

Figure 2.10 Illustration de l'importance du contexte sur une image du dataset Mstex : un patch de 5×5 (gauche) ne montre qu'une texture ambiguë, tandis que le patch de 27×27 (droite) révèle la structure du caractère textuel entier

Pour démontrer plus concrètement l'impact d'un champ réceptif plus étendu, la Figure 2.10 compare deux extraits d'une même image de la base de donnée Mstex. A gauche, la taille du patch correspond au champ récepteur du modèle de base, qui ne capture qu'un fragment de

texture non identifiable. L'information est trop locale pour être interprétée. A droite, la taille du patch, qui représente la portée du modèle avec LKA, révèle la lettre **M** dans son intégralité. Cette comparaison simple illustre l'avantage du module LKA. Ce dernier permet au réseau une fenêtre suffisamment grande pour accéder à un contexte sémantique pertinent permettant de reconnaître des caractères textuels. Le modèle est alors libre d'identifier les relations qui lui sembles pertinentes afin de segmenter l'image proposée.

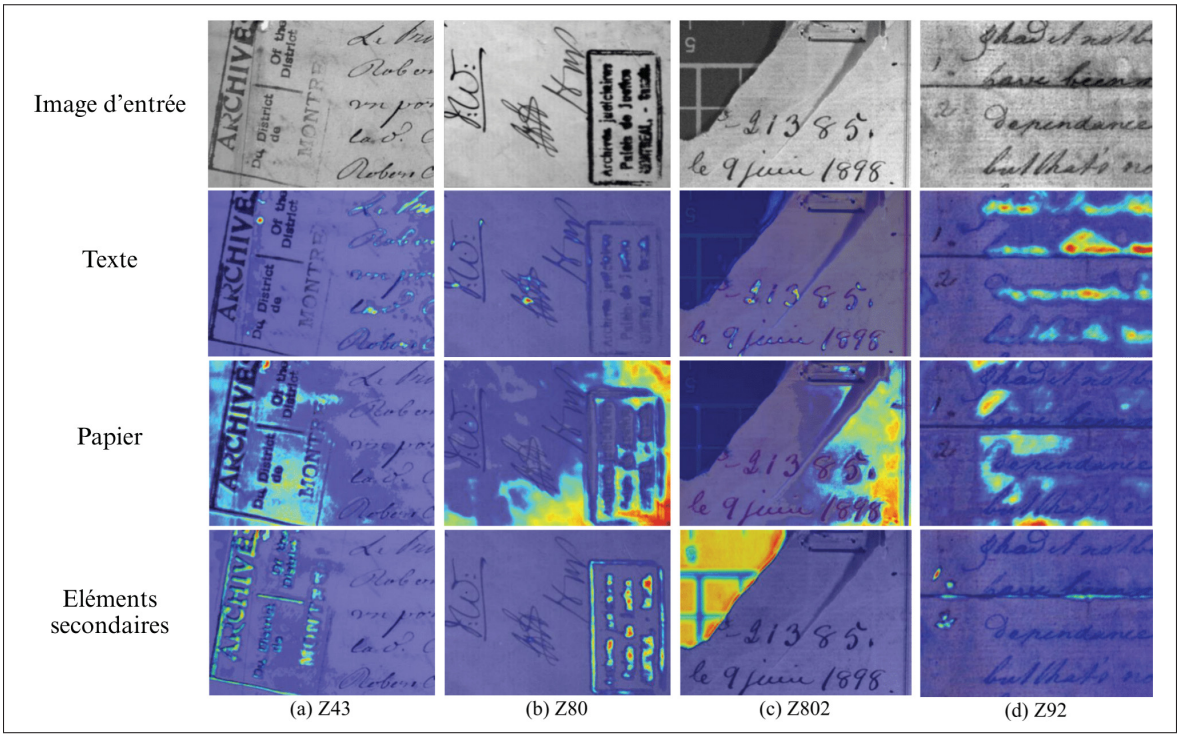

Figure 2.11 Visualisation de cartes d'attention en sortie du bloc d'attention LKA pour différentes images de documents. La première rangée montre les images d'entrée, tandis que les rangées suivantes illustrent certaines caractéristiques sur lesquelles le modèle porte son attention

La Figure 2.11 illustre concrètement cette capacité d'attention appliquée aux images de document. La première rangée expose les images d'entrée, tandis que les suivantes montrent plusieurs cartes d'attention mettant en valeur certaines caractéristiques clefs lors de l'apprentissage du modèle. On observe que pour la catégorie « Texte », l'attention se concentre précisément sur les caractères manuscrits. Pour la classe « Papier », le modèle isole le support du document, et

enfin, pour les « Éléments secondaires », il parvient à identifier des objets distincts comme les tampons d'archive, les encres secondaires ou encore l'arrière-plan.

Cette visualisation démontre que le modèle n'utilise pas seulement son large champ réceptif pour *voir* plus loin, mais qu'il apprend également à utiliser le contexte pour mieux discriminer les différents éléments spécifiques. Cette compétence est essentielle pour une segmentation sémantique précise.

2.4.4 Influence de l'orthogonalité et de la température

Les hyperparamètres de régularisation jouent un rôle déterminant dans la qualité de la décomposition spectrale obtenue par notre méthode. Deux paramètres clés contrôlent le comportement du modèle : la température softmax $\tau_{softmax}$, qui régit la netteté des cartes d'abondances, et le coefficient de régularisation orthogonale λ_{orth}, qui impose l'indépendance spatiale entre les composantes extraites. Pour analyser leur influence respective et leurs interactions, nous avons conduit une étude ablative sur le document manuscrit historique présenté précédemment.

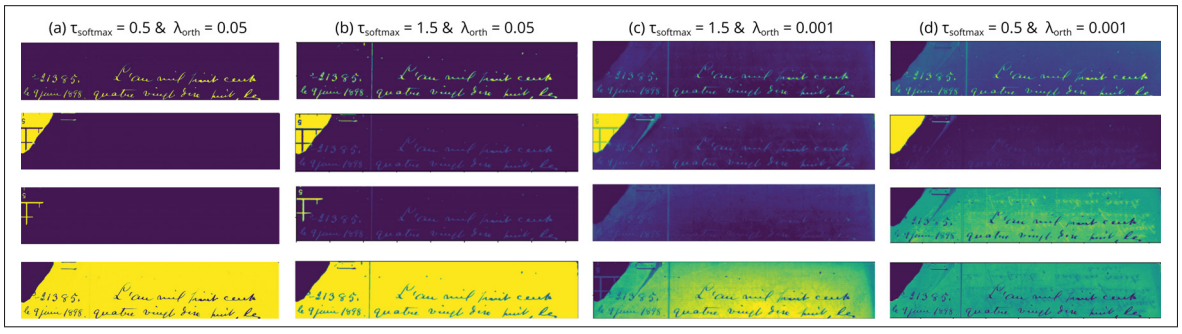

Figure 2.12 Influence des paramètres de température $\tau_{softmax}$ et d'orthogonalité λ_{orth} sur la décomposition. Une température élevée combinée à une faible régularisation orthogonale produit des cartes d'abondances à forte entropie, tandis qu'une température basse avec une forte pénalité orthogonale favorise une séparation nette des composantes

La Figure 2.12 présente les résultats de décomposition pour quatre configurations représentatives, obtenues en croisant deux valeurs de température ($\tau_{softmax} \in \{0.5, 1.5\}$) avec deux niveaux de pénalité orthogonale ($\lambda_{orth} \in \{0.001, 0.05\}$). Cette grille expérimentale permet d'observer le

comportement du modèle, avec des décompositions pouvant être plus diffuses ou extrêmement segmentées en fonction du choix des paramètres.

Chaque configuration produit quatre cartes d'abondances correspondant à des composantes physiques différentes. Bien que la décomposition (d) soit plus redondante, en ayant deux composants représentant le papier et ne séparant pas l'arrière plan en deux composantes, elle arrive à identifier le texte, visible par transparence, écrit à l'arrière du papier. Il est alors intéressant de noter que le choix de ces paramètres peut être adapté en fonction des besoins. Dans le cadre de notre étude, des composantes distinctes et directement interprétables sont préférées afin d'obtenir des segmentations nettes des images multispectrales. Les paramètres seront donc fixés suivant le scénario (a) avec $\tau_{softmax} = 0.5$ et $\lambda_{orth} = 5 \times 10^{-2}$.

2.5 Limites du modèle et problématique du rang

Malgré ses **avantages** significatifs en terme de **temps de calcul** et de **qualité de décomposition**, notre approche partage une **limite fondamentale** avec la NMF classique et les différentes méthodes d'apprentissage machine (voir Tab. 2.1) : la nécessité de **fixer manuellement le rang** r, soit le nombre de sources à extraire. Ce prérequis est le principal obstacle à une automatisation complète du modèle, et surtout, à son application à des données réelles où le nombre de matériaux est inconnu à priori. Le choix du rang est pourtant une tâche critique, une mauvaise estimation conduit inévitablement à une sous-représentation (rang trop faible) ou à une sur-segmentation des sources (rang trop élevé), compromettant la fiabilité de la décomposition. Pour adresser cette problématique et permettre une analyse de données entièrement autonome, le chapitre suivant introduit un **mécanisme d'estimation automatique du rang**.

CHAPITRE 3

MÉCANISME DE SÉLECTION AUTOMATIQUE DU RANG

La **détermination du rang**, soit le nombre optimal de composantes latentes dans un modèle, est une étape fondamentale et souvent complexe dans de nombreuses méthodes d'apprentissage automatique et de traitement du signal, tel que la NMF. Que ce soit pour déterminer le nombre de sujets dans un corpus de textes, le nombre de sources dans une scène observée, ou la complexité intrinsèque d'un réseau de neurones, le choix du rang r a un impact direct sur la capacité du modèle à généraliser, son **interprétabilité** et sa **performance**. Un rang sous-estimé peut conduire à un modèle trop simple, incapable de capturer la richesse des données (sous-apprentissage), tandis qu'un rang sur-estimé peut entraîner un sur-apprentissage, où le modèle s'ajuste au bruit et perd son pouvoir de généralisation, tout en augmentant inutilement le coût de calcul. L'importance d'une sélection de rang autonome, c.-à-d. d'une méthode capable de déduire le rang optimal directement à partir des données sans supervision manuelle, est donc évidente. Ce chapitre explore d'abord les approches classiques pour cette tâche avant de présenter une méthode dynamique combinant **l'élagage** et le principe de la **Longueur de Description Minimale**.

3.1 Revue des méthodes existantes pour la détermination rang, soit le nombre de composantes

La question de la sélection du modèle, et plus particulièrement du choix de sa dimensionnalité, est un problème classique toujours non résolu. Les approches pour y répondre se divisent principalement en deux grandes familles. La première, celle des méthodes de sélection *a posteriori*, requiert d'entraîner plusieurs modèles avec différents rangs pour ensuite choisir le meilleur en appliquant un critère externe. À l'opposé, la seconde famille regroupe les méthodes de sélection en ligne (ou online), qui visent à déterminer le rang optimal au sein d'un unique processus d'apprentissage, généralement en partant d'un modèle sur-paramétré dont la complexité est ajustée dynamiquement durant l'entraînement.

3.1.1 Méthodes de sélection du rang *a posteriori*

3.1.1.1 Critères statistiques basés sur la théorie de l'information

Ces méthodes formulent la sélection du rang comme un problème d'optimisation cherchant à équilibrer la fidélité du modèle aux données et sa complexité. L'idée centrale est de pénaliser les modèles plus complexes pour éviter le sur-apprentissage.

Le **Critère d'Information d'Akaike (AIC)** est l'un des plus connus (Akaike, 1974). Il est défini par :

$$AIC = 2k - 2\ln(\hat{L}) \tag{3.1}$$

où k est le nombre de paramètres du modèle (directement lié au rang r) et $\hat{L}$ est la vraisemblance maximale du modèle. Dans un cadre général, l'AIC est un outil très apprécié dont l'avantage majeur est sa capacité à sélectionner le modèle offrant la meilleure performance prédictive, une qualité qui devient optimale avec de grands volumes de données. Cependant, sa pénalité pour la complexité d'un modèle est relativement faible, ce qui crée une tendance naturelle à favoriser des modèles plus complexes. Ce défaut est particulièrement handicapant lorsqu'il existe plus d'un bon modèle candidat, l'AIC étant alors reconnu comme ni efficace ni convergent pour identifier la structure la plus simple (Zhang, Yang & Ding, 2023).

Pour remédier à cela, le Critère d'information d'Akaike corrigé (AICc) a été développé. L'AICc introduit un terme correctif d'ordre deux qui renforce la pénalisation de la complexité du modèle, en particulier lorsque le nombre d'observations est faible par rapport au nombre de paramètres. La formule de l'AICc est la suivante :

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \tag{3.2}$$

Ce terme correctif supplémentaire est d'autant plus important que n est petit. Lorsque la taille de l'échantillon n devient grande, le terme de correction tend vers zéro et l'AICc converge

vers l'AIC. Une règle empirique largement adoptée suggère d'utiliser l'AICc chaque fois que le rapport entre la taille de l'échantillon et le nombre de paramètres, n/k, est inférieur à 40.

Le **Critère d'Information Bayésien (BIC)**, ou critère de Schwarz, propose une pénalité plus forte pour la complexité du modèle, surtout pour les grands ensembles de données (Schwarz, 1978) :

$$BIC = k \ln(n) - 2 \ln(\hat{L}) \tag{3.3}$$

où n est le nombre d'échantillons de données. De manière générale, le critère BIC est plus parcimonieux que l'AIC car il pénalise plus sévèrement la complexité des modèles, ce qui limite la tendance au surajustement souvent associée à l'AIC.

Le principe de la **Longueur de Description Minimale (MDL)**, introduit par Rissanen (1978), est une autre approche puissante issue de la théorie de l'information. Il postule que le meilleur modèle est celui qui permet la compression la plus courte des données. La longueur de description totale est la somme de la longueur du code pour décrire le modèle lui-même, $L(H)$, et la longueur du code pour décrire les données étant donné le modèle, $L(D|H)$. Le BIC est souvent considéré comme une approximation du MDL. Des travaux récents continuent d'explorer des variantes du MDL, par exemple en utilisant la Vraisemblance Maximale Normalisée (NML) pour des factorisations non-négatives (Ito, Oeda & Yamanishi, 2016).

Cependant, les approches basées sur le MDL nécessitent généralement une sélection du rang *a posteriori*, après avoir exécuté plusieurs factorisations NMF avec différents rangs (Squires, Prügel-Bennett & Niranjan, 2017). Ce processus implique de comparer la longueur totale de description pour chaque rang candidat afin de choisir celui qui la minimise. Bien que rigoureux, il peut être coûteux en temps de calcul, car il ne permet pas une estimation directe du rang au cours de l'optimisation elle-même.

3.1.1.2 Méthodes basées sur la stabilité des solutions et l'erreur de reconstruction

Une approche plus heuristique mais très intuitive consiste à examiner la courbe de l'erreur de reconstruction en fonction du rang. En théorie, cette courbe devrait présenter un «coude» (ou "elbow" en anglais) au niveau du rang optimal, après lequel l'ajout de nouvelles composantes n'apporte qu'un gain marginal (voir Fig. 3.1). Cependant, le coude est souvent ambigu et difficile à identifier de manière automatique avec des données réelles, pouvant rendre cette méthode très subjective. L'erreur de reconstruction seule n'est alors pas suffisante pour juger de l'optimalité d'une décomposition, comme illustrée sur la Figure 3.1, où les rangs supérieurs continuent d'améliorer cette dernière malgré un rang optimal synthétique fixé à trois composants.

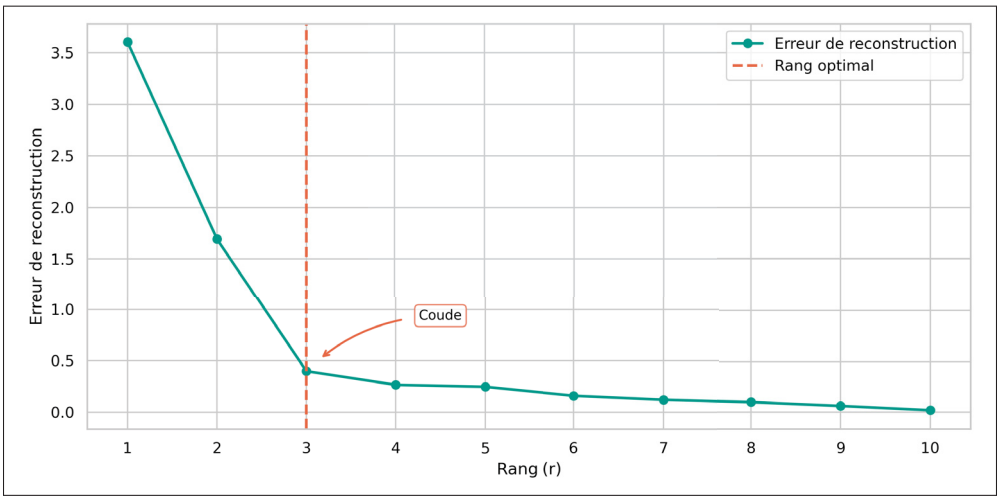

Figure 3.1 Méthode du coude pour la sélection du rang pour une NMF sur des données synthétiques. L'erreur de reconstruction diminue rapidement jusqu'au coude, puis se stabilise. Le rang optimal correspond au point d'inflexion où l'amélioration devient négligeable avec l'ajout de composantes, permettant d'équilibrer la reconstruction et la complexité du modèle

Pour surmonter cette subjectivité, des **méthodes basées sur la stabilité** ont été développées. L'idée fondamentale est qu'un modèle avec le rang correct devrait produire des solutions stables si les données d'entrée sont légèrement perturbées. L'approche de *Stability Selection* Meinshausen & Bühlmann (2010), formalise cette idée en entraînant le modèle sur de nombreux sous-échantillons des données et en ne retenant que les composantes (ou variables) qui

apparaissent de manière stable à travers les différentes exécutions. Cette approche suggère un rang en analysant la variabilité des solutions obtenues à partir de différentes initialisations aléatoires, postulant que le rang optimal correspond à une région de plus grande stabilité des solutions. En revanche, cette méthode nécessite le lancement de plusieurs itérations d'un même modèle pour différents rangs, demandant des coûts de calculs importants.

3.1.2 Méthodes de sélection du rang en ligne

3.1.2.1 Inférence bayésienne et Détermination Automatique de la Pertinence (ARD)

L'approche bayésienne formule la sélection du rang comme un problème d'inférence statistique, unifiant l'estimation des paramètres et le choix de la complexité du modèle. La **Détermination Automatique de la Pertinence (ARD)** est un cadre particulièrement puissant à cet égard, notamment mis en œuvre dans les modèles de NMF variationnels (*p. ex.*, VBONMF par Rahiche & Cheriet (2022)). Le principe fondamental de l'ARD consiste à placer un ***a priori* hiérarchique** sur les paramètres du modèle. Typiquement, chaque composante latente k (par exemple, la colonne $\mathbf{m_k}$ de la matrice de dictionnaire) est gouvernée par un hyperparamètre de précision λ_k (où $\lambda_k = 1/\sigma_k^2$) . On assigne un *a priori* gaussien de moyenne nulle aux poids de la composante, et un a priori Gamma Γ sur son paramètre de précision

$$p(\mathbf{m}_k|\lambda_k) = \mathcal{N}(\mathbf{m}_k|0, \lambda_k^{-1}\mathbf{I}) \tag{3.4}$$

$$p(\lambda_k) = \Gamma(\lambda_k|a_0, b_0) \tag{3.5}$$

Durant l'apprentissage par inférence variationnelle, le modèle optimise les distributions de ces hyperparamètres. Si une composante k est jugée superflue pour expliquer les données, la valeur attendue de sa précision $\mathbb{E}[\lambda_k]$ sera poussée vers l'infini. En conséquence, la distribution a posteriori de $\mathbf{m}_k$ se concentre massivement à zéro, ce qui équivaut à une suppression de la composante associée. Le rang effectif du modèle est alors simplement le nombre de composantes dont la pertinence reste finie.

Cette méthode, initialisée avec un nombre de composant initial r_{init} élevé, détermine ainsi le rang optimal de manière automatique. Cependant, sa mise en œuvre présente des défis, notamment pour la sensibilité du choix des hyperparamètres de l'*a priori* Γ (*i.e.*, a_0, b_0) et un coût de calcul qui, bien qu'amélioré par des algorithmes d'inférence rapides, peut rester élevé (voir Fig. 2.7).

3.1.2.2 Approches d'élagage dans les réseaux de neurones et leur pertinence

L'élagage (ou *pruning*) dans les réseaux de neurones profonds est un ensemble de techniques visant à réduire la taille du modèle en supprimant des poids, des neurones ou des filtres redondants, afin de diminuer le coût de calcul et d'améliorer la généralisation. Ces techniques sont particulièrement pertinentes pour notre discussion car elles offrent un mécanisme pour ajuster dynamiquement la complexité du modèle.

On distingue principalement deux types d'élagage (Cheng, Zhang & Shi, 2024) :

- **L'élagage non structuré**, qui supprime des poids individuels dans les matrices de poids. Cela crée des matrices creuses mais ne réduit pas directement la «largeur» ou la «profondeur».
- **L'élagage structuré**, qui supprime des groupes entiers de poids, comme des canaux de convolution, des couches des blocs de neurones.

A partir d'un réseau sur-paramétré, l'élagage réduit le nombre de paramètres progressivement, en se basant sur des critère de pertinence ou de redondance pour sélectionner les poids, ou filtres dans les CNNS, à supprimer. L'élagage est alors majoritairement appliqué après l'entraînement (*post-training*), mais plusieurs méthodes récentes s'intéressent a des application dites «*online*», c.-à-d. pendant l'entraînement (Elkerdawy, Elhoushi, Zhang & Ray, 2022). En revanche, bien que ces méthodes en ligne effectuent une réduction de la complexité du modèle, elles sont dépourvues d'un objectif explicite interprétable visant à déterminer un rang optimal. Le processus s'arrête généralement lorsqu'un taux de compression prédéfini est atteint (Park, Kim, Kim, Choi & Lee, 2023; Anagnostidis *et al.*, 2023). Même si certains critères de sélection ciblent la redondance pour ne garder que les éléments les plus importants (He, Wu, Liang & Lam, 2021), les méthodes d'élagage n'ont pas directement été conçu pour directement estimer un rang de décomposition.

Toutefois, un parallèle peut être fait entre l'élagage dans un réseau de neurone pour une réduction de dimensionnalité et la sélection d'un rang pour la NMF. Comme vu avec la similarité avec les auto-encodeurs présentée au Chapitre 2, la suppression d'un canal ou de neurones dans l'espace latent revient à réduire la dimensionnalité et donc à une modification du rang.

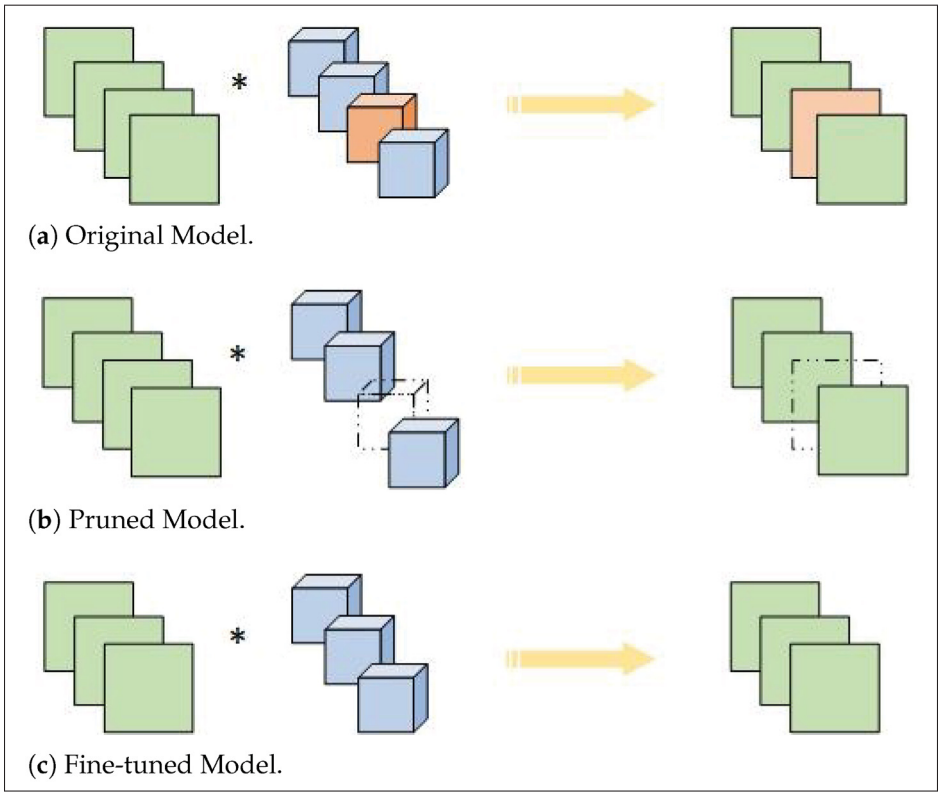

Figure 3.2 Exemple schématique du processus d'élagage (pruning) de filtres dans un modèle CNN, tiré de Shao *et al.* (2021). **(a)** Modèle initial avec tous les filtres actifs. **(b)** Modèle élagué où certains filtres redondants ou peu utiles sont supprimés. **(c)** Modèle réajusté (fine-tuning) pour compenser la perte induite par l'élagage. Appliqué aux filtres produisant les cartes d'abondances d'un modèle, l'élagage correspond à une réduction du rang effective

3.1.2.3 Réseaux de neurones avec sélection du nombre de composantes

Les architectures récentes basées sur des transformeurs visuels (ViT) ont profondément transformé la segmentation d'images, mais la question de la sélection automatique du rang reste largement ouverte. Des modèles comme Mask2Former (Cheng, Misra, Schwing, Kirillov & Girdhar,

2022), atteignent des performances de pointe pour la segmentation d'images, mais reposent sur un apprentissage supervisé et un nombre de classes fixé prédéfini. DINOv2 (Oquab *et al.*, 2024) propose un pré-entraînement non-supervisé à grande échelle de ViT, produisant des représentations visuelles robustes réutilisables pour diverses tâches. Cependant, l'application à la segmentation nécessite l'ajout de modules dédiées, eux-même entraînés sur des données annotées, empêchant une estimation directe du nombre de classes.

Kirillov *et al.* (2023) introduisent SAM, proposant une segmentation « universelle ». Un ViT génère des masques de segmentation en réponse à des indications visuelles (points, boîtes). Une segmentation automatique est alors proposée en générant des nombreux points quadrillant l'image, chacun produisant un masque. Ces masques candidats sont alors fusionnés selon des scores de qualité et via une suppression non maximale (NMS) basée sur leur superposition. Ainsi, bien que le nombre de composants varie automatiquement, cette sélection repose uniquement sur des critères locaux de redondance, sans prise en compte d'une structure sémantique globale. Pour des images de texte, SAM peut produire un masque par lettre. Chacun de ces masques est jugé valide individuellement, mais n'ayant aucun chevauchement, le modèle ne les regroupe pas en un seul composant qui pourrait correspondre à un mot ou une ligne de texte (voir Fig. 3.3).

(a) Image RGB (b) Segmentation Automatique (SAM)

Figure 3.3 Exemple de segmentation automatique d'une image de documents historique par SAM. Le ViT segmente chaque lettre de manière automatique, mais ne réussit pas à les identifier comme un même composant

Dans le domaine médical, MS-Former (Karimijafarbigloo, Azad, Kazerouni & Merhof, 2024) se distingue comme l'unique approche à combiner segmentation non supervisée et ajustement dynamique du rang. Il s'appuie sur une régularisation auto-supervisée (cohérence intra/inter-classe et entropie croisée) pour estimer le nombre de segments sans supervision humaine.

Toutefois, il reste spécialisé sur des images médicales et n'a pas été validé sur d'autres domaines comme les images de documents. L'utilisation des patchs pour la segmentation de documents manuscrits peut être sous-optimal, comme observé sur la Figure 1.5

3.2 Sélection dynamique du rang basée sur l'élagage et le principe MDL

En nous inspirant des approches d'élagage et d'identification du rang, nous proposons une méthode qui détermine le rang optimal non pas en testant une série de rangs discrets, mais en adaptant le modèle lui-même de manière dynamique.

3.2.1 Principe : De la surcomplétude à l'optimalité

Au lieu de chercher à deviner le bon rang *a priori*, notre stratégie part d'un état initial délibérément sur-complet, c'est-à-dire avec plus de cartes d'abondance (r_{init}) que d'objets potentiellement présents dans les données. Cette sur-complétude initiale permet au réseau de capturer un ensemble exhaustif de toutes les composantes possibles, comme cela a été fait dans des travaux antérieurs (Karimijafarbigloo *et al.*, 2024; Rahiche & Cheriet, 2022).

À partir de ces r_{init} cartes d'abondance de départ, une stratégie d'entraînement itérative est employée. Le but est de progressivement combiner les cartes d'abondances les plus redondantes, jusqu'à obtenir un rang optimal. A la différence de SAM, la combinaison des composantes se produit durant l'entraînement. En intégrant un principe d'élagage au modèle hybride présenté précédemment, il est alors possible de réduire le rang, tout en gardant des composantes interprétables ayant un sens sémantique.

Deux composantes majeures au fonctionnement de cet algorithme sont alors nécessaires : (1) Un **critère de similarité** permettant d'identifier les composantes à supprimer, et (2) un **critère de sélection** pour le rang optimal.

3.2.2 Développement d'un critère de similarité inter-composantes

Pour identifier la carte la moins informative, un score de similarité par paire est calculé entre toutes les cartes non élaguées afin d'identifier la paire la plus similaire. De cette paire, la carte ayant la plus petite norme de Frobenius (*i.e.*, la carte représentant la plus petite proportion de l'image) est ajoutée à l'ensemble des cartes élaguées. Pour chaque carte de cet ensemble, les poids de la convolution ponctuelle qui la produisent sont masqués. De même, les poids correspondants du décodeur, qui représentent la signature spectrale associée sont aussi désactivés. Cela permet de sélectionner la carte la moins informative. Le modèle est forcé de réorganiser ses poids pour combiner cette carte d'abondance avec les cartes restantes. Le modèle transfert les connaissances apprises vers les cartes d'abondances les plus proches ressemblantes, produisant une nouvelle décomposition.

Pour évaluer de manière exhaustive la relation entre nos cartes, nous avons développé un score de similarité qui intègre deux aspects cruciaux : la **similarité spatiale** et la **similarité spectrale**. Si une approche purement spatiale, telle que la suppression non-maximale (NMS), permet d'identifier des composants similaires dans les cartes d'abondance, cela n'est pas suffisant comme discuté à la section 3.1.2.3. C'est pourquoi une composante de similarité spectrale est ajoutée. Celle-ci est essentielle pour pénaliser les éléments similaires situés à des localisations spatiales différentes.

3.2.2.1 Prise en compte de la similarité spatiale

Notre mesure de similarité spatiale s'inspire de la corrélation croisée normalisée, qui évalue la ressemblance entre deux cartes d'abondance A_i et A_j. Sa formulation de base est la suivante :

$$\mathcal{S}_{\text{corr}}(i,j) = \frac{\langle A_i, A_j \rangle}{\|A_i\| \|A_j\|}, \tag{3.6}$$

où $\langle \cdot, \cdot \rangle$ est le produit scalaire de Frobenius et $\| \cdot \|$ est la norme de Frobenius. Cependant, en raison de la contrainte d'orthogonalité ajoutée, les cartes d'abondances générées ont peu de zones de chevauchement. Lors des premières itérations du modèle, le modèle est forcé de

fragmenter des mêmes éléments en plusieurs parties. Une simple corrélation ne peut alors pas identifier la ressemblance entre ces cartes d'abondances. Pour contrer cet effet de fragmentation d'objets, nous appliquons d'abord une opération de lissage spatial à l'aide d'un noyau gaussien, obtenant $\mathcal{A}_i$ et $\mathcal{A}_j$, avant de mesurer une possible corrélation.

Un autre manque de la corrélation simple peut être observé lorsqu'un objet similaire est divisé en deux parties égales. Ainsi, un petit objet à l'intérieur d'un plus grand était souvent plus pénalisé qu'un même objet divisé en deux parties égales et séparées (*p. ex.*, le texte et le papier sont plus corrélés que le même support papier divisé en deux zones). Dans ce cas, seulement une fine zone de contact est mesurée par la corrélation, résultant en l'identification de deux matériaux distincts. Pour mieux gérer cet effet, la différence absolue des activations totales entre les deux cartes est utilisée comme facteur pénalisant.

$$\frac{1}{|\langle \mathcal{A}_i - \mathcal{A}_j, \mathbf{1} \rangle| + \epsilon} \tag{3.7}$$

où $\langle \cdot, \cdot \rangle$ est le produit scalaire de Frobenius, $|\cdot|$ est la valeur absolue et ϵ est une constante évitant une division par zéro. Ce terme **amplifie** la similarité des paires de cartes ayant des distributions d'abondance très proches en termes de magnitudes et inversement. Deux éléments représentant une grande proportion de l'image, et ce même avec une faible corrélation spatiale, sont alors plus pénalisés que deux éléments corrélés mais ayant une grande différence d'activation.

3.2.2.2 Prise en compte de la similarité spectrale

En plus de la similarité spatiale, il est impératif de considérer la ressemblance spectrale des matériaux eux-mêmes, représentés par leurs signatures spectrales. Pour ce faire, nous utilisons l'angle spectral, une mesure qui évalue la similarité de forme entre deux signatures spectrales, indépendamment des variations d'illumination. Cette mesure standard, connue sous le nom de Distance Angulaire Spectrale (SAD), est une mesure de distance. Or, pour l'intégrer de manière cohérente à notre score de similarité spatiale, nous devons la convertir en une mesure

de similarité (où une valeur élevée indique une forte ressemblance). Nous utilisons donc l'angle complémentaire de la distance angulaire (SAD′) :

$$\mathrm{SAD}'(\mathbf{E}_i, \mathbf{E}_j) = \frac{\pi}{2} - \arccos\left(\frac{\langle \mathbf{E}_i, \mathbf{E}_j \rangle}{\|\mathbf{E}_i\| \|\mathbf{E}_j\|}\right) \tag{3.8}$$

où $\mathbf{E}_i$ et $\mathbf{E}_j$ sont les i-ème et j-ème colonnes de la matrice des signatures spectrales E. Cette transformation assure qu'une similarité spectrale élevée se traduit par une valeur de SAD′ élevée, rendant cette mesure directement additive à la composante de similarité spatiale.

La formule finale de similarité, capturant à la fois la similarité spatiale et spectrale entre les matériaux, est alors exprimée comme suit :

$$\mathcal{S}_{i,j} = \frac{1}{|\langle \mathcal{A}_i - \mathcal{A}_j, \mathbf{1} \rangle| + \epsilon} \frac{\langle \mathcal{A}_i, \mathcal{A}_j \rangle}{\|A_i\| \|A_j\|} + \lambda_{\mathrm{SAD}'} \mathcal{L}_{SAD'}(E_i, E_j), \tag{3.9}$$

où $\mathcal{A}_i$ et $\mathcal{A}_j$ sont les cartes d'abondance lissées, E_i et E_j sont les signatures correspondantes, $\langle \cdot, \cdot \rangle$ est le produit scalaire, et $\lambda_{SAD'}$ est un paramètre équilibrant la similarité spectrale et spatiale.

3.2.3 Guidage par le principe de la Longueur de Description Minimale (MDL)

Une fois le critère de similarité $S_{i,j}$ établi, combiné avec l'élagage, il permet une suppression itérative de la composante la plus redondante. La paire obtenant le score le plus élevé est identifiée comme la plus redondante. De cette paire, la composante ayant la plus faible norme, représentant donc la plus petite fraction de l'image, est désignée pour être élaguée. Cela permet d'adapter le rang de manière dynamique durant l'entraînement. Cependant, une question demeure : *comment identifier le rang de décomposition optimal ?*

Pour cela, notre approche s'appuie sur le principe de la **Longueur de Description Minimale (MDL)**, un cadre théorique robuste pour la sélection de modèles (Squires *et al.*, 2017; Rissanen, 1978). Le principe MDL postule que le meilleur modèle est celui qui permet la description la

plus concise des données. L'objectif est donc de minimiser la somme de deux termes :

$$\min_{H} \{L(H) + L(D \mid H)\}, \tag{3.10}$$

où $L(D \mid H)$ est la qualité de description des données D étant donné le modèle H, et $L(H)$ représente longueur de description du modèle lui-même (*i.e.*, sa complexité).

Dans le contexte de notre modèle de type hybride proposé, cette formulation trouve une analogie directe :

- **La fidélité des données** $L(D \mid H)$ correspond à notre fonction objective utilisée, $\mathcal{L}_{SAD} + \lambda_1 \mathcal{L}_{MSE}$, qui mesure l'erreur de reconstruction à partir du modèle H.
- **La complexité du modèle** $L(H)$ est capturée par le coût de description des facteurs {A,S,E}. Ce dernier est influencé à la fois par la structure et la taille des facteurs. Cela est représenté par le rang, qui diminue à chaque étape d'élagage et réduit la taille du modèle, ainsi que par la contrainte d'orthogonalité, $\lambda_2 \mathcal{L}_{orth}$ qui réduit la complexité structurelle.

La fonction de coût globale $\mathcal{L}_{\text{MDL}}$, que notre algorithme cherche à minimiser, est donc définie comme suit :

$$\mathcal{L}_{\text{MDL}} = \underbrace{(\mathcal{L}_{\text{SAD}} + \lambda_1 \mathcal{L}_{\text{MSE}})}_{\text{Fidélité aux données } L(D|H)} + \underbrace{(\lambda_{\text{orth}} \mathcal{L}_{\text{orth}} + \lambda_r r)}_{\text{Complexité du modèle } L(H)} \tag{3.11}$$

En suivant ce principe, il est possible de calculer le coût MDL total à chaque itération du modèle. Au début de l'entraînement, le rang est haut, les composantes ne sont pas orthogonales et le modèle peine à reconstruire les données. Le coût MDL associé est donc naturellement plus élevé. Au fil de l'entraînement, le modèle apprend à mieux représenter les données, réduisant donc $L(D \mid H)$. La contrainte d'orthogonalité ainsi que l'élagage progressif réduisent alors la complexité du modèle $L(H)$, faisant ainsi chuter le coût total MDL. Ce coût atteint alors un minimum lorsque le modèle atteint un équilibre optimal entre simplicité et fidélité. Lorsque l'élagage se poursuit au-delà de ce point, le modèle devient trop simple pour représenter les données, et l'erreur de reconstruction $L(D \mid H)$ augmente. Cela entraîne une remontée du coût MDL total. Ce minimum empirique observé sur la courbe du coût MDL sert alors de critère de

sélection objectif. Le rang correspondant à ce point est considéré comme le rang optimal, car il représente le compromis idéal entre la compression du modèle et la préservation de l'information contenue dans les données.

3.2.4 Algorithme d'élagage progressif des composantes redondantes

L'algorithme alterne entre deux phases : une phase d'entraînement du modèle jusqu'à un critère d'arrêt précoce, et une phase d'élagage qui supprime la carte jugée la moins informative.

Algorithme 3.1 : Algorithme de la méthode d'élagage progressif proposée (PRISM)

Input : Image MS à décomposer $\mathbf{Y}$;
Rang initial r_{init} et nombre de composantes minimum r_{min} ;
Hyperparamètres $\lambda_1, \lambda_{\text{orth}}, \lambda_r, \lambda_{SAD'}$ et kernel de lissage K ;
Output : Cartes d'abondances optimales $\mathbf{A}_{\text{opt}}$, Signatures spectrales optimales $\mathbf{E}_{\text{opt}}$, Rang optimal r

Initialisation du modèle PRISM avec r_{init} composantes (toutes actives);
active_map_indices $\leftarrow \{1, 2, \ldots, r_{init}\}$;
best_overall_mdl_cost $\leftarrow \infty$, best_model_config $\leftarrow$ null, et best_rank $\leftarrow r_{init}$;
for *k from* r_{init} *down to* r_{min} **do**
    Entraînement du modèle PRISM jusqu'au critère d'early stopping;
    Soit trained_model$_k$ le modèle après convergence;
    $\mathbf{A}_k, \mathbf{E}_k, \mathbf{S}_k, \hat{\mathbf{Y}} \leftarrow \text{trained\_model}_k(\mathbf{Y})$;
    $L_{\text{recon}} \leftarrow \mathcal{L}_{SAD}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_1 \mathcal{L}_{MSE}(\mathbf{Y}, \hat{\mathbf{Y}})$ `// Erreur de reconstruction` $L(D|H)$
    $L_{\text{struct}} \leftarrow \lambda_2 \mathcal{L}_{orth}(\mathbf{A}_k)$ `// Complexité structurelle de` $L(H)$
    $L_{\text{rank\_penalty}} \leftarrow \lambda_3 \cdot k$ `// Terme de pénalité de rang de` $L(H)$
    current_mdl_cost $\leftarrow L_{\text{recon}} + L_{\text{struct}} + L_{\text{rank\_penalty}}$;
    if *current_mdl_cost* $<$ *best_overall_mdl_cost* **then**
        best_overall_mdl_cost $\leftarrow$ current_mdl_cost;
        best_model_config $\leftarrow$ trained_model$_k$;
        best_rank $\leftarrow k$;
    if $k > r_{min}$ **then**
        Soit $\mathbf{A}_{active}$ le set des cartes d'abondances correspondants aux active_map_indices dans $\mathbf{A}_k$;
        Soit $\mathbf{E}_{active}$ le set des signatures spectrales correspondants aux active_map_indices dans $\mathbf{E}_k$;
        for *each map* A_m *in* $\mathbf{A}_{active}$ **do**
            $\mathcal{A}_m \leftarrow K * A_m$; `// Convolution spatiale (lissage)`
        for *each pair of distinct maps* $(\mathcal{A}_i, \mathcal{A}_j)$ *(and corresponding* E_i, E_j *from* $\mathbf{E}_{active}$*)* **do**
            Calcul de la similarité S_{ij} à partir de $\mathcal{A}_i, \mathcal{A}_j, E_i, E_j$ (en utilisant l'Eq. 3.9);
        $(i^*, j^*) \leftarrow$ indices de la paire la plus similaire;
        map_to_prune $\leftarrow \arg\min_{m \in \{i^*, j^*\}} \|A_m\|_F$;
        Supprimer map_to_prune du set active_map_indices;
        Élagage des poids produisant les cartes inactives dans la convolution pré-abondances;
        Élagage des signatures spectrales associées;
Charger le modèle PRISM avec la configuration et les paramètres de best_model_config;
$\mathbf{A}_{\text{opt}}, \mathbf{E}_{\text{opt}} \leftarrow \text{Optimal\_Model}(\mathbf{Y})$;
$r \leftarrow$ best_rank;
return $\mathbf{A}_{opt}, \mathbf{E}_{opt}, r$

À chaque étape d'élagage, le modèle est contraint de s'adapter en fusionnant, déplaçant ou supprimant des éléments, distillant ainsi efficacement ses connaissances dans une représentation plus compacte. Cette stratégie offre l'avantage significatif de fournir une interprétabilité sur la focalisation du modèle et de permettre un contrôle sur son processus de décision.

Le coût MDL permet alors de trouver un équilibre entre la qualité de la reconstruction et la complexité du modèle. Cet équilibre représente le rang optimal estimé de la décomposition.

3.2.5 Intégration dans l'architecture hybride proposée

L'architecture complète du modèle, appelé PRISM, est présentée à la Figure 3.4. L'architecture intègre tous les composants décrits précédemment. Cette architecture hybride de NMF profonde établit un pont entre l'apprentissage profond non supervisé et les contraintes physiques des modèles NMF. Elle permet à la fois la performance computationnelle, l'interprétabilité physique des résultats et une sélection adaptative du rang.

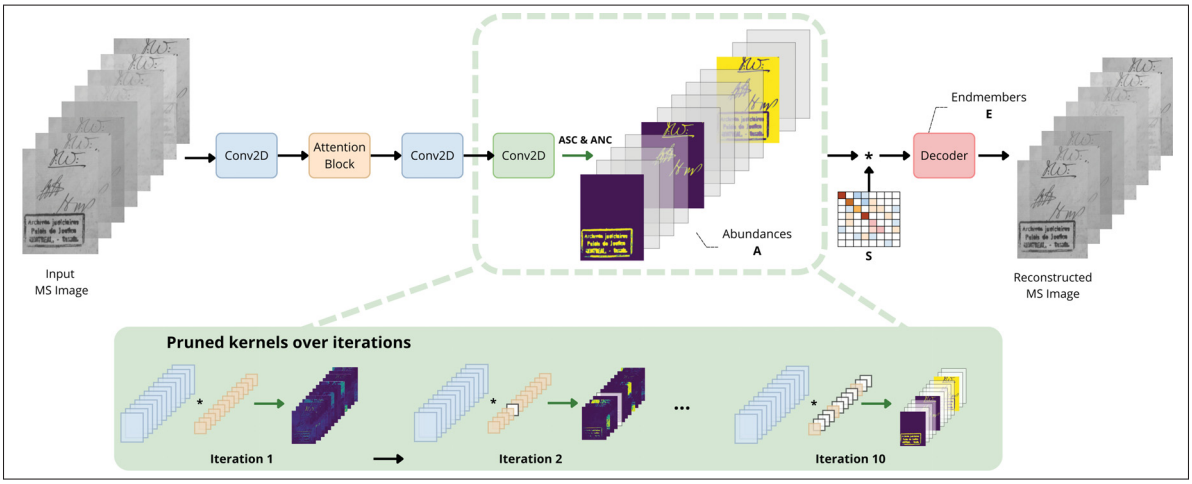

Figure 3.4 Schéma de l'architecture hybride avec sélection adaptative du rang. En vert est illustré l'élagage progressif des composantes. Au fil des itérations, les poids de la convolution pré-abondance sont élagués, résultant en la suppression des carte d'abondance redondantes associées

L'implémentation du modèle à l'aide de la librairie PyTorch permet la création de cette architecture, tout en permettant une optimisation des différents modules par rétro-propagation.

Afin d'obtenir un élagage non-structuré des poids, c'est-à-dire sans utiliser des ratios aléatoires de suppressions, le module `torch.nn.utils.prune.custom_from_mask` est utilisé. Les masques de pruning sont spécifiquement conçus pour supprimer les poids neuronaux responsables de la génération des cartes d'abondances sélectionnées lors de la procédure d'élagage progressif. L'implémentation conjointe de contraintes de paramétrisation et d'élagage personnalisé par masque pour des couches de convolution étant non-standard, nous renvoyons le lecteur intéressé par les détails techniques vers les tutoriels et la documentation officielle de Pytorch [3] [4] [5].

L'utilisation du principe MDL pour la sélection automatique du rang constitue une contribution majeure de notre approche. La Figure 3.5 illustre l'évolution du coût MDL en fonction du rang pour une image du jeu de données MStex (voir section 4.1). A chaque itération, la carte

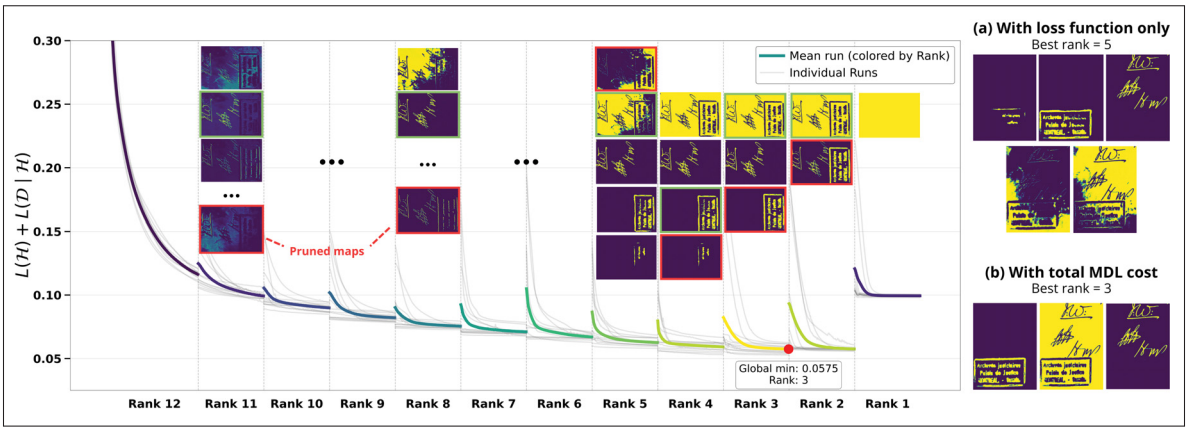

Figure 3.5 Visualisation de l'élagage progressif sur une image de document. Après chaque itération, une carte est élaguée et le modèle s'entraîne à adapter la décomposition

d'abondance la moins pertinente est supprimée par élagage des poids correspondants. Afin de respecter la contrainte ASC, le modèle est contraint de réorganiser ses connexions durant l'entraînement, ce qui se manifeste par une augmentation transitoire du coût MDL suivie d'une

[3] Paramétrisation des modules PyTorch : https://docs.pytorch.org/tutorials/intermediate/parametrizations.html

[4] Tutoriel d'élagage et re-paramétrisation : https://docs.pytorch.org/tutorials/intermediate/pruning_tutorial.html

[5] Documentation `custom_from_mask` : https://docs.pytorch.org/docs/stable/generated/torch.nn.utils.prune.custom_from_mask.html

convergence vers une valeur inférieure. Au-delà d'un certain rang, la capacité du modèle devient insuffisante pour maintenir simultanément une reconstruction fidèle des données et la distinction entre les composantes. Cette dégradation se traduit par une augmentation du coût MDL, signalant que le compromis entre complexité du modèle et qualité de reconstruction n'est plus optimal.

3.3 Validation de la sélection dynamique du rang

Cette section présente une analyse approfondie des différentes composantes de notre approche de sélection dynamique du rang. Nous évaluons l'impact de chaque module sur les performances du système à travers une série d'études d'ablation et d'analyses paramétriques.

3.3.1 Ablation des composantes du critère de similarité

Afin d'évaluer la contribution de chaque terme dans notre mesure de similarité, nous conduisons une étude d'ablation systématique sur des paires de cartes d'abondances extraites de données réelles. La Figure 3.6 présente trois scénarios représentatifs : (a) cartes d'abondances représentant du texte avec l'effet de fragmentation, (b) cartes d'abondances présentant une composante de papier fragmentée en deux parties égales, et (c) cartes d'abondances de texte et de papier.

Pour chaque scénario, nous évaluons l'impact progressif de l'ajout des différentes composantes du critère de similarité :

- **Corrélation seule** (ligne 1) : Utilisation du coefficient simple de corrélation entre les cartes d'abondances
- **Corrélation et différence d'activation** (ligne 2) : Ajout du terme mesurant les différences d'intensité d'activation
- **Configuration complète** (ligne 3) : Intégration du lissage spatial des cartes d'abondances via le noyau de convolution K

Les résultats quantitatifs démontrent l'importance cruciale de chaque composante. Avec la corrélation seule, les trois scénarios obtiennent des scores de similarité très faibles (0.04%, 4.5% et 0.05% respectivement), ne permettant pas de distinguer les paires similaires des dissimilaires.

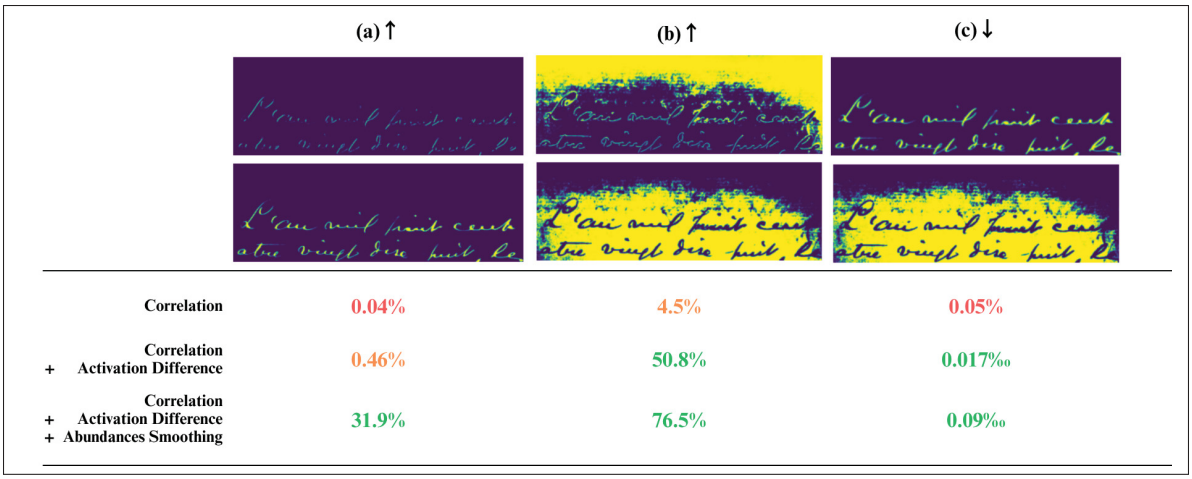

Figure 3.6 Étude d'ablation des composantes du critère de similarité sur trois paires de cartes d'abondances représentatives. (a) Paire de cartes hautement similaires issues du même matériau avec variations effet de fragmentation. (b) Paire de cartes hautement similaires mais partageant seulement une fine région de contact. (c) Paire de cartes dissimilaires correspondant à des matériaux distincts. Les pourcentages indiquent le score de similarité calculé pour chaque configuration

L'ajout du terme de différence d'activation améliore significativement la discrimination. Pour les scénarios (a) et (b), les scores augmentent à 0.46% et 50.8%, tandis que pour le scénario (c), représentant deux matériaux distincts, le score diminue de 0.05% à 0.017‰, confirmant sa dissimilarité. En revanche, le scénario (a) reste tout de même sous-évalué avec seulement 0.46%, bien que représentant le même texte fragmenté.

L'intégration finale du lissage spatial complète efficacement le critère. Les scénarios (a) et (b) atteignent des scores de 31.9% et 76.5% respectivement, permettant leur identification correcte comme paires similaires. Le scénario (c) maintient un score négligeable (0.09‰), confirmant la robustesse du critère pour identifier les véritables dissimilarités.

Au-delà de ces composantes spatiales, le critère intègre également une composante de similarité spectrale. Cette composante est essentielle pour l'identification et le regroupement de cartes d'abondances correspondant au même matériau mais n'ayant aucun chevauchement spatial, un phénomène observé précédemment pour la méthode SAM avec suppression NMS (voir Fig. 3.3).

3.3.2 Apport du coût MDL pour la sélection du rang

Comme illustré dans la Figure 3.7, l'utilisation de la fonction de perte seule, sans régularisation MDL, conduit à un modèle sur-complet. Dans ce cas, les meilleurs résultats se concentrent autour de rangs sous-optimaux de 4, 5 et 6. Qualitativement, ces solutions de rang élevé produisent des composantes avec des matériaux textuels fragmentés, indiquant une décomposition moins interprétable du point de vue physique.

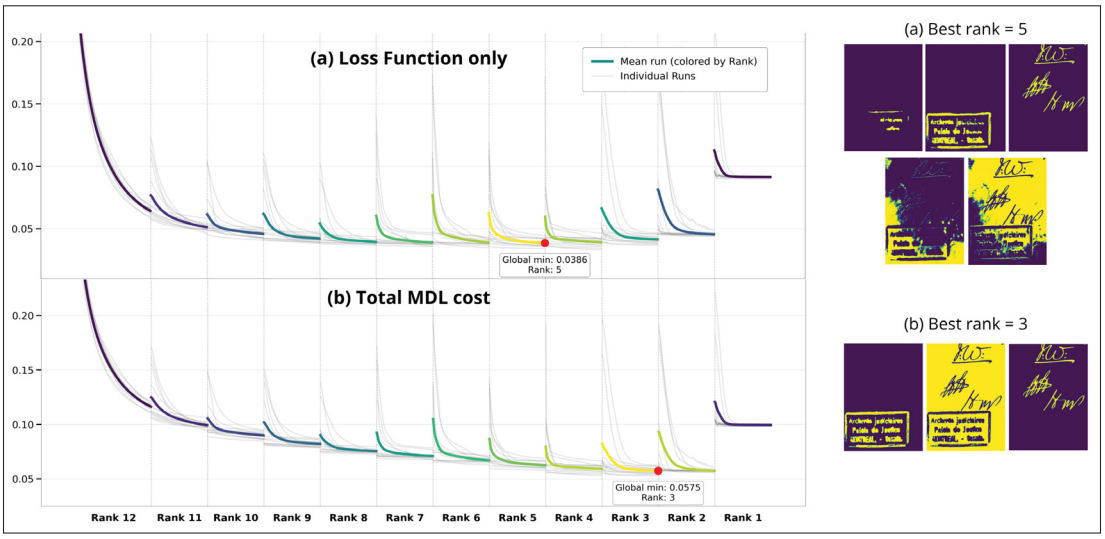

Figure 3.7 Ablation sur la sélection du rang avec et sans le coût MDL. La courbe du haut représente l'évolution de la fonction de perte seule, montrant une préférence pour des rangs plus élevés. La courbe du bas intègre le coût MDL complet, révélant un minimum global au rang 3 et une sélection plus claire

L'incorporation du coût MDL total améliore non seulement les propriétés de convergence, mais guide également le modèle vers une solution plus robuste. Le coût MDL identifie correctement une plage stable et acceptable de solutions autour des rangs 2, 3 et 4, avec un minimum global clair au rang 3. Cette observation confirme que le coût MDL est crucial pour prévenir le sur-apprentissage du modèle et pour déterminer correctement le nombre réel de composantes latentes dans les données. Sans le terme $L(D \mid H)$, le modèle tend à favoriser des décompositions complexes qui maximisent la fidélité de reconstruction au détriment de l'interprétabilité.

3.3.3 Influence du rang initial et sélection du rang minimal

Contrairement aux approches traditionnelles où le rang sélectionné influence directement la décomposition finale, le choix du rang initial r_{init} dans notre méthode n'a qu'un impact limité sur la solution finale. Nos expériences révèlent que ce paramètre influence principalement le temps de calcul total, avec une relation linéaire entre le nombre d'itérations d'élagage $(r_{init} - r_{final})$ et le temps d'exécution. La qualité de la décomposition finale reste généralement invariante, l'algorithme combinant de manière hiérarchique les différentes composantes selon leur pertinence. Dans les rares cas où une composante importante serait supprimée prématurément, l'adoption d'une stratégie de sélection basée sur plusieurs exécutions (typiquement 10) permet de garantir la convergence vers le rang optimal, indépendamment de l'initialisation.

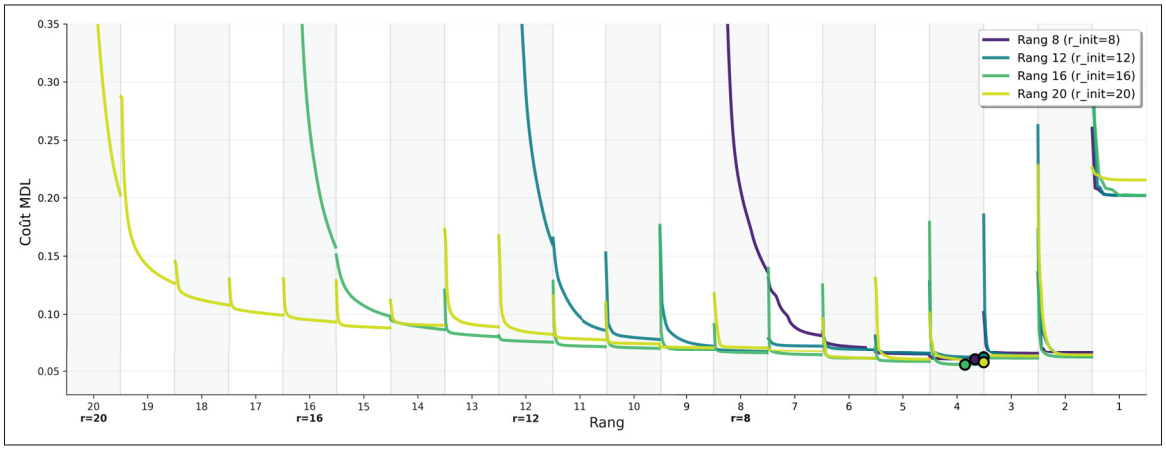

Figure 3.8 Évolution du coût MDL en fonction du rang pour différentes initialisations. Les courbes correspondent à des rangs initiaux de 8, 12, 16 et 20, montrant la convergence vers un minimum commun aux rang 4 pour une même image

La Figure 3.8 illustre cette robustesse en présentant les trajectoires du coût MDL pour différentes valeurs de $r_{init} \in \{8, 12, 16, 20\}$. Toutes les courbes convergent vers le même minimum global, confirmant que le processus d'élagage guidé par le critère MDL identifie de manière fiable les composantes les moins pertinentes, indépendamment du point de départ. La seule contrainte pratique est que r_{init} doit être supérieur ou égal au nombre réel de composantes dans les données. Dans nos expériences, nous fixons $r_{init} = 12$, correspondant au nombre de bandes spectrales du jeu de données MSBin, assurant ainsi une initialisation suffisamment riche pour tous les scénarios

testés, tout en gardant un temps d'exécution plus bas. Comme le montre la Figure 3.9, ce choix représente un bon compromis entre capacité de représentation et efficacité computationnelle. Le temps d'exécution croît linéairement avec le rang initial, passant d'environ 1 minutes pour $r_{init} = 8$ à plus de 3 minutes pour $r_{init} = 20$.

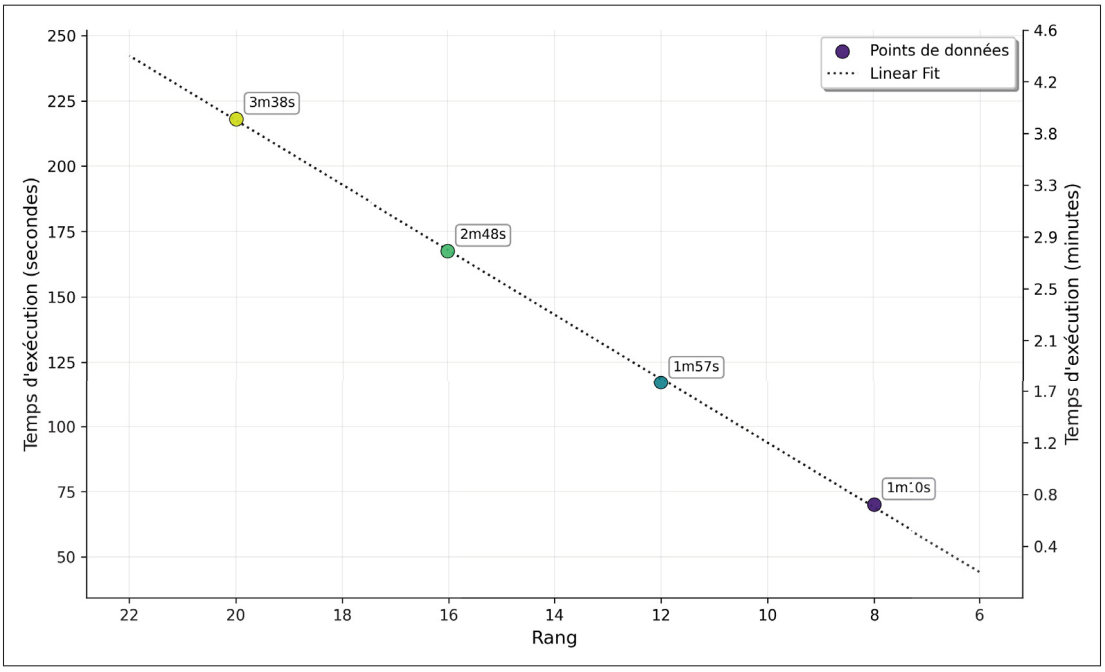

Figure 3.9 Temps d'exécution de l'algorithme en fonction du rang initial. Les points montrent les temps mesurés pour les rangs 8, 12, 16 et 20, avec une relation linéaire claire (ligne pointillée)

Le rang minimal r_{min} est lui fixé à 2, représentant la configuration minimale physiquement significative pour le démélange spectral : une composante pour le matériau d'intérêt (typiquement le texte) et une pour l'arrière-plan. Une valeur de 1 conduirait à une solution triviale (*i.e.*, une image uniforme) violant les hypothèses fondamentales d'un mélange linéaire. Cette configuration minimale permet à l'algorithme d'explorer des solutions avec peu de composantes, tout en maintenant l'interprétabilité physique des résultats. Cela permet d'optimiser l'efficacité computationnelle en supprimant la dernière itération superflue.

3.4 Discussion des avantages et des défis de la sélection dynamique du rang

La **sélection dynamique du rang** présente plusieurs avantages significatifs par rapport aux approches traditionnelles à rang fixe. Premièrement, elle élimine le besoin de validation croisée exhaustive pour déterminer le nombre optimal de composantes, **réduisant** ainsi considérablement **le coût computationnel**. Deuxièmement, l'intégration du critère MDL fournit une justification théorique solide pour le compromis entre complexité du modèle et qualité de reconstruction.

Cependant, plusieurs défis demeurent. Les différents hyperparamètres du critère MDL nécessitent un réglage, bien que nos expériences, présentées au chapitre suivant, suggèrent que des valeurs par défaut robustes puissent être établies. Une limitation importante de notre approche réside dans la conception du critère de similarité, initialement optimisé pour des données textuelles où la redondance spatiale est clairement définie. Pour des données hyperspectrales de nature différente, ce critère pourrait potentiellement éliminer des composantes spectralement proches mais physiquement distinctes et importantes pour l'interprétation.

Cette observation suggère que l'approche est optimalement déployée dans un contexte semi-supervisé, où l'expert peut valider les décisions d'élagage. En effet, l**'analyse des décompositions** sur-complètes **avant élagage** offre un éclairage pertinent sur la structure des données, révélant **comment le modèle** identifie et **sépare** initialement les différentes régions de l'image. Cette capacité d'exploration s'avère particulièrement pertinente pour des scènes complexes (voir Section 4.3.3), où la visualisation de l'évolution du rang permet de comprendre les relations hiérarchiques entre composantes. Le chapitre suivant présente une évaluation exhaustive de notre approche sur divers jeux de données, démontrant sa capacité de généralisation tout en illustrant ces considérations pratiques pour différents types de scènes multispectrales et hyperspectrales.

CHAPITRE 4

EXPÉRIMENTATIONS ET GÉNÉRALISATION DES APPLICATIONS DU MODÈLE

Ce chapitre évalue la capacité de généralisation de l'approche proposée, dénommée **PRISM**, à travers trois objectifs expérimentaux : (1) valider la robustesse sur des **images multibandes de documents** avec diverses configurations spectrales, (2) positionner PRISM par rapport aux auto-encodeurs de **démélange HS** sur des données de télédétection, et (3) s'attaquer à **l'extension au cas sous-déterminé**, avec des images RVB de documents et des représentations profondes d'images naturelles traitées comme données HS.

4.1 Résultats sur la base d'image MS de documents MStex

Le premier jeu de données que nous utilisons dans le cadre de cette étude est une combinaison des collections appelées MStex 1 et 2 [6]. Ces collections, fournies par Bibliothèque et Archives nationales du Québec (BAnQ) et digitalisés par Hedjam & Cheriet (2013), regroupent deux ensembles distincts de dix cubes MS de manuscrits historiques couvrant une période s'étendant du XVIIe au XXe siècle. Chaque cube est composé de 8 bandes spectrales, capturées dans une plage allant de 340 nm à 1100 nm, comme présentées dans le Tab. 4.1. Les acquisitions ont été réalisées à l'aide d'une caméra Chroma KAF 6303E (Kodak) offrant une résolution de 6 mégapixels. Les données acquises présentent alors une large résolution spatiale de 3072×2048 pixels par bande spectrale, chaque pixel couvrant une zone de 9×9 µm. Chaque document a ensuite été segmenté en zones d'intérêt, générant des images de résolutions spatiales variables.

Tableau 4.1 Détail des bandes spectrales utilisées pour la collection MStex

Bande	**F1s**	**F2s**	**F3s**	**F4s**	**F5s**	**F6s**	**F7s**	**F8s**
Longueur d'onde (nm)	340	500	600	700	800	900	1000	1100
Filtre de lumière	UV	Bleu	Vert	Rouge	IR 1	IR 2	IR 3	IR 4

Il est à noter que l'image z58 du premier jeu de données n'a pas été retenue dans notre analyse en raison de dégradations importantes observées sur l'une de ses bandes spectrales.

[6] Ces collections ont été introduites dans le cadre du concours MSTex (Hedjam, Nafchi, Moghaddam, Kalacska & Cheriet, 2015). Données disponibles en ligne : https://tc11.cvc.uab.es/datasets/SMADI_1

La vérité terrain (ground-truth) associée à ces images classifie alors les pixels en deux catégories : texte ou non-texte. La qualité de décomposition peut alors seulement être évaluée par la qualité du texte extrait, en partant du principe que si la décomposition est bonne, alors le texte extrait devrait l'être aussi. En effet, une décomposition optimale des composantes spectrales doit conduire à une séparation efficace entre les différents matériaux constitutifs du document (encre, papier, agents de dégradation, etc.). Si cette séparation est réalisée de manière satisfaisante, l'extraction et la lisibilité du texte s'en trouvent nécessairement améliorées. Par conséquent, les métriques de qualité textuelle agissent comme un indicateur direct et fiable de la performance globale du processus de décomposition. Cette approche présente l'avantage de fournir une évaluation objective et quantifiable, où une amélioration des scores de qualité textuelle traduit directement une meilleure séparation des composantes spectrales.

4.1.1 Métriques d'évaluation

Pour quantifier cette qualité d'extraction du texte et ainsi mesurer l'efficacité de différentes approches, quatre métriques complémentaires issues de la littérature en binarisation de document manuscrits ont été adoptées. Ces métriques ont été sélectionnées pour leur capacité à capturer différentes caractéristiques complémentaires de la qualité de binarisation.

La première métrique évalue directement l'exactitude de la classification binaire. La **F-Mesure (FM)**, largement reconnue dans la communauté scientifique, constitue notre métrique principale de performance. Elle s'exprime comme la moyenne harmonique entre la précision et le rappel :

$$\text{FM} = \frac{2 \times \text{Rappel} \times \text{Précision}}{\text{Rappel} + \text{Précision}}, \tag{4.1}$$

où le Rappel = $\frac{VP}{VP+FN}$ quantifie la capacité du modèle à identifier l'ensemble des pixels textuels, tandis que la Précision = $\frac{VP}{VP+FP}$ mesure la justesse de ces identifications, avec VP, FP et FN représentant respectivement les vrais positifs, faux positifs et faux négatifs. Cette métrique, exprimée en pourcentage, offre un équilibre entre sur-détection et sous-détection du texte.

La seconde métrique s'attache à caractériser les distorsions introduites lors du processus de binarisation. La **Métrique de Taux Négatif (NRM)** quantifie spécifiquement les erreurs de classification au niveau des pixels :

$$NRM = \frac{TR_{FN} + TR_{FP}}{2}, \tag{4.2}$$

où $TR_{FN} = \frac{FN}{FN+VP}$ et $TR_{FP} = \frac{FP}{FP+VN}$ représentent respectivement les taux de faux négatifs et de faux positifs. Cette métrique, variant de 0 à 1, est particulièrement sensible aux erreurs qui peuvent affecter la lisibilité du texte extrait.

De manière complémentaire, la **Distorsion de Distance Réciproque (DRD)** évalue les distorsions perceptuelles en considérant un voisinage spatial de chaque pixel :

$$DRD = \frac{\sum_{k=1}^{N} DRD_k}{NUBN}, \tag{4.3}$$

où DRD_k représente la distorsion pondérée dans un voisinage 5×5 centré sur le k-ième pixel erroné, et NUBN dénombre les blocs 8×8 non-uniformes dans l'image de référence. Cette métrique, également bornée entre 0 et 1, mesure efficacement l'impact visuel des erreurs de binarisation sur la lecture, les erreurs isolées étant moins pénalisées que les clusters d'erreurs.

Enfin, le **Rapport Pic de Signal sur Bruit (PSNR)** offre une mesure globale de fidélité entre l'image binarisée et la vérité terrain :

$$PSNR = 10 \log_{10} \left(\frac{I_{max}^2}{EQM} \right), \tag{4.4}$$

où I_{max} représente l'intensité maximale possible et EQM l'erreur quadratique moyenne. Exprimé en décibels (dB), le PSNR fournit une quantification logarithmique de la qualité de reconstruction, particulièrement pertinente pour évaluer la préservation des détails fins du texte.

L'utilisation conjointe de ces quatre métriques permet une évaluation quantitative de la performance : tandis que la FM et le PSNR (valeurs élevées souhaitables) quantifient la

qualité globale de l'extraction, la NRM et la DRD (valeurs faibles souhaitables) caractérisent les erreurs commises. Cette approche garantit une évaluation objective avec les méthodes de l'état de l'art, tout en capturant les nuances essentielles à l'évaluation de la qualité du texte extrait.

4.1.2 Méthodes comparées

Étant donné que la vérité terrain (GT) n'existe que pour les composantes textuelles, nous évaluons la qualité de la décomposition à l'aide d'un banc d'essai de binarisation de texte. D'une part, comme référence de l'état de l'art en binarisation de documents MS, nous utilisons la méthode de **Howe** (Howe, 2013). Cinq modèles traditionnels conçus pour l'extraction de texte sont également comparés : **SKKHM** (Spatial Kernel K-Harmonic Means), qui exploite des noyaux spatiaux et des moyennes harmoniques pour regrouper les pixels de texte (Li *et al.*, 2007) ; **GMM** (Gaussian Mixture Models), qui modélise les distributions d'intensité par un mélange de gaussiennes (Hollaus *et al.*, 2018) ; **SAE** (Selectional Autoencoder), un autoencodeur profond entraîné à extraire une représentation discriminative du texte (Calvo-Zaragoza & Gallego, 2019) ; et **ACE v1 & v2** (Adaptive Coherence Estimator), qui mesurent la cohérence spectrale et spatiale pour détecter le texte dans les images multispectrales (Hollaus, Diem & Sablatnig, 2015b; Diem, Hollaus & Sablatnig, 2016). Nous confrontons également ces approches à deux modèles NMF pour la décomposition d'images multispectrales de documents : **MA-ONMF**, un NMF bi-orthogonal, et **VBONMF**, un NMF orthogonal bayésien variationnel, présentant les résultats de l'état-de-l'art sur le jeu de données MStex (Rahiche & Cheriet, 2021, 2022).

D'autre part, pour explorer l'apport des techniques basées sur le modèle linéaire de mélange, cinq auto-encodeurs conçus pour le démélange hyperspectral ont été ré-implémentés : **EndNet**, un autoencodeur estimant simultanément les spectres purs et les abondances (Ozkan, Kaya & Akar, 2019) ; **MTAEU** (Multi-Task Autoencoder for Unmixing), qui combine reconstruction et régularisation spatiale pour un démélange multitâche (Palsson *et al.*, 2019) ; **DAEU** (Denoising Autoencoder for Unmixing), intégrant un module de débruitage pour renforcer la robustesse face au bruit (Palsson *et al.*, 2018) ; **OSPAEU** (Optimized Spatially-Aware Autoencoder for Unmixing), qui optimise une pénalité spatiale afin de préserver la cohérence locale (Dou, Gao,

Zhang, Wang & Wang, 2020) ; et **CNNAEU** (Convolutional Neural Network Autoencoder for Unmixing), exploitant des couches convolutionnelles pour capturer efficacement les dépendances spatiales et étant à la base de l'architecture de PRISM (Palsson *et al.*, 2021). Enfin, nous évaluons **MSFormer**, un modèle ViT à rang capable d'ajuster dynamiquement le nombre de composants spécialisé dans la segmentation d'images médicales MS (Karimijafarbigloo *et al.*, 2024).

4.1.3 Résultats quantitatifs

Le Tableau 4.2 présente les performances moyennes des différentes méthodes sur l'ensemble des 20 images de la collection MSTEx. La méthode proposée PRISM obtient les meilleures performances sur la majorité des métriques indiquant une méthode qui généralise aux différentes configurations d'images, avec notamment un score FM de 86.47%.

Tableau 4.2 Performances moyenne sur les 20 images MS de MSTEx

Method	**FM**(%)↑	**DRD**(10^{-3})↓	**NRM**(10^{-2})↓	**PSNR**↑
Méthodes issues de domaines connexes				
MSFormer	57.15	19.76	16.79	11.55
Endnet	61.23	32.20	14.08	11.87
CNNAEU	65.43	31.02	17.99	13.22
DAEU	71.49	12.33	14.88	14.50
OSPAEU	73.62	18.74	12.98	14.78
MTAEU	73.30	17.01	8.30	13.95
Méthodes pour les images de documents				
SAE	64.98	9.94	14.33	13.45
SKKHM	71.78	11.09	13.37	14.35
Howe	76.96	6.63	8.94	14.94
GMM	80.72	5.12	10.42	16.05
ACE v1	83.85	4.12	9.11	-
MAONMF	85.09	3.59	7.52	-
ACE v2	85.15	3.66	8.29	-
VBONMF	**85.70**	**3.55**	**6.31**	**17.12**
PRISM	**86.47**	**3.11**	**7.30**	**17.24**

Pour mieux comprendre ces résultats moyens et évaluer la robustesse des différentes approches, nous présentons dans les Tableaux 4.3 et 4.4, les résultats détaillés sur chaque image pour les méthodes qui ont pu être réimplémentées. Comme observé sur ces deux tableaux, une grande variabilité peut-être observée dans les résultats pour chaque méthode.

Tableau 4.3 Comparaison de différentes méthodes sur les images de MSTEx1. Les meilleurs résultats sont montrés en **gras noir**, les deuxièmes meilleurs en **gras bleu**

Image	Métrique	Méthodes										
		Howe	SAE	SKKHM	GMM	CNNAEU	EndNet	DAEU	OSPAEU	MSFormer	MTAEU	Ours
z64	FM↑	82.71	72.43	82.21	**87.28**	64.53	76.34	85.88	84.43	66.11	85.24	**86.60**
	DRD↓	4.48	9.24	3.62	**3.14**	6.01	7.97	3.31	3.46	10.76	3.84	**2.90**
	NRM↓	**4.14**	**3.70**	12.24	5.97	24.15	4.59	7.79	9.42	10.56	5.10	7.80
	PSNR↑	15.44	12.66	16.22	**17.25**	13.72	13.63	16.84	16.52	11.96	16.32	**17.00**
z37	FM↑	84.80	75.21	87.22	83.74	78.15	84.25	**88.20**	86.72	68.03	86.44	**88.73**
	DRD↓	4.37	10.13	2.98	3.86	5.13	3.49	**2.69**	3.06	15.46	3.90	**2.62**
	NRM↓	5.98	**5.47**	9.54	12.43	16.51	12.93	9.27	10.03	8.34	**5.36**	7.59
	PSNR↑	13.63	10.74	15.07	14.17	12.98	14.32	**15.36**	14.84	9.19	14.14	**15.38**
z35	FM↑	70.30	66.14	91.07	78.75	78.14	81.78	**92.15**	89.97	82.22	91.14	**91.80**
	DRD↓	17.33	10.01	2.32	5.99	8.32	7.88	**2.20**	2.70	5.49	2.59	**2.06**
	NRM↓	4.55	22.86	3.73	15.06	7.45	**2.53**	4.02	7.38	8.00	**3.04**	4.77
	PSNR↑	11.67	13.32	18.22	15.04	13.80	14.31	**18.76**	17.94	15.04	18.08	**18.63**
z80	FM↑	78.97	79.62	71.61	82.45	65.27	62.72	64.18	**94.46**	69.93	81.55	**94.39**
	DRD↓	10.18	6.83	10.51	4.63	11.61	11.54	8.89	**1.67**	16.53	9.23	**1.71**
	NRM↓	4.91	9.29	15.87	14.45	21.53	23.87	26.35	**3.43**	6.17	4.77	**2.29**
	PSNR↑	13.69	14.37	13.21	15.89	12.63	12.53	13.38	**20.20**	11.49	14.32	**20.03**
z38	FM↑	83.74	13.37	82.17	55.26	**87.72**	86.42	74.31	87.23	68.09	85.59	**88.14**
	DRD↓	3.91	14.73	3.92	14.36	2.98	**2.94**	5.08	2.96	11.58	4.15	**2.71**
	NRM↓	**5.08**	46.45	13.35	26.04	7.38	9.81	20.26	7.53	14.52	**3.85**	6.67
	PSNR↑	13.47	10.29	14.78	12.27	**15.70**	15.70	13.40	15.70	11.55	15.14	**16.12**
z68	FM↑	83.84	72.41	84.45	81.28	69.43	70.63	80.31	81.03	62.09	**88.44**	**90.37**
	DRD↓	5.09	9.31	4.34	6.24	7.00	9.73	4.82	5.38	14.17	**2.57**	**2.05**
	NRM↓	**2.91**	5.59	5.94	6.71	13.81	10.22	12.20	10.78	11.91	4.53	**3.62**
	PSNR↑	16.50	12.73	17.10	15.62	13.61	13.23	15.11	15.45	11.29	**18.45**	**19.25**
z70	FM↑	85.34	57.59	79.67	77.53	59.31	79.98	75.18	74.73	48.13	**85.10**	**91.99**
	DRD↓	4.64	14.52	5.64	5.34	5.48	4.23	7.81	7.63	13.17	**3.15**	**1.99**
	NRM↓	**3.98**	14.53	8.70	13.67	28.29	13.14	13.35	14.16	30.49	8.08	**3.96**
	PSNR↑	16.04	12.33	15.36	15.17	13.28	16.27	14.02	14.35	11.90	**17.42**	**19.74**
z76	FM↑	90.18	54.15	73.45	63.81	82.18	86.94	73.70	68.57	48.12	**88.39**	**92.82**
	DRD↓	2.48	20.11	6.72	10.79	4.24	3.26	9.46	19.53	21.08	**2.89**	**1.82**
	NRM↓	6.43	22.74	17.35	29.52	8.82	5.49	10.77	6.82	21.56	**3.57**	**3.07**
	PSNR↑	16.72	10.83	14.31	12.56	16.19	17.26	13.81	11.81	11.07	**17.84**	**19.35**
z82	FM↑	78.12	80.47	76.47	86.00	62.86	73.99	77.09	77.28	77.33	**86.74**	**88.86**
	DRD↓	7.29	8.70	6.26	**3.60**	6.36	8.63	5.85	25.79	7.16	3.78	**2.59**
	NRM↓	6.35	**2.91**	10.72	8.68	21.05	6.31	15.06	**2.90**	12.69	6.02	5.16
	PSNR↑	14.02	13.77	14.38	16.50	13.27	13.88	14.64	10.96	13.65	**16.80**	**17.91**
z58	FM↑	72.33	67.94	84.04	80.28	77.30	82.45	87.71	82.49	77.74	**88.80**	**88.75**
	DRD↓	10.55	10.18	4.15	5.68	5.02	5.10	3.04	4.74	8.38	**2.94**	**2.71**
	NRM↓	7.24	13.72	9.46	12.54	17.03	7.68	7.61	11.71	7.02	**5.23**	**4.99**
	PSNR↑	12.92	13.20	16.48	15.35	15.22	15.97	17.28	15.65	14.11	**17.42**	**17.70**

Note : FM = F-Measure, DRD = Distance Reciprocal Distortion, NRM = Negative Rate Metric, PSNR = Peak Signal-to-Noise Ratio. ↑ indique une valeur optimale élevée, ↓ indique une valeur optimale basse.

Tableau 4.4 Comparaison de différentes méthodes sur les images de MSTEx2. Les meilleurs résultats sont montrés en **gras noir**, les deuxièmes meilleurs en **gras bleu**

Image	Métrique	Méthodes										
		Howe	SAE	SKKHM	GMM	CNNAEU	EndNet	DAEU	OSPAEU	MSFormer	MTAEU	Ours
z27	FM↑	78.57	49.24	70.59	**81.67**	64.26	79.92	80.56	79.73	72.64	77.86	**82.14**
	DRD↓	6.96	9.73	10.35	**4.86**	8.08	6.04	5.40	5.32	9.38	6.22	**5.28**
	NRM↓	**10.20**	33.29	14.50	11.67	24.56	11.01	11.90	13.51	11.95	13.90	**8.90**
	PSNR↑	12.44	10.59	10.99	**13.57**	11.26	12.82	13.13	13.12	10.98	12.64	**13.23**
z31	FM↑	66.22	63.43	59.34	**86.25**	51.15	54.68	56.61	68.78	44.69	57.50	**86.82**
	DRD↓	12.48	14.52	13.01	**2.77**	12.10	12.76	6.36	4.30	16.70	21.53	**2.41**
	NRM↓	9.29	8.64	14.23	**5.15**	25.51	22.32	30.10	23.52	25.76	**5.92**	8.90
	PSNR↑	14.45	13.82	14.12	**19.21**	13.84	13.93	15.88	**16.92**	12.64	12.31	13.22
z43	FM↑	68.05	70.55	50.01	**81.00**	57.97	54.53	60.49	60.47	51.49	64.71	**81.75**
	DRD↓	10.78	11.74	13.56	**4.56**	19.03	12.33	10.97	11.66	20.04	11.26	**3.07**
	NRM↓	13.33	**6.53**	27.58	**9.14**	11.96	25.06	21.43	20.38	18.05	16.69	10.90
	PSNR↑	13.39	13.17	12.19	**16.08**	11.10	12.48	12.99	12.81	10.65	13.10	**16.44**
z582	FM↑	19.93	7.21	18.13	**79.09**	73.33	17.05	70.26	15.98	43.16	18.91	**83.12**
	DRD↓	11.51	12.19	71.34	**4.71**	7.37	173	5.05	141	12.95	149	**3.44**
	NRM↓	44.38	48.13	36.54	11.21	**10.29**	26.73	22.54	33.31	32.19	24.05	**11.10**
	PSNR↑	13.42	13.19	6.89	**16.85**	15.14	3.05	16.20	3.92	12.91	3.68	**17.82**
z592	FM↑	81.53	78.43	72.92	**86.46**	80.72	16.87	59.90	80.85	60.97	67.52	**86.50**
	DRD↓	5.51	5.69	10.08	**3.19**	4.49	124	6.34	4.87	11.74	14.50	**2.90**
	NRM↓	6.42	8.94	7.09	**4.67**	9.53	32.92	28.50	10.58	16.91	**5.45**	7.80
	PSNR↑	16.95	16.36	14.79	**18.40**	16.95	4.51	15.28	17.11	13.41	13.34	**18.60**
z65	FM↑	**84.35**	81.56	74.05	82.27	47.79	63.71	29.09	59.86	55.02	71.48	**84.06**
	DRD↓	**4.21**	4.35	6.55	4.45	9.02	16.94	86.93	21.94	20.62	10.72	**4.07**
	NRM↓	**6.68**	**6.27**	14.61	9.28	33.33	8.67	21.1	8.11	14.63	9.54	7.40
	PSNR↑	**15.96**	15.01	14.03	15.58	12.27	10.84	4.11	9.98	9.08	12.65	**15.82**
z802	FM↑	81.28	62.04	81.52	**92.73**	7.77	34.43	68.72	88.02	37.76	56.86	**91.57**
	DRD↓	2.35	20.57	5.25	**1.53**	443.71	81.81	15.18	2.92	53.5	26.96	**1.78**
	NRM↓	7.31	4.39	10.28	**4.12**	46.71	9.48	**3.46**	6.53	11.6	4.42	4.38
	PSNR↑	17.87	13.68	18.76	**22.77**	0.72	7.92	14.79	20.46	9.61	12.55	**21.9**
z822	FM↑	**84.64**	81.67	79.09	78.26	76.64	23.95	**84.98**	75.11	38.99	83.61	83.7
	DRD↓	**2.64**	3.42	3.93	4.41	4.06	94.93	2.86	3.87	13.55	**2.73**	2.90
	NRM↓	10.41	**9.44**	14.13	13.24	15.53	19.05	**8.46**	19.28	33.28	11.99	9.60
	PSNR↑	**19.01**	18.12	17.96	17.63	17.36	6.25	**18.92**	17.52	13.16	18.87	18.64
z90	FM↑	**84.64**	68.38	51.36	**83.77**	45.64	49.98	41.75	54.59	26.58	47.94	83.2
	DRD↓	**2.64**	7.72	26.81	4.68	41.21	24.8	52.94	24.47	106.5	40.46	**4.13**
	NRM↓	10.41	18.74	15.43	**7.75**	12.35	19.87	10.29	14.70	16.47	9.31	**9.22**
	PSNR↑	13.86	15.51	11.67	**18.01**	10.03	11.93	8.98	12.08	5.96	10.12	**17.89**
z92	FM↑	72.11	63.88	69.98	69.01	75.59	54.98	76.65	**81.64**	61.55	71.27	**80.44**
	DRD↓	8.68	9.55	10.74	10.09	5.44	25.79	7.56	**4.06**	16.53	11.01	**4.61**
	NRM↓	10.46	17.81	8.93	10.47	14.59	9.42	**6.39**	9.71	10.85	**6.18**	8.5
	PSNR↑	12.96	11.49	11.59	11.58	13.5	8.19	12.62	**14.52**	9.78	11.38	**14.02**

Note : FM = F-Measure, DRD = Distance Reciprocal Distortion, NRM = Negative Rate Metric, PSNR = Peak Signal-to-Noise Ratio. ↑ indique une valeur optimale élevée, ↓ indique une valeur optimale basse.

L'analyse de ces tableaux révèle plusieurs observations importantes. Premièrement, on constate que la collection MSTEx2 présente des défis significativement plus importants que MSTEx1. La plupart des méthodes montrent une dégradation notable de leurs performances sur MSTEx2, avec des cas extrêmes comme l'image z582 où EndNet, OSPAEU, MTAEU et même Howe obtiennent des valeurs catastrophiques, indiquant un échec quasi-total de la binarisation. De manière surprenante, GMM constitue une exception notable, maintenant des performances relativement stables entre les deux collections et obtenant même certains de ses meilleurs résultats sur MSTEx2 (FM de 92.73% sur l'image z802). La méthode de Howe illustre parfaitement le problème de variabilité des performances. Bien qu'elle obtienne d'excellents résultats sur certaines images (z822 avec un FM de 84.64% et un DRD de 2.64×10^{-3}), elle échoue complètement sur d'autres, comme z31 où son FM chute à 66.22% avec un DRD élevé de 12.48×10^{-3}. Cette inconsistance rend certaines méthodes peu fiables face aux variations des documents inhérentes à leur nature.

Parmi les méthodes basées sur les auto-encodeurs, MTAEU se distingue comme étant la plus performante, obtenant régulièrement des résultats compétitifs. Par exemple, sur l'image z68, MTAEU atteint un FM de 88.44% avec un excellent DRD de 2.57×10^{-3}, surpassant toutes les autres approches d'apprentissage profond. Cependant, même MTAEU présente une forte variabilité, avec des échecs notables sur certaines images difficiles de MSTEx2.

En contraste, notre méthode PRISM démontre une remarquable stabilité à travers l'ensemble des images. Elle obtient systématiquement des performances parmi les meilleures sur chaque image, avec des résultats dépassant tous le seuil de 80% en FM. Sur les 20 cube MS testés, PRISM se classe première ou deuxième pour la métrique FM dans 17 cas, et maintient toujours un DRD inférieur à 5.28×10^{-3}. Cette constance, reflétée dans les métriques moyennes, confirme la robustesse de notre approche face à la variabilité des contenus manuscrits, quelle que soit sa structure spatiale, l'âge du document, ou son état de dégradation.

4.1.4 Résultats qualitatifs

L'analyse qualitative présentée dans la Figure 4.1 révèle des différences marquantes entre les performances visuelles des différentes méthodes sur l'image z31 de MSTEx-1. Cette image, particulièrement représentative des défis rencontrés dans l'extraction de texte manuscrit, permet d'observer concrètement les limitations et avantages des différentes approches.

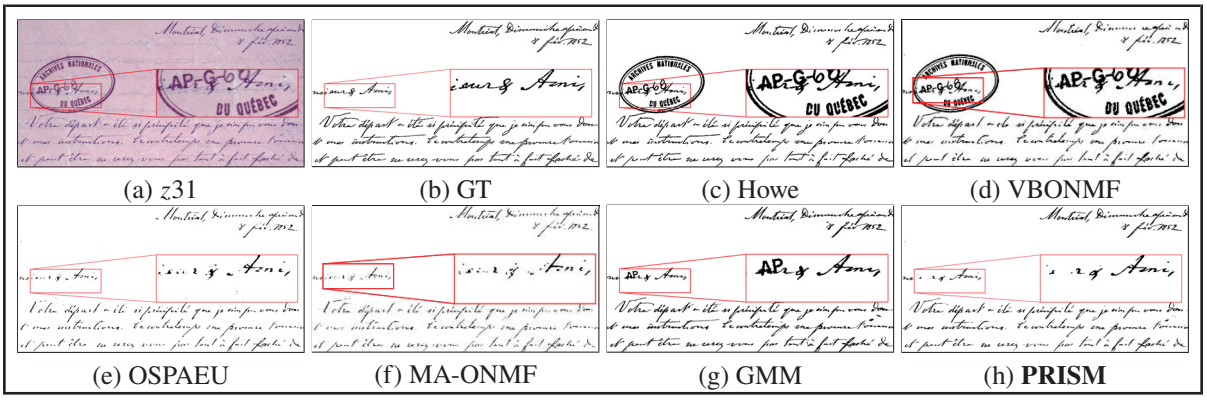

Figure 4.1 Composantes de texte extraites par différentes méthodes, classées par score FM, sur l'image z31 de MSTEx-1. Visualisation suivant Rahiche & Cheriet (2021, 2022)

L'image en pseudo-couleur (a) ainsi que la vérité terrain associée (b) illustrent la complexité de la tâche. Les caractères manuscrits présentant des variations d'épaisseur, et différentes connexions entre lettres. Un tampon des archives nationales a aussi été ajouté sur le texte, rendant la tâche de segmentation encore plus difficile. La méthode de Howe (c) présente des résultats décevants sur cette image, confirmant les observations quantitatives précédentes. L'extraction apparaît incomplète, avec le tampon qui n'est pas séparé du texte et de l'arrière-plan. Cette dégradation visuelle explique directement le score FM particulièrement faible de 66.22% obtenu sur cette image, illustrant parfaitement les problèmes de robustesse de cette approche. VBONMF (d) ne montre pas d'amélioration notable par rapport à Howe sur cette image, bien que présentant des résultats moyens plus haut sur MStex. La méthode ne réussit pas à différencier le tampon du texte. Sur cette image, OSPAEU (e) obtient les meilleurs résultats parmi les méthodes AE ré-implémentées. Cela révèle les manques inhérents aux méthodes d'apprentissage profond se basant seulement sur les pixels. La méthode parvient à distinguer le tampon du texte, mais

a plus de mal à bien identifier les bordures du texte ainsi que les traits plus fins, rendant la lecture plus difficile. MA-ONMF (f) démontre une performance intermédiaire intéressante. L'extraction préserve mieux la connectivité des traits manuscrits que les méthodes précédentes, mais présente encore des difficultés dans la zone sous le tampon. On observe une tendance à lisser excessivement certains détails fins, ce qui peut compromettre la lisibilité de caractères particulièrement stylisés ou dégradés. La méthode GMM (g) surprend par sa robustesse relative sur cette image difficile. L'approche statistique semble bien adaptée à la nature de la distribution des intensités dans cette image, résultant en une extraction relativement propre et cohérente. Cette performance visuelle corrobore les résultats quantitatifs favorables obtenus par GMM sur certaines images de la collection MStex. Le seul défaut, consiste en une partie du tampon qui n'est pas correctement discriminée, résultant en un score FM plus bas.

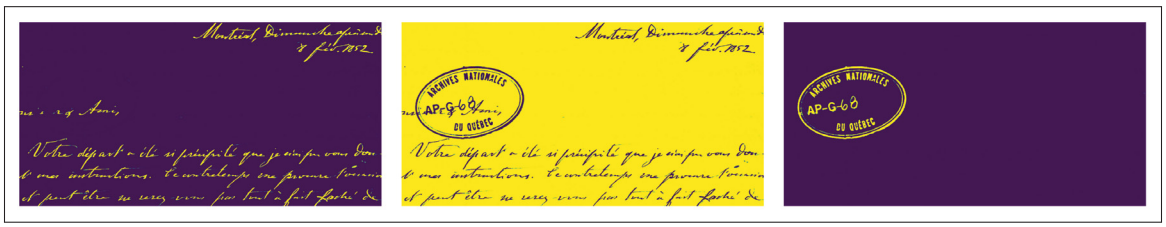

Figure 4.2 Décomposition de PRISM sur l'image z31 de MStex, montrant le texte, le tampon et l'arrière-plan bien décomposé

Enfin, la méthode PRISM (h) se distingue nettement par la qualité de l'extraction obtenue. Le résultat présente une remarquable fidélité à la vérité terrain, étant une des seules méthodes à correctement séparer le tampon du texte. A la différence de OSPAEU et MA-ONMF, la composante extraite préserve les détails fins et les traits manuscrits du texte. Les caractères apparaissent alors nets et complets, améliorant la lisibilité du texte par rapport à d'autres méthodes. Cette qualité visuelle supérieure explique directement les performances quantitatives élevées et constantes de PRISM à travers l'ensemble des images testées. Comme montré sur la Figure 4.2, la méthode proposée décompose l'image en trois composant distincts : le texte, le tampon des archives ainsi que l'arrière plan. Cette analyse qualitative montre que la supériorité quantitative de PRISM pour l'extraction de texte se traduit effectivement par une amélioration visuelle de la décomposition, permettant une meilleur compréhension de l'image décomposée.

4.2 Généralisation de l'approche à différentes configurations d'images multibandes

4.2.1 MSBin : Jeu de données MS de documents anciens

Afin d'évaluer la capacité de généralisation de notre méthode PRISM, il est crucial de la tester sur des configurations d'acquisition multispectrale variées. En effet, les systèmes d'imagerie diffèrent considérablement en termes de nombre de bandes spectrales, de leur répartition et de la plage spectrale couverte. Cette variabilité représente un défi majeur pour les méthodes de traitement, qui doivent s'adapter à des caractéristiques spectrales hétérogènes tout en maintenant leurs performances. Nous évaluons donc notre approche sur la collection MSBin, qui présente une configuration spectrale distincte avec 12 bandes étroites au lieu de 8, permettant ainsi de vérifier la robustesse de PRISM face à différentes modalités d'acquisition. Cette collection comprend 30 cubes multispectraux acquis à l'aide d'une caméra achromatique Phase One IQ260 d'une résolution de 60 mégapixels, capturés sur douze bandes spectrales étroites s'étendant de 365 nm à 940 nm (Hollaus *et al.*, 2019). Celles-ci sont montrées sur le Tab. 4.5.

Tableau 4.5 Détail des bandes spectrales utilisées pour le jeu de données MSbin

Bande	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**
Longueur d'onde (nm)	Toutes	365	450	465	505	535	570	625	700	780	870	940
Filtre de lumière	Blanc	UV	· · · Visible · · ·							IR 1	IR 2	IR 3
Temps d'exposition (s)	0.066	10	0.125	0.1	0.05	0.066	0.166	0.033	0.166	0.2	0.2	0.5

Les images proviennent de la numérisation de deux manuscrits : le Bitola-Triodion ABAN 38 (livre BT) et l'Enina-Apostolus NBMK 1144 (livre EA), ce dernier présentant un état de dégradation sévère. La vérité terrain classe chaque pixel en trois catégories : texte au premier plan (encre ferro-gallique), arrière-plan, ou régions incertaines jugées trop complexes pour une annotation manuelle (ces dernières étant exclues de l'évaluation). Certaines images comportent également un second plan textuel à l'encre rouge. Ces images possèdent une taille spatiale et spectrale plus élevé que celles de la collection MStex.

4.2.2 Résultats quantitatifs

Le Tableau 4.6 présente les performances moyennes des différentes méthodes sur l'ensemble des 30 cubes MS de la collection MSBin. La méthode proposée obtient les meilleures performances sur l'ensemble des métriques, avec notamment un score FM de 82.57%, confirmant sa capacité à traiter efficacement des documents manuscrits aux caractéristiques variées.

Tableau 4.6 Performances moyenne sur les 30 cubes MS de MSBin

Method	**FM**$_{(\%)}\uparrow$	**DRD**$_{(10^{-3})}\downarrow$	**NRM**$_{(10^{-2})}\downarrow$	**PSNR**$\uparrow$
Méthodes issues de domaines connexes				
CNNAEU	67.10	39.52	19.35	10.32
Endnet	67.62	41.71	20.61	10.57
MSFormer	69.87	46.62	12.55	13.18
OSPAEU	71.59	52.90	13.14	11.77
DAEU	72.68	29.73	16.50	11.77
MTAEU	74.02	34.25	**12.81**	11.35
Méthodes pour les images de documents				
MAONMF	-	-	-	-
VBONMF	-	-	-	-
SAE	40.83	50.86	32.42	8.95
Howe	41.50	38.68	31.27	10.48
SKKHM	73.42	26.78	17.43	11.74
GMM	80.00	**20.35**	-	13.18
ACE v2	81.25	20.80	-	13.27
ACE v1	**81.28**	22.03	-	**13.28**
PRISM	**82.57**	**18.71**	**10.45**	**13.74**

Parmi les méthodes spécifiquement conçues pour les documents, on observe une grande disparité de performances. SAE et Howe, deux approches classiques, obtiennent des résultats catastrophiques avec des scores FM de 40.83% et 41.50% respectivement, suggérant leur inadéquation face à la complexité des manuscrits de MSBin. Ces échecs sont particulièrement visibles dans leurs métriques NRM dépassant 30×10^{-2}, témoignant d'une incapacité à préserver correctement les structures du texte. Les résultats pour les méthodes MAONMF et VBONMF, qui obtiennent de bons résultats sur MStex, n'ont pas pu être produits en raison des difficultés computationnelles, renforçant l'idée que MSBin représente un défi significatif.

La plupart des méthodes obtiennent des résultats inférieurs à ceux du jeu de données MStex. En revanche, les méthodes proposant une décomposition parviennent à maintenir des résultats élevés en comparaison avec les méthodes de binarisation uniquement. C'est le cas de SKKHM, qui, bien qu'obtenant des résultats plus bas, affiche un score FM de 73,42 % supérieur à celui de MStex (71,78 %). GMM semble indiquer une certaine robustesse, avec 80 % de FM sur les deux collections. Les deux versions d'ACE obtiennent des performances quasi identiques (FM de 81,25 % et 81,28 %), plus basses que pour MStex. Étrangement, ACE v2 obtient des résultats légèrement plus bas que ACE v1 pour le score FM et le PSNR, mais pas pour la DRD. Cela semble démontrer que l'ensemble de ces métriques doit être pris en compte pour mieux évaluer la qualité du texte extrait, chacune étant complémentaire. Cela suggère également une grande variabilité dans les images MS, chacune présentant ses propres défis.

Les méthodes issues de domaines connexes, bien qu'ayant démontré leur efficacité sur d'autres types de données, peinent manifestement à s'adapter aux spécificités des manuscrits historiques. Cependant, malgré des paramètres qui n'ont pas été adaptés, elles obtiennent des résultats similaires, indiquant une bonne généralisation des méthodes d'apprentissage profond de démélange HS. Étonnamment, MS Former est la méthode dont les résultats augmentent le plus entre les deux jeux de données. Cela peut s'expliquer par le fait que le texte de MSBin est significativement plus large. Il est donc mieux traité par l'architecture transformer qui prend des patchs en entrée, ce qui permet d'obtenir une plus grande précision.

Enfin, notre méthode PRISM démontre sa supériorité en surpassant toutes les approches concurrentes sur l'ensemble des métriques. Avec un FM de 82.57%, elle dépasse d'environ 1.3 points de pourcentage les meilleures méthodes alternatives. Plus significatif encore, elle obtient simultanément le meilleur DRD (18.71×10^{-3}) et le meilleur NRM (10.45×10^{-2}), ce qui témoigne de sa capacité à équilibrer précision de détection et qualité visuelle. Le PSNR de 13.74, supérieur de 0.46 points à ACE v1, confirme la qualité globale supérieure des images binarisées. Cette domination sur l'ensemble des métriques, contrairement aux autres méthodes qui excellent généralement sur certains aspects au détriment d'autres, souligne la robustesse et l'équilibre de notre approche face à la diversité des défis présents dans la collection MSBin.

4.2.3 Résultats qualitatifs

L'analyse qualitative présentée dans la Figure 4.3 offre une perspective complémentaire sur les performances des différentes méthodes d'extraction de texte manuscrit sur la base de données MSBin. Cette image, présente plusieurs défis caractéristiques : un texte manuscrit avec deux différentes couleurs, un support papier avec des variations d'intensité, ainsi qu'une petite déchirure ajoutant un fond complexe visible en haut à gauche dans l'image pseudo-couleur (b).

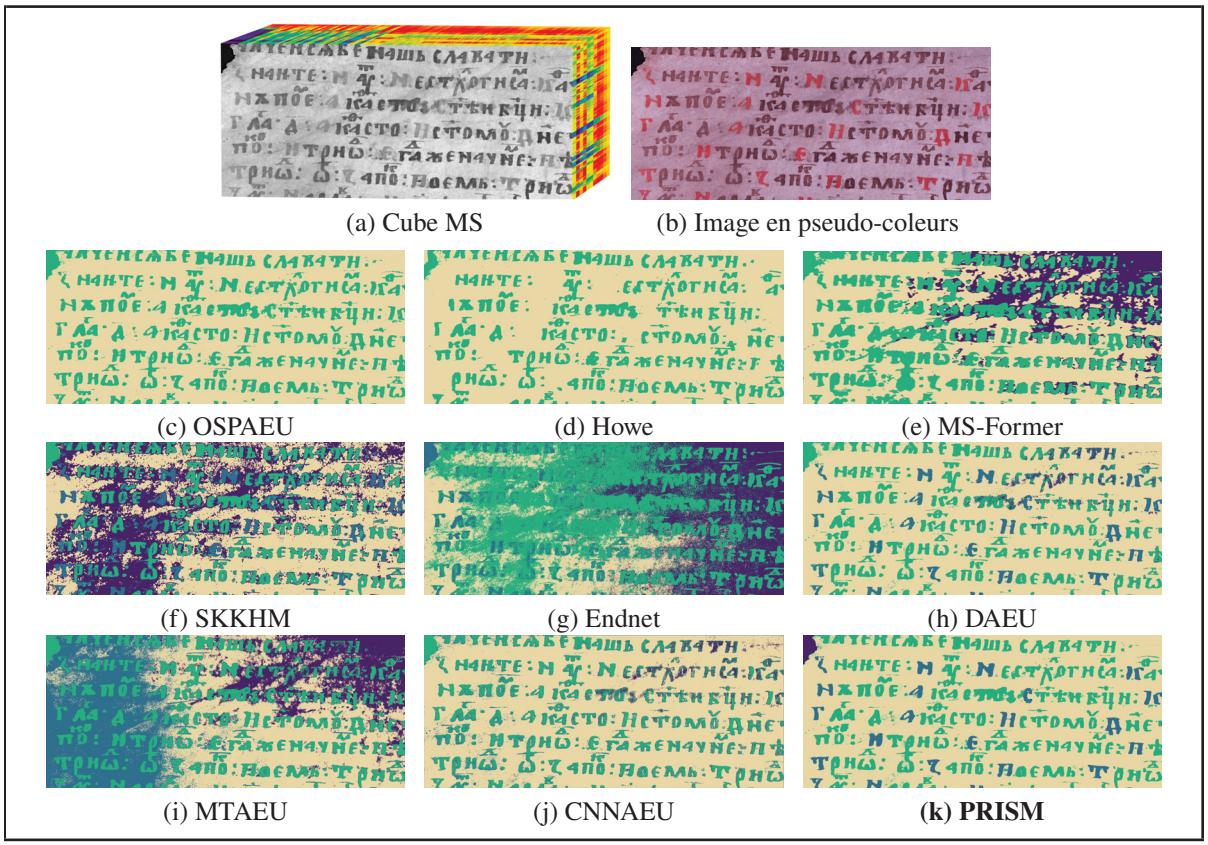

Figure 4.3 Décomposition de PRISM sur l'image BT56 de MSBin

Le cube MS (a) ainsi que l'image en pseudo-couleurs correspondante, révèlent la richesse de l'information spectrale disponible, que les différentes méthodes exploitent avec des degrés de succès variables. Parmi ces différentes méthodes, toutes ont vu leur rang fixé à 4 à l'exception de Howe (d), qui peut seulement identifier un texte et un arrière-plan, ainsi que de MS-Former (e), et de PRISM (k), qui proposent une sélection du rang dynamique.

On peut alors observer plusieurs comportements pour la décomposition du texte. Certaines méthodes réussissent à faire la distinction entre le texte principal et la deuxième encre, comme Howe, SKKHM, DAEU ou PRISM, tandis que les autres ne distinguent pas ces encres différentes.

OSPAEU (c) produit une extraction relativement propre mais ne distingue que deux composantes importantes. Cette limitation suggère une capacité insuffisante du modèle à discriminer finement les signatures spectrales. La méthode de Howe (d), bien qu'étant une approche classique respectée, montre ici ses limites face à la complexité spectrale de l'image. L'extraction apparaît incomplète avec une perte significative de détails, la perte de la distinction de la deuxième encre et une confusion entre certaines parties du texte et de l'arrière-plan, confirmant les observations faites sur MSTEx. MS-Former (e) démontre les défis d'une architecture transformer pour les tâches de segmentation fine. Les caractères extraits sont grossiers, et des artefacts d'illuminations sont identifiés comme des composants à part entière, suggérant que l'attention multi-tête seule ne suffit pas à capturer toutes les nuances spectrales nécessaires à une séparation optimale. SKKHM (f) présente une contamination similaire par les variations d'illuminations, représenté par la composantes violette, illustrant les défis inhérents aux approches de clustering appliquées aux données MS. Les méthodes basées sur les auto-encodeurs ; Endnet (g), DAEU (h), MTAEU (i) et CNNAEU (j), montrent des performances variables. Endnet et MTAEU souffrent particulièrement d'une mauvaise séparation spectrale, avec des composants extraits qui ne sont pas interprétables, compromettant la lisibilité. CNNAEU présente un compromis intéressant avec moins d'artefacts colorés mais ne réussit pas à distinguer les deux couleurs de texte. DAEU offre une amélioration notable avec une extraction plus nette mais ne réussit pas à différencier l'arrière-plan du texte principal (en vert).

Finalement, la méthode proposée PRISM (k) se distingue une fois de plus par la qualité de l'extraction obtenue. Le texte apparaît net et complet, avec une séparation des deux couleurs du texte. Parmi toutes les méthodes, PRISM est la seule à réussir à distinguer l'arrière-plan (en violet) du texte principale. L'absence quasi-totale d'artefacts colorés et la préservation fidèle de tous les détails du manuscrit, y compris les traits les plus fins et les variations d'intensité subtiles, confirment la robustesse de l'approche proposée. Cette supériorité visuelle corrobore les résultats

quantitatifs et démontre la capacité de PRISM à exploiter efficacement l'information MS pour produire des décompositions de haute qualité et interprétables. L'avantage de la décomposition par rapport à une approche de binarisation directe est illustré dans la Figure 4.4.

(a) Image pseudo-couleur (b) Texte binarisé par Howe **(c) PRISM**

Figure 4.4 Comparaison entre décomposition PRISM et binarisation seule (Howe) sur l'image EA58, montrant l'avantage de la décomposition pour révéler les textes dégradés

La méthode de Howe, référence en binarisation MS, inverse les composantes en classifiant l'arrière-plan comme texte. L'approche proposée réussit à identifier le texte, tout en proposant une décomposition physiquement interprétable. Cette supériorité qualitative se traduit quantitativement par un gain de performance de 28,4 points de FM, moyenné sur l'ensemble de MSBin et MSTex.

4.3 Application et validation sur des données hyperspectrales satellitaires

L'analyse comparative présentée dans les sections précédentes établit la supériorité de la méthode PRISM dans le contexte de l'extraction de texte à partir d'images multispectrales, démontrant des performances quantitatives et qualitatives supérieures aux approches concurrentes. Différentes architectures hybrides de démélange HS ont aussi été testées en référence, réussissant à généraliser sans adaptations des paramètres. Cette observation empirique conduit à une problématique scientifique fondamentale concernant la proximité entre ces deux domaines : *dans quelle mesure une approche développée spécifiquement pour la séparation de composantes textuelles peut-elle être généralisée aux problèmes classiques de démélange hyperspectral ?*

Cette question de recherche trouve sa justification dans l'analyse des fondements théoriques communs aux deux domaines. En effet, tant l'extraction de texte MS que le démélange HS reposent sur le même cadre mathématique de séparation aveugle de source (SAS), formalisé par le modèle linéaire de mélange (LMM). L'évaluation de PRISM sur des données HS constitue donc une validation cruciale permettant d'établir la robustesse de l'algorithme face à des caractéristiques spectrales différentes de celles des manuscrits. Cela pourrait démontrer une généralisation au-delà du domaine d'application initial, soulignant sa force d'adaptation.

4.3.1 Initialisation par VCA

L'algorithme Vertex Component Analysis (VCA) est une méthode largement utilisée pour extraire des signatures spectrales dans les données HS sous l'hypothèse de l'existence de «pixels purs» (Nascimento & Dias, 2005). La VCA repose sur des projections itératives dans un espace de dimension réduite, de façon à faire apparaître tour à tour les sommets (vertices) du simplexe englobant les données. Concrètement, à chaque itération on projette l'ensemble des spectres le long d'une direction aléatoire orthogonale au sous-espace déjà identifié, puis on sélectionne le pixel extrême (maximum de la projection) comme nouvelle signature spectrale pure.

Initialiser les méthodes par des signatures spectrales extraites par VCA permet de placer le point de départ de l'algorithme de démélange dans une configuration physiquement plausible, ce qui accélère la convergence et réduit le risque d'être piégé dans un optimum local. Plusieurs travaux ont montré que, comparé à une initialisation aléatoire ou fondée sur une simple réduction de dimension (PCA, ICA), l'utilisation de VCA améliore significativement la précision de l'estimation des abondances et est devenu un standard pour l'initialisation dans ce domaine.

Ainsi, pour l'application de PRISM aux images hyperspectrales, une étape d'initialisation par VCA est ajoutée pour assurer à la fois la rapidité de convergence et la fiabilité des résultats, tout en exploitant pleinement les hypothèses initiales propres à ce domaine.

4.3.2 Jasper Ridge : Jeu de données de démélange HS

Le site de Jasper Ridge, situé en Californie du Nord, fournit une image hyperspectrale largement utilisée pour évaluer les algorithmes de démélange spectral. L'acquisition a été réalisée à l'aide d'un imageur Aisa Eagle, couvrant 224 bandes spectrales dans la gamme 400–1000nm, avec une résolution spatiale de 100×100 pixels et une taille de pixel au sol d'environ 17m. Quatre matériaux dominants sont annotés sur l'image observée : la forêt, l'eau, le sol nu et l'asphalte. Ces classes servent de références pour valider la précision des cartes d'abondance estimées.

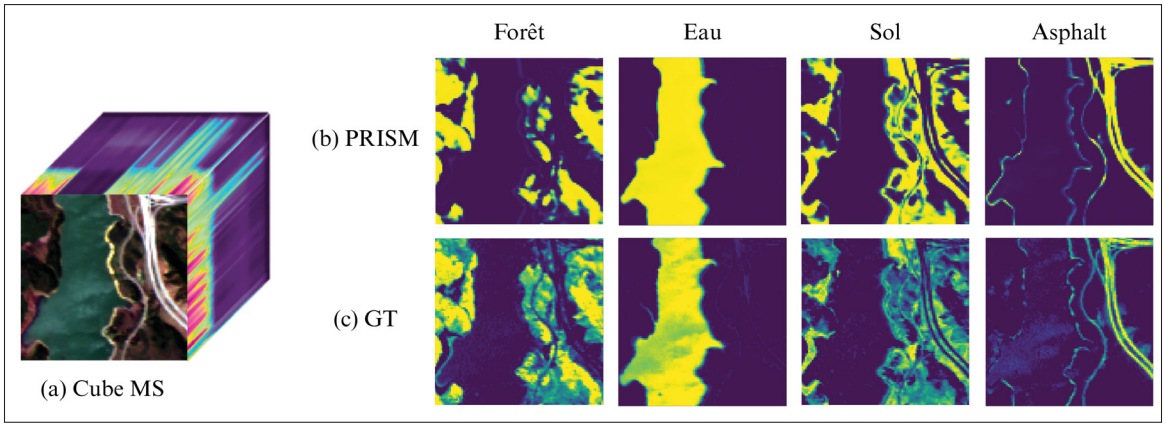

Figure 4.5 Décomposition de PRISM sur l'image HS de Jasper Ridge

La Figure 4.5 montre les cartes d'abondance extraites par PRISM pour chaque materiel, avec :

- **Forêt** : la zone boisée sur la rive gauche est clairement délimitée, avec un contraste élevé entre la végétation dense (valeurs d'abondance élevées) et les zones non forestières (RMSE= 0.06).
- **Eau** : la rivière centrale apparaît avec une abondance très homogène et maximale, attestant de la pureté spectrale de l'eau selon PRISM (RMSE= 0.22).
- **Sol** : les terrains dégagés environnants sont bien extraits, même dans les zones de transition où le sol se mélange à la végétation (RMSE= 0.1049).
- **Asphalte** : les différentes routes est reconstituée avec précision, y compris ses bordures fines, démontrant la capacité de PRISM à capturer des signatures discrètes (RMSE= 0.20).

La racine de l'erreur quadratique moyenne (RMSE) sur l'ensemble des classes est de 0.146, ce qui se situe dans la plage basse des valeurs observées dans la littérature (entre $[0.0838; 0.2943]$).

Les résultats qualitatifs et quantitatifs confirment ainsi l'adéquation des cartes d'abondance produites par PRISM avec les annotations du sol, illustrant sa robustesse pour le démélange HS.

4.3.3 Urban : Jeu de données de démélange HS multi-composantes

Le jeu de données Urban est acquis par le capteur HYDICE au-dessus de la ville de Houston. C'est un jeu de référence pour le domaine du démélange HS. Il se compose d'une image de 307×307 pixels, couvrant 210 bandes spectrales dans la gamme 400–2500 nm. Ce jeu de données est caractérisé par une forte hétérogénéité spectrale et spatiale, notamment en raison de la coexistence de matériaux aux signatures proches comme le béton, l'asphalte, les toits ou la végétation. Une particularité intéressante de ce jeu de données réside dans la disponibilité de plusieurs vérités terrain annotées, définissant trois scénarios de démélange correspondant respectivement à 6, 5 ou 4 matériaux. Ces trois variantes permettent de tester la robustesse des méthodes de démélange face à des définitions de classes plus ou moins précises, et offrent une opportunité d'analyser l'adéquation du rang dynamique proposé par PRISM. Suivant la méthode de Palsson *et al.* (2022), la RMSE entre les abondances générées et la référence GT correspondante a été calculée, en retenant le meilleur résultat parmi dix exécutions indépendantes afin d'atténuer les effets d'initialisation pour les trois scénarios. Dans cette configuration, pour notre méthode, le rang minimal possible r_{min} a été fixé à quatre pour chaque exécution, et les cartes d'abondance correspondantes à chaque scénario GT ont été sauvegardées. Les résultats sont présentés ci-dessous dans les Tab. 4.7, 4.8, et 4.9 et sur la Fig. 4.6.

Tableau 4.7 Comparaison quantitative avec les méthodes de démélange HS sur le jeu de données Urban pour SIX éléments. Les meilleurs résultats en **gras** et les deuxièmes en **bleu**

Métrique	Élément	CNNAEU	Endnet	DAEU	OSPAEU	MTAEU	PRISM
RMSE ↓	Asphalte	0.2270	0.1528	**0.1322**	0.2994	**0.1517**	0.1786
	Herbe	0.3622	0.2141	0.2352	**0.1782**	**0.1862**	0.2080
	Arbre	0.1972	**0.0939**	0.1492	**0.1358**	0.1152	0.1492
	Toit	0.1252	0.1060	**0.0915**	0.1460	0.1152	**0.0852**
	Sol	0.2096	0.2017	0.2195	0.2354	**0.1395**	**0.1732**
	Métal	0.1710	0.2508	**0.1606**	**0.0844**	0.1808	0.1881
	Moyenne	0.2154	0.1699	0.1647	0.1847	**0.1515**	**0.1637**

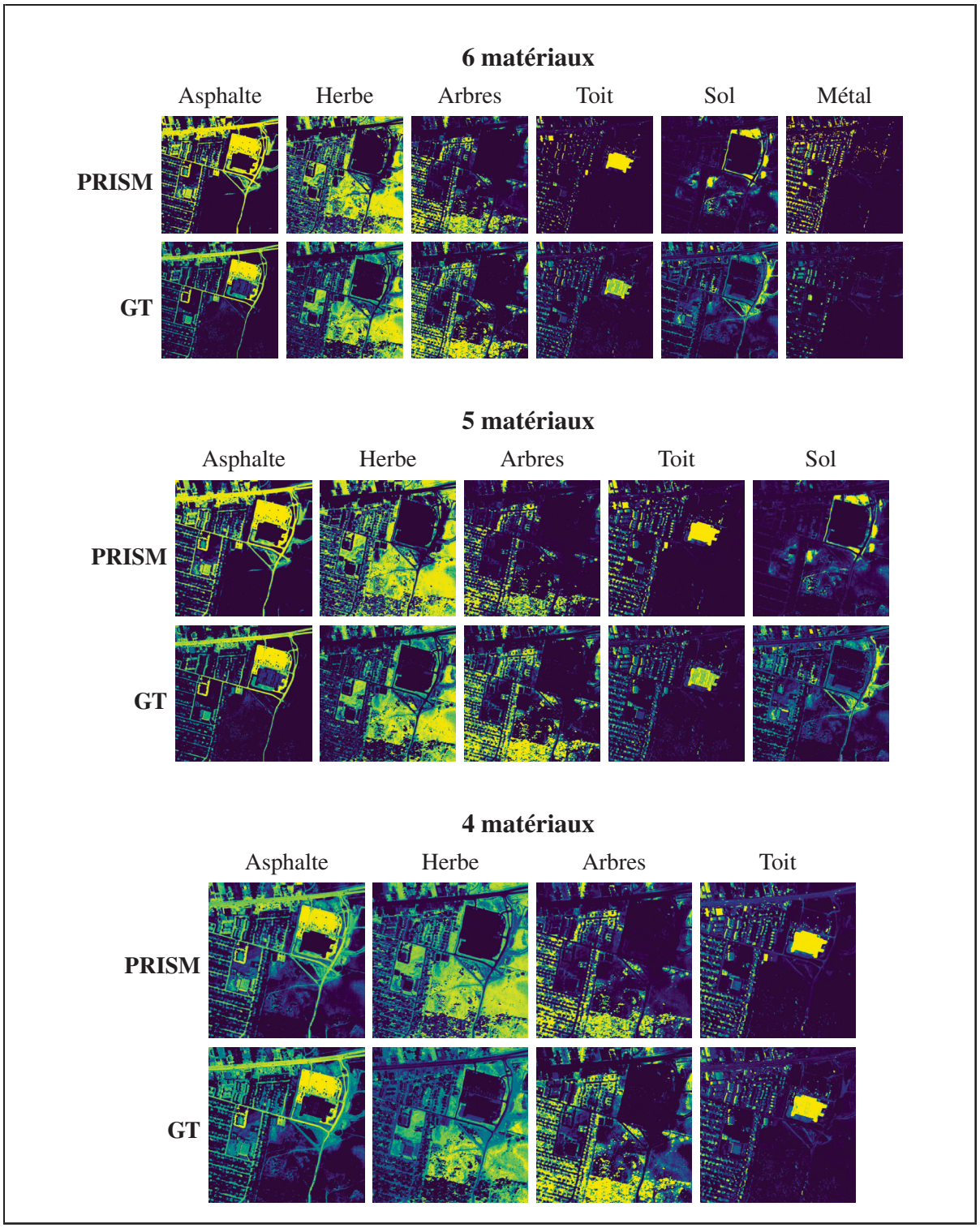

Figure 4.6 Évaluation qualitative de PRISM sur le dataset Urban pour 6, 5 et 4 matériaux : chaque groupe commence par la décomposition proposée suivie de la ligne GT

Tableau 4.8 Comparaison quantitative avec les méthodes de démélange HS sur le jeu de données Urban pour CINQ éléments. Les meilleurs résultats sont en **gras** et les deuxièmes meilleurs en **bleu**

Métrique	Élément	CNNAEU	Endnet	DAEU	OSPAEU	MTAEU	PRISM
RMSE ↓	Asphalte	0.2499	**0.1102**	**0.1266**	0.3159	0.1295	0.2221
	Herbe	0.2563	**0.1688**	0.1856	0.2064	**0.1620**	0.2113
	Arbre	0.2022	**0.1082**	0.1169	0.1644	**0.1105**	0.1197
	Toit	0.1212	0.0870	0.1022	0.1563	**0.0693**	**0.0591**
	Sol	0.2641	**0.1534**	0.1815	0.2410	**0.1157**	0.2326
	Moyenne	0.2187	**0.1255**	0.1426	0.2168	**0.1117**	0.1689

Tableau 4.9 Comparaison quantitative avec les méthodes de démélange HS sur le jeu de données Urban pour QUATRE éléments. Les meilleurs résultats sont en **gras** et les deuxièmes meilleurs en **bleu**

Métrique	Élément	CNNAEU	Endnet	DAEU	OSPAEU	MTAEU	PRISM
RMSE ↓	Asphalte	0.2369	**0.1084**	0.1703	0.3028	0.1426	**0.0948**
	Herbe	0.2756	0.1660	0.1678	0.2688	**0.1346**	**0.1562**
	Arbre	0.2070	0.1019	**0.0762**	0.2134	**0.0951**	0.1252
	Toit	0.1876	**0.0845**	0.0867	0.2876	0.0904	**0.0560**
	Moyenne	0.2268	**0.1152**	0.1253	0.2682	0.1157	**0.1081**

PRISM démontre des performances compétitives par rapport aux méthodes de démélange HS sur ce jeu de données, avec des résultats particulièrement solides pour les cartes d'abondance représentant le toit. Bien que PRISM n'obtienne pas la RMSE la plus basse pour toutes les classes de matériaux, il surpasse systématiquement la méthode de référence CNNAEU qui a servi de base à son architecture. La RMSE plus élevée, notamment pour les composants Asphalte et Sol, peut être attribuée à une similitude entre ces deux matériaux, PRISM classant certaines sentiers hors route comme Asphalte plutôt que Sol. Cette confusion est atténuée lorsque ces composants sont combinés en un seul matériau dans le dernier scénario, produisant même les meilleurs résultats quantitatifs. PRISM, grâce à son adaptation dynamique du rang, est aussi la seule méthode capable de générer les résultats pour les trois scénarios dans un même entraînement. Cela démontre la force de son système d'élagage pour l'interprétation des décompositions.

4.4 Exploration d'applications aux Images RVB et monocanales

4.4.1 Problème du cas sous-déterminé

Dans le cadre de l'analyse d'images multibandes, comme celles issues de documents historiques, on cherche souvent à retrouver un certain nombre de sources latentes à partir d'un nombre donné de bandes spectrales. Ce problème devient particulièrement complexe lorsque le nombre de sources à estimer dépasse le nombre de bandes disponibles : on parle alors de cas sous-déterminé. Intuitivement, cela signifie que l'on dispose de moins d'informations que de variables inconnues, ce qui rend le problème mal posé sans hypothèses ou contraintes supplémentaires. Mathématiquement, on considère la matrice $\mathbf{Y} \in \mathbb{R}_+^{b \times n}$, où b est le nombre de bandes spectrales et n le nombre de pixels. L'objectif est de factoriser cette matrice avec $\mathbf{U} \in \mathbb{R}_+^{b \times r}$ représentant les signatures spectrales estimées et $\mathbf{V} \in \mathbb{R}_+^{r \times n}$ les cartes d'abondance correspondantes. Lorsque le rang r dépasse le nombre de bandes b (*i.e.*, $r > b$), le système devient sous-déterminé : le nombre d'inconnues est alors supérieur au nombre d'observations disponibles par pixel.

Dans ce contexte, des techniques spécifiques deviennent indispensables pour restreindre l'espace des solutions à une abondance parcimonieuse réaliste. Ce cadre souligne la nécessité du contexte spatial : chaque pixel n'est plus considéré seul mais dépend aussi de ses différents voisins.

4.4.2 DIBCO : Jeu de données RVB de documents

Le concours Document Image Binarization Contest (DIBCO), propose un ensemble de données annotées pour l'évaluation des méthodes de binarisation de documents. L'édition 2018 contient des images de documents numérisés en couleur (RVB), dégradés par du bruit, des taches, des plis ou une faible lisibilité, avec leurs masques binaires de vérité terrain correspondants. Les documents incluent des textes manuscrits, imprimés ou mixtes, posant ainsi un défi varié pour les méthodes de séparation de texte et de fond.

Figure 4.7 Décomposition de PRISM sur l'image 1 de DIBCO 2018

Sur la Figure 4.7, on observe la décomposition obtenue par le modèle PRISM appliqué à la première image du jeu de données DIBCO 2018. L'image originale (en haut à gauche) présente un document manuscrit avec du texte encre noire, quelques annotations en encre plus claire représentant le texte du verso de la feuille, des taches visibles, ainsi que différentes dégradations. Les autres sous-figures montrent différentes composantes extraites par PRISM : certaines mettent en évidence les structures textuelles principales (en haut centre), d'autres isolent des artefacts ou des éléments de fond (en bas à gauche et au centre). La dernière image (en bas à droite) illustre une abondance mettant en évidence le texte du verso, proche d'un masque de binarisation. PRISM parvient ainsi à dissocier efficacement les différentes sources présentes dans l'image, en particulier le texte manuscrit, malgré la présence de quelques dégradations visuelles.

Le Tableau 4.10 présente les performances de PRISM comparées à celles des autres participants au concours DIBCO 2018, selon trois métriques de binarisation. PRISM se positionne nettement au-dessus des méthodes classiques de binarisation RVB (Sauvola, Otsu) et dépasse la plupart des méthodes du concours. La méthode gagnante combine plusieurs pré/post-traitements (*e.g.*, transformée bottom-hat, binarisation de Howe, et élimination du bruit) ainsi qu'un pré-entraînement sur les images des éditions précédentes similaires (DIBCO 2017). Ces choix lui assurent les meilleurs scores, au prix d'une approche spécialisée et dépendante des données d'entraînement. En comparaison, PRISM, conçu comme une méthode générique et non supervisée, propose une décomposition bout-en-bout qui extrait plusieurs composantes sans

Tableau 4.10 Résultats de PRISM parmi les méthodes du concours DIBCO 2018

Methode	**FM**(%)↑	**DRD**↓	**PSNR**↑
Méthodes du concours DIBCO 2018			
Méthode gagnante (supervisée)	**88.34**	**4.92**	**19.11**
PRISM	**76.87**	**8.99**	**15.34**
Méthode 7	73.45	26.24	14.62
Méthode 2	70.01	17.45	13.58
Méthode 3b	64.52	16.67	13.57
PRISM s/ attention spatiale	56.96	19.13	12.53
Méthode 5	56.08	28.99	11.50
Méthode 6	46.35	24.56	11.79
Méthode 3a	43.36	40.80	10.42
Méthode 4	41.87	37.36	10.38
Binarisation classique RVB			
Sauvola	67.81	17.69	13.78
Otsu	51.45	59.07	9.74

être spécifiquement adapté au dataset DIBCO. Malgré cette absence d'optimisation, PRISM obtient la deuxième place, surpassant plusieurs approches spécialisées. Ces résultats soulignent la robustesse de PRISM dans un contexte de décomposition sous-déterminée et l'apport du contexte spatial (+19.9 FM) pour compenser la perte d'information spectrale (*voir* Tab. 4.10).

4.4.3 Différences entre images de documents et images naturelles

Les images de documents et les images naturelles représentent deux catégories d'images fondamentalement distinctes, dont les caractéristiques influencent directement la conception et l'efficacité des modèles de vision par ordinateur. Au-delà des différences évidentes de structure et de contenu, la distinction la plus fondamentale entre les images de documents et les images naturelles réside dans leurs propriétés statistiques. Alors que les images naturelles sont complexes, organiques et riches en occlusions, les documents présentent une nature structurée et prévisible. Cette différence est particulièrement manifeste dans le domaine de Fourier. Dans cet espace fréquentiel, les images naturelles démontrent une décroissance en loi de puissance $1/f^{\alpha}$ (où α est proche de 2), propriété universelle observable dans de nombreux phénomènes naturels. Cela illustre et reflète l'organisation hiérarchique et fractale du monde naturel, où le spectre de puissance suit exactement une relation linéaire en représentation log-log, caractéristique d'une

distribution énergétique décroissant progressivement avec la fréquence. À l'opposé, les images de documents, constituées d'artefacts répétitifs (*e.g.*, lignes de texte équidistantes, caractères uniformes, structures tabulaires), présentent un spectre marqué par des pics à différentes fréquences spécifiques. Ces singularités spectrales correspondent directement aux périodicités spatiales inhérentes à la mise en page documentaire, illustrées sur la Fig. 4.8.

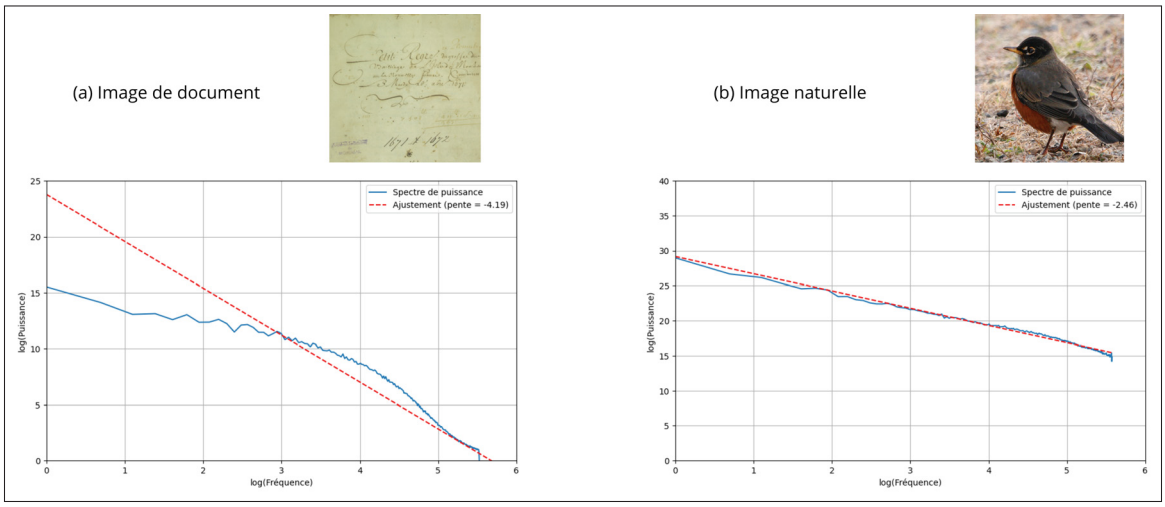

Figure 4.8 Analyse spectrale de Fourier en représentation log-log pour une image de document (a), et une image naturelle (b). L'ajustement linéaire (rouge) révèle que seules les images naturelles suivent une loi de puissance

Cette distinction éclaire d'ailleurs les performances remarquables des petits CNN simples comme PRISM, pour traiter les images de documents. Le théorème de convolution établit que cette opération dans le domaine spatial équivaut à une multiplication dans le domaine de Fourier. Les filtres convolutifs peuvent ainsi efficacement capturer les motifs périodiques et les structures régulières caractéristiques des documents, expliquant pourquoi des architectures CNN relativement peu profondes atteignent des bonnes performances. En revanche, la complexité spectrale des images naturelles, avec leur continuum de fréquences et l'absence de régularités exploitables, nécessite des architectures plus sophistiquées pour les décomposer adéquatement.

4.4.4 Analyse des représentations d'encodeurs pré-entraînés comme cubes HS

Face aux limitations des CNN légers pour les images RVB naturelles, une approche alternative pourrait consister à exploiter la puissance des architectures transformers pré-entraînées. Ces modèles, ayant appris des représentations riches et complexes sur des millions d'images naturelles, réussisent a encoder une compréhension profonde des structures visuelles. La Figure 4.9 illustre la différence de qualité entre les représentations CNN et Dino v2 pré-entraînés. L'image originale (a) montre des bananes sur un support en papier journal. Lorsqu'on applique un algorithme de k-moyennes sur les embeddings produits par un CNN classique (b), la segmentation résultante se limite principalement à une distinction basée sur les couleurs de l'image et des textures locales sans cohérence sémantique claire. En revanche, le même algorithme appliqué aux embeddings de Dino v2 (c) produit une segmentation nettement plus structurée et sémantiquement cohérente, distinguant les différentes régions de l'image (*i.e.*, bananes, journal, mangues, et arrière-plan).

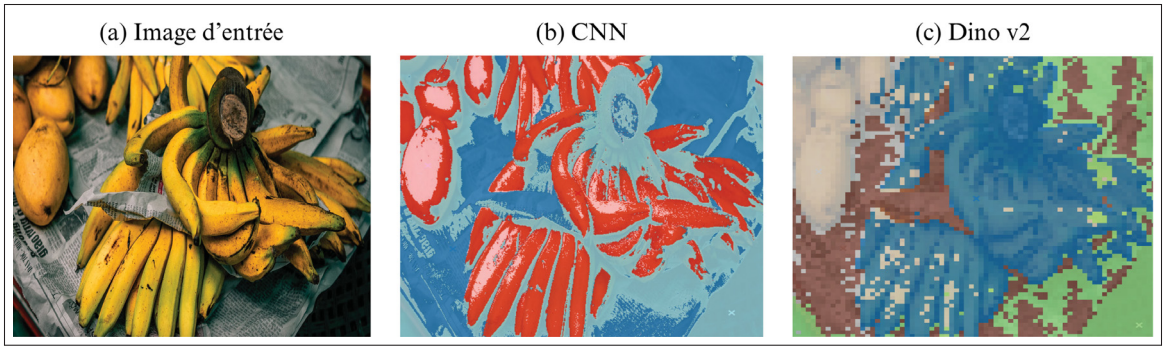

Figure 4.9 Analyse comparative des représentations apprises sur des images naturelles : (a) Image naturelle originale, (b) k-moyennes sur les embeddings d'un CNN pré-entraîné montrant une segmentation limitée, (c) k-moyennes sur les embeddings de Dino v2 démontrant une segmentation sémantique riche

L'idée est alors de traiter les représentations produite par des modèles ViT comme Dino v2, en les considérant comme des cubes HS. Chaque dimension de l'embedding joue alors le rôle d'une bande spectrale, permettant d'appliquer notre méthode PRISM directement sur ces représentations de haut niveau. Ainsi, plutôt que d'analyser les pixels bruts d'images naturelles complexes, PRISM peut opérer sur des représentations déjà enrichies sémantiquement, combinant la simplicité de notre approche avec la puissance des ces architectures pré-entraînés.

Figure 4.10 Résultats de segmentation binaire obtenus par l'application de PRISM sur les embeddings Dino v2. Chaque ligne présente une image naturelle sélectionnée aléatoirement dans le dataset ImageNet, suivie de différents masques binaires extraits automatiquement par la combinaison PRISM + Dino v2, illustrant la capacité de la méthode à identifier diverses structures sémantiques sans supervision

La Figure 4.10 démontre l'application de PRISM sur les embeddings Dino v2 pour segmenter des images naturelles complexes. Pour chaque image test (motos, antenne parabolique, bus scolaire, poubelle, chien), PRISM extrait automatiquement plusieurs segmentations binaires pertinentes sans aucune supervision. Les résultats révèlent que les embeddings Dino v2, traités comme des cubes HS, contiennent une richesse d'information sémantique exploitable. Par exemple, pour l'image des motards, PRISM parvient à isoler séparément les silhouettes des pilotes, les

motos, et même des éléments d'arrière-plan. L'antenne parabolique est segmentée, distinguant la parabole du support et du ciel. Pour le bus scolaire, la méthode sépare efficacement le véhicule de son environnement urbain, tandis que la poubelle est extraite proprement de l'arrière-plan. Particulièrement remarquable, la segmentation du husky démontre la capacité de PRISM à capturer des formes complexes de différentes couleurs et textures. Cette diversité de segmentations pertinentes, malgré quelques artefacts résiduels, est obtenue sans apprentissage spécifique sur ces catégories d'objets, validant l'hypothèse que les représentations des transformers pré-entraînés peuvent être efficacement exploités comme cube HS. Ainsi, cela pourrait ouvrir l'application du modèle pour la segmentation d'images naturelles ou même pour l'analyse d'embeddings de ViT.

4.5 Synthèse expérimentale

Ces évaluations expérimentales ont permis de valider la capacité de **généralisation de notre approche hybride** tout en révélant des pistes d'application qui dépassent le cadre initialement envisagé pour ce mémoire. Sur les images multibandes de documents historiques, PRISM remplit son objectif premier avec des performances supérieures aux méthodes existantes sur diverses configurations spectrales. Ces résultats confirment l'adéquation de l'approche pour son domaine d'application cible. L'**application aux données HS** de télédétection démontre qu'une architecture unique peut alors traiter efficacement des données historiques et satellitaires, démontrant la généralité du LMM pour le démélange spectral. Si la séparation de sources sur images RVB de documents reste dans la continuité logique des travaux, l'application aux représentations profondes de Dino v2 ouvre une perspective non anticipée. En traitant les embeddings comme des cubes HS, PRISM produit des segmentations sémantiquement cohérentes sans apprentissage spécifique des catégories d'objets. Bien que préliminaires, ces résultats suggèrent que PRISM pourrait avoir des applications dans la vision par ordinateur moderne, pour l'**interprétabilité** des représentions profondes ou même la **génération non-supervisée d'annotations**.

Au-delà de la validation des trois axes de recherche initiaux, ces expériences suggèrent donc que l'architecture proposée possède une flexibilité qui pourrait s'étendre à d'autres domaines nécessitant une décomposition interprétable de données multi-dimensionnelles.

CONCLUSION ET RECOMMANDATIONS

Ce mémoire s'est attaqué à un défi majeur de l'analyse d'images de documents historiques : développer une méthode de décomposition d'images multispectrales qui soit à la fois **performante**, **automatique** et **interprétable**. Face aux limitations des approches existantes, nous avons proposé **PRISM**, une **architecture hybride** qui combine l'interprétabilité physique de la factorisation matricielle non-négative avec la **puissance de modélisation** des réseaux de neurones convolutifs.

5.1 Synthèse des contributions

Les trois besoins fondamentaux identifiés en introduction : interprétabilité, amélioration des performances et non supervision, ont guidé notre démarche de recherche. L'analyse systématique de l'état de l'art (Chapitre 1) a confirmé ces besoins tout en révélant des lacunes spécifiques suggérant qu'une approche hybride serait la plus prometteuse pour y répondre. Ces constats ont structuré nos trois contributions principales :

1. **Synergie NMF-Autoencodeur** : Notre première contribution réside dans la conception d'une architecture qui intègre naturellement les contraintes de non-négativité au sein d'un autoencodeur convolutif. Cette approche a démontré sa capacité à préserver l'interprétabilité physique des composantes extraites tout en bénéficiant de la modélisation spatiale et non-linéaire des réseaux profonds. Les résultats expérimentaux ont confirmé que cette synergie produit des décompositions plus précises et plus robustes que les méthodes traditionnelles, avec des gains de performance dans les coûts de calculs.
2. **Sélection automatique du rang** : La deuxième innovation majeure concerne le mécanisme d'estimation automatique du nombre de sources. En combinant une stratégie d'élagage progressif avec le principe de Longueur de Description Minimale, nous avons développé une approche qui libère l'utilisateur de la contrainte de spécifier a priori le nombre de composantes. Cette automatisation représente une avancée significative vers une analyse véritablement non supervisée des images de documents.

3. **Généralisation multi-domaines** : La troisième contribution, qui s'est révélée plus riche que prévu, concerne la capacité de généralisation de PRISM. Au-delà de son domaine d'application initial, la méthode a démontré son efficacité sur des données hyperspectrales de télédétection et sur des représentations profondes d'images naturelles.

5.2 Discussion et positionnement scientifique

Les résultats obtenus positionnent PRISM à l'intersection de plusieurs domaines de recherche. D'une part, la méthode s'inscrit dans la continuité des travaux sur la NMF et sur le démélange spectral, apportant une solution qui unifie les approches classiques des méthodes d'apprentissage profond. D'autre part, l'application réussie aux représentations de transformers pré-entraînés permet un lien entre les méthodes plus classiques et la vision par ordinateur moderne.

Cette approche arrive à un moment où l'interprétabilité redevient une priorité dans la recherche en intelligence artificielle. Les architectures d'auto-encodeurs connaissent un regain d'intérêt pour leur capacité à produire des représentations compactes et significatives. Cela reflète une évolution plus large vers des approches génériques et adaptatives en traitement du signal. Alors que les méthodes traditionnelles étaient souvent conçues pour des types de données spécifiques, ce travail démontre qu'il est possible de développer des architectures suffisamment flexibles pour s'adapter à des contextes variés tout en maintenant des garanties d'interprétabilité.

5.3 Limites et considérations pratiques

Malgré ces avancées, plusieurs limites doivent être soulignées. Premièrement, bien que le mécanisme de sélection automatique du rang fonctionne de manière satisfaisante, il reste perfectible pour les cas les plus complexes où des sources spectralement proches mais sémantiquement distinctes coexistent. Dans ces situations, une validation experte demeure précieuse pour affiner les résultats. Deuxièmement, bien que les cartes d'abondance et les

signatures spectrales produites par PRISM soient physiquement cohérentes, leur interprétation en termes de matériaux spécifiques (*e.g.*, types d'encres, dégradations) requiert des connaissances en sciences des matériaux et en conservation afin de s'assurer de la véracité. Troisièmement, bien que PRISM présente une forte capacité de généralisation sans adaptation, il accuse une legère baisse de performance en F-score lors du passage d'un cube multispectral à une représentation RVB, réduisant la capacité de discrimination fine des matériaux. Pour compenser cette perte d'information spectrale, plusieurs améliorations peuvent être envisagées, comme un pré-entraînement sur des images RVB comme proposé par la méthode gagnante du concours DIBCO, un ajustement du module d'attention pour mieux exploiter l'information spatiale disponible, ou encore l'augmentation des données via l'ajout d'autres espaces de représentation (*e.g.*, HSL, LAB) pour simuler l'information spectrale manquante. Quatrièmement, le passage à l'échelle pour des collections massives de documents reste un défi. Bien que PRISM soit plus efficace que de nombreuses méthodes concurrentes, le traitement de milliers d'images multispectrales haute résolution nécessite encore des ressources computationnelles significatives.

5.4 Perspectives futures

Les développements récents en intelligence artificielle, notamment l'émergence de modèles fondationnaux comme Spectral-GPT (Hong *et al.*, 2024) ou AlphaEarth Foundations (Brown *et al.*, 2025) par Google DeepMind, ouvrent des perspectives fascinantes pour l'analyse HS. L'intégration de modèles comme PRISM dans un cadre plus large, où un modèle pré-entraîné sur des millions d'images hyperspectrales pourrait être affiné spécifiquement pour des besoins spécifiques représente une direction de recherche prometteuse. Cette approche pourrait considérablement améliorer la généralisation tout en réduisant les besoins en données annotées.

Au-delà de la décomposition, l'intégration de modules d'identification et de localisation automatique des matériaux constitue une extension naturelle de nos travaux. Si le problème de passage à l'échelle est résolu, en combinant PRISM avec des modèles de langage multimodaux,

il devient envisageable de développer des systèmes capables non seulement de décomposer les images, mais aussi de générer automatiquement des rapports d'analyse compréhensibles par les conservateurs et les historiens. Cette approche de type «Visual Question Answering» pour les documents historiques pourrait révolutionner l'accès et l'exploitation des collections numérisées.

La capacité démontrée de PRISM à traiter les représentations de transformers suggère des applications bien au-delà du patrimoine documentaire. En médecine, l'analyse d'images multibandes pourrait bénéficier de cette approche pour la création automatique d'annotations. En agriculture, la décomposition d'images satellitaires pourrait permettre un suivi des terrains et des cultures. Plus généralement, toute application nécessitant une décomposition interprétable de données multibandes pourrait potentiellement bénéficier de l'architecture proposée.

Ce mémoire a donc présenté PRISM, une nouvelle approche de NMF profonde pour la décomposition non supervisée d'images multibandes qui combine performance et interprétabilité. Les contributions techniques avec une architecture hybride, la sélection dynamique du rang, et la généralisation multi-domaines, constituent des avancées pour le domaine. Plus important encore, la flexibilité et la robustesse démontrées ouvrent la voie à différentes applications futures, suggérant que les principes développés ici pourraient avoir un impact bien au-delà de l'analyse de documents. Alors que nous entrons dans une époque où l'intelligence artificielle joue un rôle croissant dans de nombreux domaines, des approches comme PRISM, qui maintiennent une interprétabilité tout en automatisant des tâches complexes, représentent un équilibre prometteur entre innovation technologique et préservation du contrôle expert sur les processus d'analyse.

L'ensemble des contributions présentées dans ce mémoire ont fait l'objet d'une publication acceptée au workshop Vision Docs de la conférence ICCV en octobre 2025 :
Declercq, Rahiche & Cheriet (2025). **PRISM : Pruning for Rank-adaptive Interpretable Segmentation Model with Application to Historical Document Multiband Images.** Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, ICCV 2025.

ANNEXE I

L'IMAGERIE MULTIBANDE : PRINCIPES ET APPLICATIONS

La capture d'une image par un appareil photo numérique moderne, qu'il soit simple ou intégré à des systèmes d'imagerie avancés, repose sur des principes physiques et mathématiques fondamentaux. Comprendre son fonctionnement est une première étape essentielle pour aborder l'imagerie multibande pour l'analyse de document historiques.

1. L'appareil photo comme système linéaire non négatif

L'appareil photo capture la lumière émise ou réfléchie par une scène grâce à son objectif, qui projette une image de cette dernière sur un capteur photosensible. Ce capteur est une mosaïque de millions de cellules appelés pixels, chacun disposant de photosite, convertissant les photons incidents en une charge électrique. La charge électrique accumulée par chaque photodiode est ensuite échantillonnée et quantifiée par un convertisseur analogique-numérique (CAN), produisant une valeur numérique discrète proportionnelle à l'intensité lumineuse pour cet emplacement. Ces valeurs brutes sont ensuite traitées par le processeur de la caméra (balance des blancs, corrections gamma, etc.) avant d'être enregistrées. La Figure I-1 illustre ce processus.

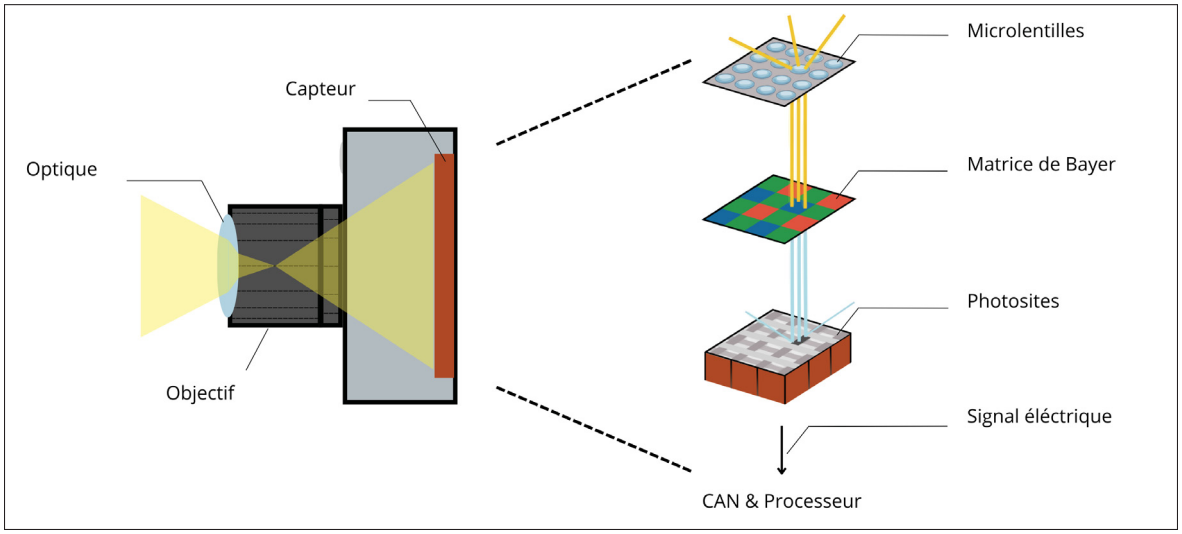

Figure-A I-1 Schéma d'un appareil photo RVB

Pour capturer la couleur, des filtres colorés (généralement rouge, vert et bleu) sont disposés au-dessus de groupes adjacents de photodiodes, selon des motifs spécifiques visant à reproduire l'appareil visuel humain. Chaque photosite mesure ainsi l'intensité lumineuse dans une bande spectrale limitée par la transmission du filtre. La réponse spectrale du capteur décrit sa sensibilité aux différentes longueurs d'onde du spectre électromagnétique. Un lien peut alors être fait entre l'intensité lumineuse physique et la mesure numérique observée. En considérant les données brutes (avant l'application de traitements non linéaires significatifs), l'appareil photo peut être modélisé comme un système linéaire. Cela signifie que la valeur numérique P_k enregistrée par le k-ième type de capteur (par exemple, le canal rouge) est approximativement proportionnelle à l'intégrale du produit de l'intensité lumineuse incidente $L(\lambda)$ et de la réponse spectrale du capteur $S_k(\lambda)$ sur l'ensemble des longueurs d'onde λ :

$$P_k \approx \int L(\lambda) S_k(\lambda) d\lambda. \tag{A I-1}$$

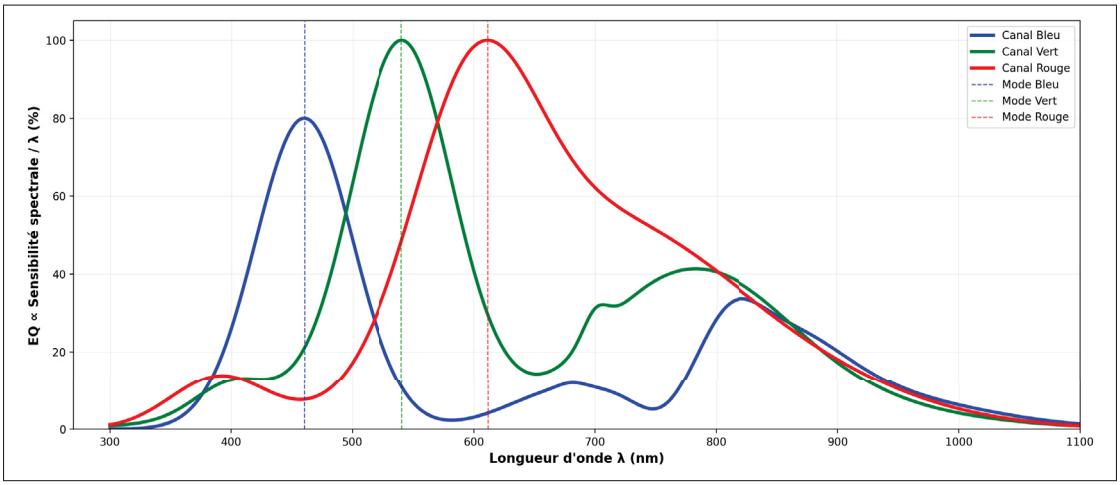

Figure-A I-2 Efficacité quantique (EQ) du capteur RVB CMOS Canon Rebel EOS. La sensibilité spectrale relative est directement liée à l'efficacité quantique [7]. Les lignes pointillées verticales marquent les pics de sensibilité (modes) pour chaque canal colorimétrique. Graphique adapté de Pan *et al.* (2021)

Enfin, il est important de souligner que ce système présente une caractéristique fondamentale : **la non-négativité**. En effet, l'intensité lumineuse, qu'elle soit mesurée en termes de radiance, de

réflectance ou de quantité de photons, ne peut physiquement pas être négative ; l'absence totale de photons représente la valeur minimale d'intensité, correspondant à zéro. Par conséquent, les valeurs numériques enregistrées par le capteur, qui représentent ces intensités, doivent également être non-négatives (supérieures ou égales à zéro). Cette propriété physique peut alors être exploitée comme une contrainte avantageuse dans le cadre du traitement ultérieur d'images. Cela permet notamment d'optimiser des algorithmes et de garantir l'interprétatibilité physique des résultats lors d'opérations de filtrage, de reconstruction ou même de restauration d'images.

2. Couleurs et spectre éléctromagnétique

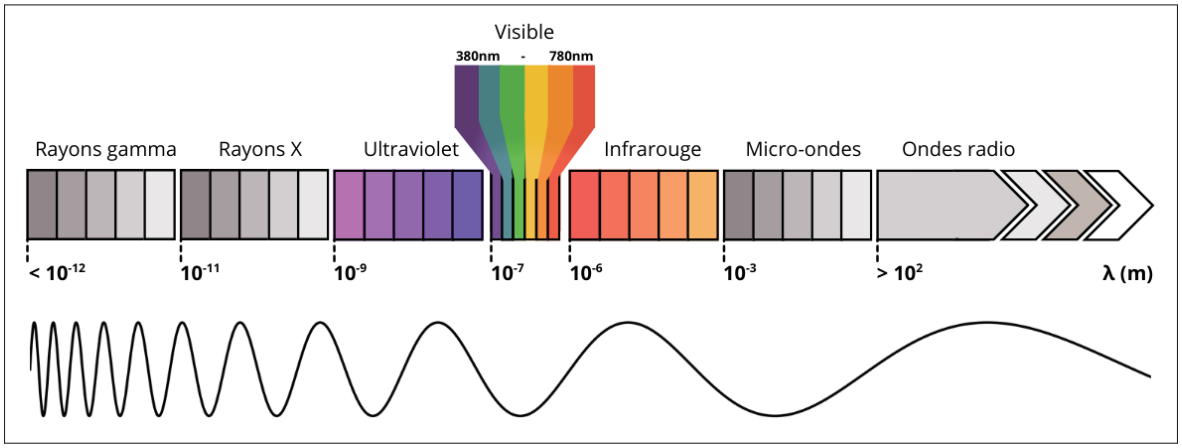

Figure-A I-3 Visualisation du spectre électromagnétique. Adapté d'une image de l'Agence spatiale canadienne (2021)

La perception des couleurs est une expérience humaine fondamentale, mais elle ne représente qu'une infime partie d'un vaste continuum d'énergies rayonnantes : le spectre électromagnétique, illustré sur la Figure I-3. La richesse des couleurs que nous percevons n'est qu'une interprétation par notre système visuel d'une petite fraction de ce spectre. Notre capacité à voir les couleurs est due à la présence dans la rétine de l'œil de trois types de cellules photoréceptrices appelées cônes, chacun sensible à une des principales gammes de longueurs d'onde du spectre visible (Rouge, Vert et Bleu). Le cerveau reçoit les signaux électriques émis par ces trois types de cônes et les combine pour produire la sensation d'une vaste palette de couleurs. Si les trois types

[7] Rapport nombre d'électrons générés par photon incident (Rogalski, Adamiec & Rutkowski, 2000)

de cônes sont stimulés de manière égale, nous percevons la lumière blanche. Il est important de noter que les humains ne sont pas sensibles aux couleurs de manière égale, notre vision étant particulièrement réceptive aux longueurs d'onde vertes-jaunes (autour de 555 nm) et moins sensible aux extrémités du spectre visible. Cette caractéristique biologique explique pourquoi les motifs de filtres pour les appareils numériques RVB, comme celui de Bayer, utilisent généralement deux photodiodes vertes pour une rouge et une bleue (*voir* Figure I-1). Les capteurs d'imagerie multibande, contrairement à l'œil humain, peuvent être conçus pour être sensibles à des longueurs d'onde bien au-delà du spectre visible, notamment dans l'ultraviolet et le proche infrarouge, voire l'infrarouge moyen et thermique. Cette capacité à ***voir l'invisible*** ouvre des perspectives considérables pour l'analyse des matériaux et des phénomènes, car de nombreuses substances possèdent des caractéristiques spectrales distinctives en dehors du domaine visible. L'apparence des objets, et en particulier leur couleur, est déterminée par l'interaction entre la lumière et la matière. Lorsqu'un rayonnement électromagnétique incident atteint un matériau, trois phénomènes principaux peuvent se produire :

- **L'absorption** (α) **:** L'énergie du rayonnement est absorbée par le matériau, souvent transformée en chaleur. La sélectivité de l'absorption à certaines longueurs d'onde est une cause majeure de la couleur des objets. Par exemple, un objet apparaît rouge parce qu'il absorbe les longueurs d'onde bleues et vertes tout en réfléchissant les rouges.
- **La réflexion** (ρ) **:** Le rayonnement est renvoyé par la surface du matériau. Elle peut être spéculaire (comme un miroir, avec un angle de réflexion égal à l'angle d'incidence) ou diffuse (la lumière est dispersée dans de multiples directions, typique des surfaces mates).
- **La transmission** (τ) **:** Le rayonnement traverse le matériau. Si le matériau est transparent ou translucide, la lumière peut le traverser, subissant potentiellement une réfraction.

Pour une longueur d'onde donnée, la somme de ces trois proportions est égale à un,

$$\alpha + \rho + \tau = 1, \tag{A I-2}$$

conformément au principe de conservation de l'énergie. Ces interactions sont fondamentales car elles sont à la base de la formation des signatures spectrales.

3. La notion de signature spectrale et son importance

La notion de signature spectrale est fondamentale dans divers domaines scientifiques et techniques, car elle permet d'identifier et de distinguer les matériaux en fonction de leur interaction unique avec le rayonnement électromagnétique à différentes longueurs d'onde. Une signature spectrale est définie comme le modèle de réflexion, d'absorption ou d'émission d'un matériau en fonction de la longueur d'onde, qui est caractéristique de sa composition et de sa structure. Ce concept est essentiel car il permet de reconnaître des matériaux spécifiques en comparant leur signature à des références connues. Dans le domaine de la télédétection, par exemple, les signatures spectrales sont utilisées pour analyser des images satellites et identifier des types de végétation, de sols ou de minéraux, aidant à des applications comme la surveillance environnementale ou l'estimation des rendements agricoles. Dans l'analyse d'œuvres d'art, la spectroscopie, qui repose sur les signatures spectrales, est employée pour identifier les pigments et les liants utilisés, ce qui est crucial pour l'authentification, la conservation et la compréhension des techniques artistiques. La signature spectrale est un outil clé pour la caractérisation des matériaux dans des contextes variés, alliant précision et non-invasivité, point essentiel pour l'étude de documents fragiles.

4. Images multispectrales (MS) contre hyperspectrales (HS)

L'imagerie multibandes exploite la manière dont les matériaux interagissent avec les diverses régions du spectre électromagnétique. En addition aux images traditionnelles RVB, on distingue principalement deux autres types d'images multibandes : les images **multispectrales (MS)** et les images **hyperspectrales (HS)**. La Figure 0.3 illustre un exemple de chacun de ces deux types.

Une image **multispectrale** (MS) est acquise en enregistrant l'information lumineuse dans un nombre limité de bandes spectrales, typiquement entre 3 à 15 bandes. Ces bandes sont généralement larges (*i.e.*, plusieurs dizaines de nanomètres de largeur) et peuvent être espacées à travers le spectre (p. ex., une bande dans le bleu, une dans le vert, une dans le rouge, et quelques-unes dans le proche infrarouge). Elles ne fournissent donc pas une couverture continue du spectre, mais permettent d'augmenter l'information disponible sur une scène observée.

Une image **hyperspectrale** (HS), en revanche, capture l'information lumineuse dans un plus grand nombre de bandes spectrales, allant de plusieurs dizaines à des centaines, voire même des milliers. La caractéristique essentielle de ces bandes est leur faible épaisseur (souvent de l'ordre de quelques nanomètres à une dizaine de nanomètres de largeur au maximum) ainsi que leur contiguïté. Elles permettent ainsi d'obtenir un continuum de réflectance pour chaque pixel de l'image, offrant une information spectrale beaucoup plus détaillée que pour les images MS.

La résolution spectrale, qui correspond à la capacité des bandes spectrales à discerner des détails fins dans le spectre, est donc significativement plus élevée pour les images HS que pour les images MS. En matière de résolution spatiale, un compromis doit souvent être établi. La première option consiste à diviser le signal lumineux en de nombreuses bandes étroites, ce qui se traduit par une réduction de la quantité d'énergie par bande. Cette approche nécessite alors l'adoption d'une résolution spatiale plus grossière afin de maintenir un bon rapport signal/bruit (HS). À l'inverse, la seconde option consiste à réduire le nombre de bandes et à les rendre plus larges, permettant d'obtenir une résolution spatiale plus fine (MS). L'imagerie HS excelle alors dans l'observation terrestre, notamment par le biais de la photographie aérienne et de la télédétection. Grâce à la capture de centaines de bandes spectrales contiguës, les caméras HS équipées sur satellites permettent de distinguer des matériaux aux signatures spectrales presque identiques, essentiel pour l'étude des ressources naturelles et le suivi climatique à grande échelle. La couverture de vastes territoires justifie le sacrifice en résolution spatiale, susceptible de générer des pixels présentant des mélanges de matériaux. Cela est compensé par la richesse spectrale permettant de caractériser en détail ces zones hétérogènes et par la moindre pertinence des détails spatiaux fins (*e.g.*, feuilles ou brins d'herbe individuels) dans l'observation terrestre. D'un autre côté, l'imagerie MS est particulièrement utile pour l'analyse de documents, permettant de distinguer encres et papiers ou de révéler des altérations grâce à quelques bandes spectrales clés avec des images de hautes résolution. Cette haute résolution spatiale est cruciale dans un contexte où chaque pixel compte, que ce soit pour la détection de détails fins comme les traits de plume ou même pour observer en détails des altérations subtiles sur des documents.

ANNEXE II

DÉVELOPPEMENT MATHÉMATIQUE DE LA NMF

1. Fonctions de Coût

L'approximation dans l'équation 1.3 est obtenue en minimisant une fonction de coût qui mesure la divergence ou l'erreur entre la matrice originale **Y** et sa reconstruction **UV**. Les deux fonctions de coût les plus couramment utilisées sont :

1. **L'erreur quadratique** (basée sur la norme de Frobenius) :

$$\min_{\mathbf{U},\mathbf{V}\geq 0} ||\mathbf{Y} - \mathbf{UV}||_F^2 = \min_{\mathbf{U},\mathbf{V}\geq 0} \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - (UV)_{ij})^2 \qquad \text{(A II-1)}$$

   Cette fonction de coût est souvent choisie lorsque le bruit dans les données est supposé être additif et gaussien.

2. **La divergence de Kullback-Leibler** (ou entropie relative généralisée) :

$$\min_{\mathbf{U},\mathbf{V}\geq 0} D_{KL}(\mathbf{Y}||\mathbf{UV}) = \min_{\mathbf{U},\mathbf{V}\geq 0} \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij}\log\frac{Y_{ij}}{(UV)ij} - Yij + (UV)_{ij}\right) \qquad \text{(A II-2)}$$

   Cette fonction est particulièrement adaptée lorsque les données peuvent être interprétées comme des comptages (p. ex., issues de distributions de Poisson) avec des applications pour le séquençage des génomes, l'extraction de texte ou le traitement de signaux audio.

Le choix de la fonction de coût peut influencer la nature des composantes extraites ainsi que la rapidité de convergence. La norme de Frobenius est sensible aux grandes valeurs, tandis que la divergence de Kullback-Leibler peut mieux gérer les données avec une grande plage dynamique. Pour les données de réflectance MS, le bruit peut avoir des caractéristiques complexes, et le choix optimal de la fonction de coût peut dépendre de l'application spécifique et des propriétés du bruit.

2. Algorithmes d'Optimisation

La minimisation de ces fonctions de coût sous les contraintes de non-négativité pour W et H est un problème d'optimisation non convexe, ce qui signifie qu'il peut exister de multiples minima locaux. Les algorithmes les plus répandus pour résoudre ce problème sont itératifs :

1. **Règles de mise à jour multiplicatives** (Multiplicative Update - MU) : Proposées initialement par Lee et Seung, ces règles consistent à mettre à jour alternativement **U** et **V** en utilisant des multiplications matricielles qui garantissent le maintien de la non-négativité à chaque itération. Elles sont relativement simples à implémenter et ont été prouvées comme convergeant vers un minimum local de la fonction de coût choisie (Lee & Seung, 2000).
2. **Moindres Carrés Alternés** (Alternating Least Squares - ALS) : Cette approche consiste à fixer alternativement l'une des matrices et à résoudre un problème de moindres carrés non-négatifs pour l'autre. Ce processus est répété jusqu'à convergence (Kim & Park, 2007).
3. **Méthodes de gradient projeté :** Ces méthodes utilisent des techniques de descente de gradient pour minimiser la fonction de coût, en projetant à chaque étape les solutions sur l'ensemble des matrices non-négatives (Lin, 2007).

Le principe commun est d'optimiser **U** and **V** alternativement, l'optimisation simultanée étant non convexe :

$$\min_{U\geq 0} \|\mathbf{Y} - \mathbf{UV}\|_F^2 \quad \text{et} \quad \min_{V\geq 0} \|\mathbf{Y} - \mathbf{UV}\|_F^2 \tag{A II-3}$$

Les **Mises à Jour Multiplicatives** (MU) de Lee et Seung, dérivées des conditions **Karush-Kuhn-Tucker (KKT)**, sont particulièrement populaires. Pour $D_F = \frac{1}{2}\|Y - UV\|_F^2$, les conditions KKT impliquent :

$$\mathbf{U} \geq 0, \quad \mathbf{V} \geq 0, \tag{A II-4}$$

$$\nabla_U D_F \geq 0, \quad \nabla_V D_F \geq 0, \tag{A II-5}$$

$$\mathbf{U} \odot \nabla_U D_F = 0, \quad \mathbf{V} \odot \nabla_V D_F = 0 \tag{A II-6}$$

Avec les gradients :

$$\nabla_U D_F = \mathbf{U}\mathbf{V}\mathbf{V}^T - \mathbf{Y}\mathbf{V}^T \tag{A II-7}$$

$$\nabla_V D_F = \mathbf{U}^T\mathbf{U}\mathbf{V} - \mathbf{U}^T\mathbf{Y} \tag{A II-8}$$

On obtient les règles de mise à jour (où $\odot$ est le produit Hadamard) :

$$\mathbf{U} = \mathbf{U} \odot \frac{\mathbf{Y}\mathbf{V}^T}{\mathbf{U}\mathbf{V}\mathbf{V}^T} \tag{A II-9}$$

$$\mathbf{V} = \mathbf{V} \odot \frac{\mathbf{U}^T\mathbf{Y}}{\mathbf{U}^T\mathbf{U}\mathbf{V}} \tag{A II-10}$$

L'algorithme de Lee et Seung (*voir* Algorithme II-1), consiste à appliquer ces règles itérativement jusqu'à convergence pour la décomposition $\mathbf{Y} \approx \mathbf{U}\mathbf{V}$, déterminée par un critère de tolérance ϵ sur la variation relative de la fonction de coût ou l'atteinte d'un nombre maximal d'itérations.

Algorithme-A II-1 : Mises à Jour Multiplicatives (Lee & Seung, 2000)

Input : Matrix $\mathbf{Y} \in \mathbb{R}_+^{m\times n}$, rank r

Output : Matrices $\mathbf{U} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{V} \in \mathbb{R}_+^{r\times n}$ such that $\mathbf{Y} \approx \mathbf{U}\mathbf{V}$

Initialize $\mathbf{U}^0 \geq 0$ and $\mathbf{V}^0 \geq 0$ randomly;

Set $k \leftarrow 0$;

while *not convergence* **do**

  Mise à jour de U : $\begin{cases} \text{Norme de Frobenius} : \mathbf{U}^{k+1} \leftarrow \mathbf{U}^k \odot \frac{\mathbf{Y}(\mathbf{V}^k)^T}{\mathbf{U}^k\mathbf{V}^k(\mathbf{V}^k)^T} \\ \text{Divergence KL} : \mathbf{U}^{k+1} \leftarrow \mathbf{U}^k \odot \frac{\frac{\mathbf{Y}}{\mathbf{U}^k\mathbf{V}^k}(\mathbf{V}^k)^T}{\mathbf{1}(\mathbf{V}^k)^T} \end{cases}$;

  Mise à jour de V : $\begin{cases} \text{Norme de Frobenius} : \mathbf{V}^{k+1} \leftarrow \mathbf{V}^k \odot \frac{(\mathbf{U}^{k+1})^T\mathbf{Y}}{(\mathbf{U}^{k+1})^T\mathbf{U}^{k+1}V^k} \\ \text{Divergence KL} : \mathbf{V}^{k+1} \leftarrow \mathbf{V}^k \odot \frac{(\mathbf{U}^{k+1})^T\frac{\mathbf{Y}}{\mathbf{U}^{k+1}\mathbf{V}^k}}{(\mathbf{U}^{k+1})^T\mathbf{1}} \end{cases}$;

  $k \leftarrow k + 1$;

end while

return $\mathbf{U}^k, \mathbf{V}^k$;

BIBLIOGRAPHIE

Agence spatiale canadienne. (2021). Le spectre électromagnétique. Repéré le 2025-05-22 à https://www.asc-csa.gc.ca/fra/multimedia/recherche/image/16886.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723. doi : 10.1109/TAC.1974.1100705.

Alayrac, J.-B., Carreira, J. & Zisserman, A. (2019, June). The Visual Centrifuge : Model-Free Layered Video Representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alfaro-Mejía, E., Manian, V., Ortiz, J. D. & Tokars, R. P. (2023). A blind convolutional deep autoencoder for spectral unmixing of hyperspectral images over waterbodies. *Frontiers in Earth Science*, Volume 11 - 2023, 1229704. doi : 10.3389/feart.2023.1229704.

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.

Anagnostidis, S., Pavllo, D., Biggio, L., Noci, L., Lucchi, A. & Hofmann, T. (2023). Dynamic context pruning for efficient and interpretable autoregressive transformers. *Advances in Neural Information Processing Systems*, 36, 65202–65223.

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K. & Ghayvat, H. (2021). CNN variants for computer vision : History, architecture, application, challenges and future scope. *Electronics*, 10(20), 2470.

Bourlard, H. & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4), 291–294.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A. & Kohli, P. (2025). *AlphaEarth Foundations : An embedding field model for accurate and efficient global mapping from sparse label data*. Rapport de recherche de Google DeepMind.

Calvo-Zaragoza, J. & Gallego, A.-J. (2019). A selectional auto-encoder approach for document image binarization. *Pattern Recognition*, 86, 37-47. doi : https://doi.org/10.1016/j.patcog.2018.08.011.

Cardoso, J.-F. (1998). Blind signal separation : statistical principles. *Proceedings of the IEEE*, 86(10), 2009–2025.

Charikar, M. & Hu, L. (2021). Approximation algorithms for orthogonal non-negative matrix factorization. *International Conference on Artificial Intelligence and Statistics*, pp. 2728–2736.

Chen, W.-S., Zhao, Y., Pan, B. & Chen, B. (2016). Supervised kernel nonnegative matrix factorization for face recognition. *Neurocomputing*, 205, 165–181.

Chen, W.-S., Zeng, Q. & Pan, B. (2022). A survey of deep nonnegative matrix factorization. *Neurocomputing*, 491, 305-320. doi : https://doi.org/10.1016/j.neucom.2021.08.152.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. (2022, June). Masked-Attention Mask Transformer for Universal Image Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290-1299.

Cheng, H., Zhang, M. & Shi, J. Q. (2024). A survey on deep neural network pruning : Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10558-10578. doi : 10.1109/TPAMI.2024.3447085.

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.

Choi, S., Cichocki, A., Park, H.-M. & Lee, S.-Y. (2005). Blind source separation and independent component analysis : A review. *Neural Information Processing-Letters and Reviews*, 6(1), 1–57.

Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S. (2009). *Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation* (éd. 1). Wiley. doi : 10.1002/9780470747278.

Davies, H. & Zawacki, A. J. (2019). Making Light Work : Manuscripts and Multispectral Imaging. *Journal of the Early Book Society*, 22, 186.

Debain, Y. (2020). *Deep Convolutional Nonnegative Autoencoders*. (Thèse de doctorat, KTH ROYAL INSTITUTE OF TECHNOLOGY, Stockholm).

Declercq, K., Rahiche, A. & Cheriet, M. (2025, October). PRISM : Pruning for Rank-adaptive Interpretable Segmentation Model with Application to Historical Document Multiband Images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 7526-7535.

Diem, M., Hollaus, F. & Sablatnig, R. (2016). Msio : Multispectral document image binarization. *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.

Ding, C. H., Li, T. & Jordan, M. I. (2008). Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1), 45–55.

Dou, Z., Gao, K., Zhang, X., Wang, H. & Wang, J. (2020). Hyperspectral Unmixing Using Orthogonal Sparse Prior-Based Autoencoder With Hyper-Laplacian Loss and Data-Driven Outlier Detection. *IEEE TGRS*, 58(9), 6550–6564. doi : 10.1109/TGRS.2020.2977819.

Duh, J., Krstić, D., Desnica, V. & Fazinić, S. (2018). Non-destructive study of iron gall inks in manuscripts. *Nuclear Instruments and Methods in Physics Research Section B : Beam Interactions with Materials and Atoms*, 417, 96–99. doi : 10.1016/j.nimb.2017.08.033.

Easton, R., Knox, K. & Christens-Barry, W. (2003). Multispectral imaging of the archimedes palimpsest. *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.*, pp. 111–116. doi : 10.1109/AIPR.2003.1284258.

Easton, R. L., Knox, K. T., Christens-Barry, W. A. & Boydston, K. (2018). Spectral Imaging Methods Applied to the Syriac Galen Palimpsest. *Manuscript Studies : A Journal of the Schoenberg Institute for Manuscript Studies*, 3(1), 69–82. doi : 10.1353/mns.2018.0003.

Elkerdawy, S., Elhoushi, M., Zhang, H. & Ray, N. (2022). Fire together wire together : A dynamic pruning approach with self-supervised mask prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12454–12463.

Febrissy, M. & Nadif, M. (2020). A consensus approach to improve nmf document clustering. *Advances in Intelligent Data Analysis XVIII : 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18*, pp. 171–183.

Flenner, J. & Hunter, B. (2017). A deep non-negative matrix factorization neural network. *Semantic Scholar*.

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. (2020). *Nonnegative matrix factorization*. SIAM.

Green, D. & Bailey, S. (2024). Algorithms for Non-Negative Matrix Factorization on Noisy Data With Negative Values. *IEEE Transactions on Signal Processing*, 72, 5187-5197. doi : 10.1109/TSP.2024.3474530.

Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M. & Hu, S.-M. (2023). Visual attention network. *Computational Visual Media*, 9(4), 733–752. doi : 10.1007/s41095-023-0364-2.

He, W., Wu, M., Liang, M. & Lam, S.-K. (2021, January). CAP : Context-Aware Pruning for Semantic Segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 960-969.

He, Z., Xie, S., Zdunek, R., Zhou, G. & Cichocki, A. (2011). Symmetric nonnegative matrix factorization : Algorithms and applications to probabilistic clustering. *IEEE Transactions on Neural Networks*, 22(12), 2117–2131.

Hedjam, R. & Cheriet, M. (2013). Historical document image restoration using multispectral imaging system. *Pattern Recognition*, 46(8), 2297–2312. doi : https://doi.org/10.1016/j.patcog.2012.12.015.

Hedjam, R., Nafchi, H. Z., Moghaddam, R. F., Kalacska, M. & Cheriet, M. (2015). ICDAR 2015 contest on MultiSpectral Text Extraction (MS-TEx 2015). *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1181-1185. doi : 10.1109/ICDAR.2015.7333947.

Hinrich, J. L. & Mørup, M. (2018). Probabilistic sparse non-negative matrix factorization. *International Conference on Latent Variable Analysis and Signal Separation*, pp. 488–498.

Hollaus, F., Diem, M., Fiel, S., Kleber, F. & Sablatnig, R. (2015a). Investigation of ancient manuscripts based on multispectral imaging. *Proceedings of the 2015 ACM Symposium on Document Engineering*, pp. 93–96.

Hollaus, F., Diem, M. & Sablatnig, R. (2015b). Binarization of multispectral document images. *Computer Analysis of Images and Patterns : 16th International Conference, CAIP 2015*, pp. 109–120.

Hollaus, F., Diem, M. & Sablatnig, R. (2018). 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. doi : 10.1109/ICDAR.2019.00091.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., Plaza, A., Gamba, P., Benediktsson, J. A. & Chanussot, J. (2024). SpectralGPT : Spectral Remote Sensing Foundation Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8), 5227-5244. doi : 10.1109/TPAMI.2024.3362475.

Hornik, K., Feinerer, I., Kober, M. & Buchta, C. (2012). Spherical k-means clustering. *Journal of statistical software*, 50, 1–22.

Howe, N. R. (2013). Document binarization with automatic parameter tuning. *International journal on document analysis and recognition (IJDAR)*, 16, 247–258.

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.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3), 626–634.

Ito, Y., Oeda, S.-i. & Yamanishi, K. (2016). Rank selection for non-negative matrix factorization with normalized maximum likelihood coding. *Proceedings Of The 2016 SIAM international conference on data mining*, pp. 720–728.

Jones, C., Duffy, C., Gibson, A. & Terras, M. (2020). Understanding multispectral imaging of cultural heritage : Determining best practice in MSI analysis of historical artefacts. *Journal of Cultural Heritage*, 45, 339–350. doi : https://doi.org/10.1016/j.culher.2020.03.004.

Joo Kim, S., Deng, F. & Brown, M. S. (2011). Visual enhancement of old documents with hyperspectral imaging. *Pattern Recognition*, 44(7), 1461-1469. doi : https://doi.org/10.1016/j.patcog.2010.12.019.

Kaarna, A., Zemcik, P., Kalviainen, H. & Parkkinen, J. (2002). Compression of multispectral remote sensing images using clustering and spectral reduction. *IEEE transactions on Geoscience and Remote Sensing*, 38(2), 1073–1082.

Karimijafarbigloo, S., Azad, R., Kazerouni, A. & Merhof, D. (2024, 10–12 Jul). MS-Former : Multi-Scale Self-Guided Transformer for Medical Image Segmentation. *Medical Imaging with Deep Learning*, 227(Proceedings of Machine Learning Research), 680–694. Repéré à https://proceedings.mlr.press/v227/karimijafarbigloo24a.html.

Kim, H. & Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12), 1495–1502.

Kingma, D. P. & Ba, J. (2017). *Adam : A Method for Stochastic Optimization*. Manuscrit publié sur arXiv. Repéré à https://arxiv.org/abs/1412.6980.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P. & Girshick, R. (2023, October). Segment Anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015-4026.

Knox, K. T., Easton Jr, R. L., Christens-Barry, W. A. & Boydston, K. (2011). Recovery of handwritten text from the diaries and papers of David Livingstone. *Computer Vision and Image Analysis of Art II*, 7869, 68–74. doi : 10.1117/12.877135.

Kong, D., Ding, C. & Huang, H. (2011). Robust nonnegative matrix factorization using l21-norm. *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 673–682.

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.

Kulkarni, S., Madurwar, R., Narlawar, R., Pandya, A. & Gawande, N. (2023). Digitization of Physical Notes : A Comprehensive Approach Using OCR, CNN, RNN, and NMF. *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pp. 1–5.

Kögel, P. R. (1920). *Die Palimpsestphotographie* (Traduit *Palimpsest Photography (Photography of Erased Writings) in its Scientific Foundations and Practical Applications)* (éd. anglaise : http://www.cis.rit.edu/~rlepci/palimpsest_imaging/Kogel_1920_English+Original_Weimar.pdf). Halle (Saale) : Knapp. Repéré à https://nbn-resolving.org/urn:nbn:de:gbv:wim2-g-2965569.

Le Roux, J., Weninger, F. J. & Hershey, J. R. (2015). Sparse NMF–half-baked or well done ? *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 11, 13–15.

Lee, D. & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13. Repéré à https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf.

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755), 788–791.

Lee, H., Cichocki, A. & Choi, S. (2009). Kernel nonnegative matrix factorization for spectral EEG feature extraction. *Neurocomputing*, 72(13-15), 3182–3190.

Leão, A. C., Costa, A. O., Almada, M., Zanibone, R., Souza, L. A. C., Costa, A. O., Almada, M., Zanibone, R. & Souza, L. A. C. (2024). Multispectral Imaging as a Preservation and Valuation Tool at the Minas Gerais Public Archive, Brazil : A Case Study on an 18th-century Illuminated Manuscript. *Archiving Conference*, 21, 28–32. doi : 10.2352/issn.2168-3204.2024.21.1.6. Publisher : Society for Imaging Science and Technology.

Li, C., Cai, R. & Yu, J. (2023). An attention-based 3D convolutional autoencoder for few-shot hyperspectral unmixing and classification. *Remote Sensing*, 15(2), 451.

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.

Licciardi, G. A. & Del Frate, F. (2011). Pixel Unmixing in Hyperspectral Data by Means of Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11), 4163-4172. doi : 10.1109/TGRS.2011.2160950.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10), 2756–2779.

Liu, Q., Zhou, F., Hang, R. & Yuan, X. (2017). Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. *Remote Sensing*, 9(12), 1330.

López-Baldomero, A. B., Martínez-Domingo, M. Á., Hernández-Andrés, J., Blanc, R., Vilchez-Quero, J., López-Montes, A., Valero, E. M. et al. (2023). Endmember Extraction for Pigment Identification Pre-and Post-intervention : A Case Study from a XVIth Century Copper Plate Painting. *Archiving Conference*. doi : 10.2352/issn.2168-3204.2023.20.1.40.

López-Baldomero, A. B., Buzzelli, M., Moronta-Montero, F., Martínez-Domingo, M. Á. & Valero, E. M. (2025). Ink classification in historical documents using hyperspectral imaging and machine learning methods. *Spectrochimica Acta Part A : Molecular and Biomolecular Spectroscopy*, 335, 125916.

Lyu, S., Liu, Y., Hou, M., Yin, Q., Wu, W. & Yang, X. (2020). Quantitative analysis of mixed pigments for Chinese paintings using the improved method of ratio spectra derivative spectrophotometry based on mode. *Heritage Science*, 8, 1–21.

López-Baldomero, A. B., Martínez-Domingo, M. A., Hernández-Andrés, J., Blanc, R., Vilchez-Quero, J. L., López-Montes, A. & Valero, E. M. (2023). Endmember Extraction for Pigment Identification Pre- and Post-intervention : A Case Study from a XVIth Century Copper Plate Painting. *Archiving Conference*, 20(1), 198–203. doi : 10.2352/issn.2168-3204.2023.20.1.40.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, 5, 281–298.

Magkanas, G., Bagán, H., Sistach, M. & García, J. (2021). Illuminated manuscript analysis methodology using MA-XRF and NMF : Application on the Liber Feudorum Maior. *Microchemical Journal*, 165, 106112.

Mathew, M., Karatzas, D. & Jawahar, C. (2021, January). DocVQA : A Dataset for VQA on Document Images. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200-2209.

Mazack, M. (2009). Non-negative Matrix Factorization with Applications to Handwritten Digit Recognition. *Department of Scientific Computation, University of Minnesota*.

Mazzocato, S., Cimino, D. & Daffara, C. (2024). Integrated microprofilometry and multispectral imaging for full-field analysis of ancient manuscripts. *Journal of Cultural Heritage*, 66, 110–116. doi : https://doi.org/10.1016/j.culher.2023.11.014.

Mei, S., Ji, J., Geng, Y., Zhang, Z., Li, X. & Du, Q. (2019). Unsupervised Spatial–Spectral Feature Learning by 3D Convolutional Autoencoder for Hyperspectral Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9), 6808-6820. doi : 10.1109/TGRS.2019.2908756.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 72(4), 417–473.

Melo, M. J., Otero, V., Nabais, P., Teixeira, N., Pina, F., Casanova, C., Fragoso, S. & Sequeira, S. O. (2022). Iron-gall inks : a review of their degradation mechanisms and conservation treatments. *Heritage Science*, 10(1), 1–11. doi : 10.1186/s40494-022-00779-2. Publisher : Nature Publishing Group.

Merrikh-Bayat, F., Babaie-Zadeh, M. & Jutten, C. (2010). Using non-negative matrix factorization for removing show-through. *Latent Variable Analysis and Signal Separation : 9th International Conference, LVA/ICA 2010, St. Malo, France, September 27-30, 2010. Proceedings 9*, pp. 482–489.

Naik, G. R., Wang, W. et al. (2014). Blind source separation. *Berlin : Springer*, 10, 978–3.

Nascimento, J. & Dias, J. (2005). Vertex component analysis : a fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4), 898-910. doi : 10.1109/TGRS.2005.844293.

Neri, J., Badeau, R. & Depalle, P. (2021). Unsupervised Blind Source Separation with Variational Auto-Encoders. *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 311-315. doi : 10.23919/EUSIPCO54536.2021.9616154.

Ng, A., Ngiam, J., Foo, C. Y., Mai, Y., Suen, C., Coates, A., Maas, A., Hannun, A., Huval, B., Wang, T. & Tandon, S. (2014). Unsupervised Feature Learning and Deep Learning. [Stanford Deep Learning Tutorial]. Repéré à http://deeplearning.stanford.edu/tutorial/.

Ngoc-Diep, H. (2008). *NONNEGATIVE MATRIX FACTORIZATION ALGORITHMS AND APPLICATIONS*. (Thèse de doctorat, École Polytechnique de Louvain).

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., oIshan Misra, Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A. & Bojanowski, P. (2024). DINOv2 : Learning Robust Visual Features without Supervision. Repéré à https://arxiv.org/abs/2304.07193.

Ozkan, S., Kaya, B. & Akar, G. B. (2019). EndNet : Sparse AutoEncoder Network for Endmember Extraction and Hyperspectral Unmixing. *IEEE TGRS*, 57(1), 482–496. doi : 10.1109/TGRS.2018.2856929.

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.

Palsson, B., Sigurdsson, J., Sveinsson, J. R. & Ulfarsson, M. O. (2018). Hyperspectral Unmixing Using a Neural Network Autoencoder. *IEEE Access*, 6, 25646-25656. doi : 10.1109/ACCESS.2018.2818280.

Palsson, B., Sveinsson, J. R. & Ulfarsson, M. O. (2019). Multitask Learning for Spatial-Spectral Hyperspectral Unmixing. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 564-567. doi : 10.1109/IGARSS.2019.8900229.

Palsson, B., Ulfarsson, M. O. & Sveinsson, J. R. (2021). Convolutional Autoencoder for Spectral–Spatial Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1), 535-549. doi : 10.1109/TGRS.2020.2992743.

Palsson, B., Sveinsson, J. R. & Ulfarsson, M. O. (2022). Blind Hyperspectral Unmixing Using Autoencoders : A Critical Comparison. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 1340-1372. doi : 10.1109/JSTARS.2021.3140154.

Pan, J., Libera, J. A., Paulson, N. H. & Stan, M. (2021). Flame stability analysis of flame spray pyrolysis by artificial intelligence. *The International Journal of Advanced Manufacturing Technology*, 114(7), 2215–2228. doi : 10.1007/s00170-021-06884-z.

Park, J.-H., Kim, Y., Kim, J., Choi, J.-Y. & Lee, S. (2023). Dynamic structure pruning for compressing cnns. *Proceedings of the AAAI conference on artificial intelligence*, 37(8), 9408–9416.

Parker, C. S., Parsons, S., Bandy, J., Chapman, C., Coppens, F. & Seales, W. B. (2019). From invisibility to readability : recovering the ink of Herculaneum. *PloS one*, 14(5), e0215775.

Patel, K. (2019). MNIST Handwritten Digits Classification using a Convolutional Neural Network (CNN). Repéré le 2024-06-01 à https://medium.com/data-science/mnist-handwritten-digits-classification-using-a-convolutional-neural-network-cnn-af5fafbc35e9.

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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Peharz, R. & Pernkopf, F. (2012). Sparse nonnegative matrix factorization with $l0$-constraints. *Neurocomputing*, 80, 38–46.

Phon-Amnuaisuk, S. (2013). Applying non-negative matrix factorization to classify superimposed handwritten digits. *Procedia computer science*, 24, 261–267.

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. doi : https://doi.org/10.1016/j.neucom.2014.02.018.

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. (2021). Blind decomposition of multispectral document images using orthogonal nonnegative matrix factorization. *IEEE Transactions on Image Processing*, 30, 5997–6012.

Rahiche, A. & Cheriet, M. (2022). Variational bayesian orthogonal nonnegative matrix factorization over the stiefel manifold. *IEEE Transactions on Image Processing*, 31, 5543–5558.

Rahiche, A., Bakhta, A. & Cheriet, M. (2019). Blind source separation based framework for multispectral document images binarization. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1476–1481.

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. doi : https://doi.org/10.1016/j.culher.2020.01.012.

Renard, N., Bourennane, S. & Blanc-Talon, J. (2008). Denoising and dimensionality reduction using multilinear tools for hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 5(2), 138–142.

Reynolds, D. A. et al. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 3.

Rhodes, A., Jiang, B. & Jiang, J. (2025). Graph Regularized Sparse L2, 1 Semi-Nonnegative Matrix Factorization for Data Reduction. *Numerical Linear Algebra with Applications*, 32(1), e2598.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471. doi : https://doi.org/10.1016/0005-1098(78)90005-5.

Rodarmel, C. & Shan, J. (2002). Principal component analysis for hyperspectral image classification. *Surveying and Land Information Science*, 62(2), 115–122.

Rogalski, A., Adamiec, K. & Rutkowski, J. (2000). *Narrow-gap semiconductor photodiodes*. Bellingham, Wash : SPIE Press.

Rossi, C., Zoleo, A., Bertoncello, R., Meneghetti, M. & Deiana, R. (2021). Application of Multispectral Imaging and Portable Spectroscopic Instruments to the Analysis of an Ancient Persian Illuminated Manuscript. *Sensors (Basel, Switzerland)*, 21(15), 4998. doi : 10.3390/s21154998.

Roux, J., Cheveigné, A. & Parra, L. (2008). Adaptive template matching with shift-invariant semi-NMF. *Advances in neural information processing systems*, 21.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. et al. (1985). Learning internal representations by error propagation. Institute for Cognitive Science, University of California, San Diego La . . ..

Salehani, Y. E. & Gazor, S. (2017). Smooth and sparse regularization for NMF hyperspectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8), 3677–3692.

Salehani, Y. E., Arabnejad, E., Rahiche, A., Bakhta, A. & Cheriet, M. (2020). MSdB-NMF : MultiSpectral document image binarization framework via non-negative matrix factorization approach. *IEEE Transactions on Image Processing*, 29, 9099–9112.

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), 79–87.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461–464. Repéré à http://www.jstor.org/stable/2958889.

Shao, L., Zuo, H., Zhang, J., Xu, Z., Yao, J., Wang, Z. & Li, H. (2021). Filter Pruning via Measuring Feature Map Information. *Sensors*, 21(19), 6601. doi : 10.3390/s21196601.

Shor, P., Manfredi, M., Bearman, G. H., Marengo, E., Boydston, K. & Christens-Barry, W. A. (2014). The Leon Levy Dead Sea Scrolls Digital Library : The Digitization Project of the Dead Sea Scrolls. *Journal of Eastern Mediterranean Archaeology and Heritage Studies*, 2(2), 71–89. Repéré à https://muse.jhu.edu/pub/2/article/548044. Publisher : Penn State University Press.

Soukup, D. & Bajla, I. (2008). Robust object recognition under partial occlusions using NMF. *Computational Intelligence and Neuroscience*, 2008(1), 857453.

Squires, S., Prügel-Bennett, A. & Niranjan, M. (2017). Rank Selection in Nonnegative Matrix Factorization using Minimum Description Length. *Neural Computation*, 29(8), 2164–2176. doi : 10.1162/NECO_a_00980.

Squires, S. E. (2019). *NON-NEGATIVE MATRIX FACTORISATION : ALGORITHMS AND APPLICATIONS*. (Thèse de doctorat, University of Southampton).

Su, L., Liu, J., Yuan, Y. & Chen, Q. (2023). A Multi-Attention Autoencoder for Hyperspectral Unmixing Based on the Extended Linear Mixing Model. *Remote Sensing*, 15(11), 2898. doi : 10.3390/rs15112898.

Su, Y., Su, Y., Li, J., Li, J., Plaza, A., Plaza, A., Marinoni, A., Marinoni, A., Gamba, P., Gamba, P., Chakravortty, S. & Chakravortty, S. (2019). DAEN : Deep Autoencoder Networks for Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 4309–4321. doi : 10.1109/tgrs.2018.2890633.

Su, Y., Zhu, Z., Gao, L., Plaza, A., Li, P., Sun, X. & Xu, X. (2024). DAAN : A Deep Autoencoder-Based Augmented Network for Blind Multilinear Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15. doi : 10.1109/TGRS.2024.3381632.

Takeuchi, K., Ishiguro, K., Kimura, A. & Sawada, H. (2013). Non-Negative Multiple Matrix Factorization. *IJCAI*, 13, 1713–1720.

Teixeira, N., Nabais, P., de Freitas, V., Lopes, J. A. & Melo, M. J. (2021). In-depth phenolic characterization of iron gall inks by deconstructing representative Iberian recipes. *Scientific Reports*, 11(1), 8811.

Thurau, C., Kersting, K., Wahabzada, M. & Bauckhage, C. (2011). Convex non-negative matrix factorization for massive datasets. *Knowledge and information systems*, 29, 457–478.

Tonazzini, A., Bedini, L. & Salerno, E. (2004). Independent component analysis for document restoration. *Document Analysis and Recognition*, 7(1), 17–27.

Tonazzini, A., Salerno, E., Abdel-Salam, Z. A., Harith, M. A., Marras, L., Botto, A., Campanella, B., Legnaioli, S., Pagnotta, S., Poggialini, F. & Palleschi, V. (2019). Analytical and mathematical methods for revealing hidden details in ancient manuscripts and paintings : A review. *Journal of Advanced Research*, 17, 31–42. doi : https://doi.org/10.1016/j.jare.2019.01.003.

Toth, M. B. (2004). The Archimedes Palimpsest. Repéré le 2025-05-15 à https://www.archimedespalimpsest.org/.

Ursescu, M., Malutan, T. & Ciovica, S. (2009). Iron gall inks influence on papers' thermal degradation FTIR spectroscopy applications. *Eur. J. Sci. Theol*, 5(3), 71–84.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, pp. 5998–6008.

Vavasis, S. A. (2010). On the complexity of nonnegative matrix factorization. *SIAM journal on optimization*, 20(3), 1364–1377.

Wang, Y.-X. & Zhang, Y.-J. (2011). Image inpainting via weighted sparse non-negative matrix factorization. *2011 18th IEEE International Conference on Image Processing*, pp. 3409–3412.

Wang, Y.-X. & Zhang, Y.-J. (2012). Nonnegative matrix factorization : A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6), 1336–1353.

Wold, S., Esbensen, K. & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37–52.

Yang, Z. & Oja, E. (2010). Linear and Nonlinear Projective Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks*, 21(5), 734-749. doi : 10.1109/TNN.2010.2041361.

Yi, Y., Wang, J., Zhou, W., Zheng, C., Kong, J. & Qiao, S. (2019). Non-negative matrix factorization with locality constrained adaptive graph. *IEEE Transactions on circuits and systems for video technology*, 30(2), 427–441.

Yoo, J.-H. & Choi, S.-J. (2010a). Nonnegative matrix factorization with orthogonality constraints. *Journal of computing science and engineering*, 4(2), 97–109.

Yoo, J. & Choi, S. (2010b). Orthogonal nonnegative matrix tri-factorization for co-clustering : Multiplicative updates on stiefel manifolds. *Information processing & management*, 46(5), 559–570.

Yu, S., Pan, Y., Zeng, Y., Doshi, P., Liu, G., Poh, K.-L. & Lin, M. (2024). An Autoencoder-Like Nonnegative Matrix Co-Factorization for Improved Student Cognitive Modeling. *Advances in Neural Information Processing Systems*, 37, 121007–121037.

Yu, X., Hu, D. & Xu, J. (2013). *Blind source separation : theory and applications*. John Wiley & Sons. doi : 10.1002/9781118679852.

Zhang, D., Zhou, Z.-H. & Chen, S. (2006). Non-negative matrix factorization on kernels. *PRICAI 2006 : Trends in Artificial Intelligence : 9th Pacific Rim International Conference on Artificial Intelligence Guilin, China, August 7-11, 2006 Proceedings 9*, pp. 404–412.

Zhang, J., Yang, Y. & Ding, J. (2023). Information criteria for model selection. *Wiley Interdisciplinary Reviews : Computational Statistics*, 15(5), e1607.

Zhao, M. & Liu, J. (2021). Robust clustering with sparse corruption via $l2,1$, $l1$ norm constraint and Laplacian regularization. *Expert Systems with Applications*, 186, 115704.

Zhao, M., Wang, M., Chen, J. & Rahardja, S. (2022). Hyperspectral Unmixing for Additive Nonlinear Models With a 3-D-CNN Autoencoder Network. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-15. doi : 10.1109/TGRS.2021.3098745.

Zheng, H., Li, Z., Sun, C., Zhang, H., Liu, H. & Wei, Z. (2024). Blind Unmixing Using Dispersion Model-Based Autoencoder to Address Spectral Variability. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-14. doi : 10.1109/TGRS.2024.3399003.

Zhou, F., Hang, R., Liu, Q. & Yuan, X. (2019). Hyperspectral image classification using spectral-spatial LSTMs. *Neurocomputing*, 328, 39-47. doi : https://doi.org/10.1016/j.neucom.2018.02.105. Chinese Conference on Computer Vision 2017.

LISTE DE RÉFÉRENCES

Torralba, A., Isola, P. & Freeman, W. T. (2024). *Foundations of computer vision*. MIT Press.